UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Extension automatique de l'annotation d'images pour la recherche et la classification

THÈSE

présentée et soutenue publiquement le 9 mai 2018

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Abdessalem BOUZAYANI

Composition du jury

Président :	Véronique EGLIN	Professeur, Université de Lyon
Rapporteurs :	Nicole VINCENT	Professeur, Université de Paris-Descartes
	Christian VIARD-GAUDIN	Professeur, Université de Nantes
Examinateur :	Sabine BARRAT	Maître de conférences, Université de Tours
Directeur de thèse :	Salvatore TABBONE	Professeur, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Remerciements

Je remercie tout d'abord Dieu tout puissant de m'avoir donné le courage, la force et la patience d'achever ce travail.

J'exprime ma sincère gratitude à mon superviseur Professeur Salvatore-Antoine Tabbone pour le soutien continu durant ma thèse, pour sa patience, sa motivation et son immense savoir. Ses qualités humaines et ses conseils m'ont été d'un grand soutien et ont largement contribué à la réalisation de cette thèse. Je n'aurais pas pu imaginer avoir un meilleur conseiller pour mes études de doctorat.

En plus de mon conseiller, j'aimerais remercier Sabine Barrat, pour ses commentaires, son encouragement et son aide précieuse.

Je tiens aussi à remercier les dirigeants et les employés de l'entreprise Xilopix de leur suivi et leur accueil chaleureux pendant mes visites.

J'exprime mes remerciements à l'ensemble des membres de mon jury : Véronique Eglin, Nicole Vincent , Christian viard-gaudin et Sabine Barrat d'avoir accepté d'évaluer ce travail.

Je remercie tous mes amis du laboratoire LORIA pour les discussions stimulantes et pour tout l'amusement que nous avons eu au cours de ces années. En particulier, Nabil et Nihel.

Un remerciement vif à toute ma famille et ma merveilleuse fiancée de m'avoir soutenu spirituellement tout au long de l'écriture de cette thèse.

Je dédie cette thèse aux plus proches de mon cœur
mon cher père,
ma chère mère,
toute ma famille,
ma loulti,
mes amis.

Sommaire

Table des figures

Liste des tableaux

Résumé

Cette thèse traite le problème d'extension d'annotation d'images. En effet, la croissance rapide des archives de contenus visuels disponibles, par exemple les sites internet de partage de photos ou vidéos, a engendré un besoin en techniques d'indexation et de recherche d'information multimédia. L'annotation d'images permet l'indexation et la recherche dans des grandes collections d'images d'une façon facile et rapide. L'indexation textuelle et manuelle conduit à de meilleurs résultats qu'une indexation visuelle et automatique, mais elle est coûteuse pour l'utilisateur. À partir de bases d'images partiellement annotées manuellement, nous souhaitons compléter les annotations de ces bases, grâce à l'annotation automatique, pour pouvoir rendre plus efficaces les méthodes de recherche et/ou classification d'images.

Pour l'extension automatique d'annotation d'images, nous avons utilisé les modèles graphiques probabilistes qui permettent de traiter le problème des données manquantes et d'offrir la possibilité de combiner plusieurs sources d'information. Le modèle proposé est un mélange de distributions multinomiales et de mélanges de Gaussiennes où nous avons combiné des caractéristiques visuelles et textuelles. Nous avons aussi utilisé ce modèle pour l'annotation et la classification de documents.

Le modèle proposé nécessite une étape préliminaire d'apprentissage où les images sont annotées manuellement. L'annotation manuelle est très longue et coûteuse pour l'utilisateur. Pour réduire le coût d'annotation manuelle et améliorer la qualité de l'annotation obtenue, nous avons intégré des retours utilisateur dans notre modèle. Les retours utilisateur ont été effectués en utilisant l'apprentissage dans l'apprentissage, l'apprentissage incrémental et l'apprentissage actif.

Pour combler le problème du fossé sémantique et enrichir l'annotation d'images, nous avons utilisé une hiérarchie sémantique en modélisant de nombreuses relations sémantiques entre les mots-clés d'annotation. Nous avons donc présenté une méthode semi-automatique pour construire une hiérarchie sémantique à partie d'un ensemble de mots-clés. Cette hiérarchie se base sur l'utilisation de trois types d'information pour chaque mot-clé : visuelle, contextuelle et sémantique. Après la construction de la hiérarchie, nous l'avons intégrée dans notre modèle d'annotation d'images. Le modèle obtenu avec la hiérarchie est un mélange de distributions de Bernoulli et de mélanges de Gaussiennes.

Mots-clés: Annotation d'images, extension d'annotation, modèle graphique probabiliste, retour utilisateur, hiérarchie sémantique.

Abstract

This thesis deals with the problem of image annotation extension. Indeed, the fast growth of available visual contents, such as websites for sharing photos or videos, has led a need for indexing and searching of multimedia information methods. Image annotation allows indexing and searching in a large collections of images in an easy and fast way. The textual and manual indexation lead to better results comparing to a visual and automatic indexation, but it is expensive. We aim, from partially manually annotated images databases, to complete automatically the annotation of these sets, in order to make more efficient image search and / or classification methods.

For automatic image annotation extension, we use probabilistic graphical models which allow to treat the problem of missing data and to offer the possibility of combining multiple sources of information. The proposed model is based on a mixture of multinomial distributions and Gaussian combining visual and textual characteristics. We also used this model for the annotation and classification of documents.

The proposed model requires a preliminary learning step where the images are annotated manually. The manual annotation is very long and expensive. To reduce the cost of manual annotation and improve the quality of the annotation obtained, we have incorporated user feedback into our model. User feedback was done using learning in learning, incremental learning and active learning.

To reduce the semantic gap problem and to enrich the image annotation, we use a semantic hierarchy by modeling many semantic relationships between keywords. We present a semi-automatic method to build a semantic hierarchy from a set of keywords. This hierarchy is based on three types of information for each keyword : visual, contextual and semantic. After building the hierarchy, we integrate it into our image annotation model. The obtained model is a mixture of Bernoulli distributions and Gaussian mixtures.

Keywords: Image annotation, annotation extension, probabilistic graphical model, user feedback, semantic hierarchy.

Chapitre 1

Introduction générale

Sommaire

1.1 Contexte

La citation célèbre de Confucius [1] *« Une image vaut mille mots »* résume bien l'importance de l'image dans notre vie. En effet, une image a une grande capacité pour transmettre rapidement autant de sens avec si peu, voire pas du tout, d'explications. Les images numériques sont devenues inévitables dans la vie professionnelle et privée des gens modernes. Ces dernières années, l'utilisation fréquente d'images numériques est devenue nécessaire dans différents domaines tels que la médecine, l'assurance et les systèmes de sécurité, la publicité, le commerce, ainsi que dans d'autres secteurs d'activité. D'autre part, dans la vie privée, les images numériques sont utilisées pour documenter les gens proches de nous, les animaux, des monuments et des événements tels que les anniversaires, les fêtes, les voyages, les excursions et les activités sportives. Cette utilisation généralisée a provoqué une augmentation rapide du nombre d'images numériques qui, aujourd'hui, sur les sites spécialisés, peuvent être comptées en milliards. Par exemple, selon des statistiques qui ont été faites début 2017 [2], 136 milles photos sont téléchargées sur Facebook toutes les 60 secondes, 95 millions de photos et vidéos sont postées sur Instagram tous les jours et plus de 20 milles photos sont partagées chaque seconde sur Snapchat.

La quantité d'images générées chaque jour par les sites Web et les archives personnelles sont en croissance permanente. Ces archives atteignent des tailles inimaginables. La popularité de ces grandes collections d'images numériques dépend de leur facilité d'utilisation. Cependant, toutes ces bases de données ne sont pas souvent équipées d'information d'indexation adéquates. Cela rend beaucoup plus difficile l'accès à des informations intéressantes pour l'utilisateur.

1. https ://fr.wikipedia.org/wiki/Confucius
2. source : https ://www.socialpilot.co/blog/151-amazing-social-media-statistics-know-2017

Ainsi, pour effectuer la recherche d'information sur les images, un mécanisme approprié est nécessaire pour caractériser le contenu de l'image de manière significative.

La recherche d'une image à partir d'une grande base de données d'images est une tâche très difficile. Nombreuses méthodes ont été proposées pour accéder à une image [YMS14, ZLT17]. L'image peut être récupérée en utilisant un contenu visuel de bas niveau telles que la forme, la couleur et la texture ou en utilisant des étiquettes ou des mots-clés qui décrivent la signification sémantique de l'image donnée. Pour accéder à des images à l'aide de fonctions visuelles de bas niveau, l'utilisateur doit donner une image en entrée d'une requête et le résultat de recherche donne un ensemble d'images qui sont visuellement similaires à l'image de requête donnée. Mais il est très difficile pour de nombreux utilisateurs d'obtenir une image de requête à chaque fois qui réponde à leur exigence. La recherche d'image basée sur le contenu (en anglais Content-Based Image Retrieval : CBIR) est une méthode qui récupère l'image en fonction de caractéristiques visuelles de bas niveau. Pour surmonter les problèmes des systèmes CBIR, une autre méthode consiste donc à attribuer des étiquettes à toutes les images de la base de données. Ces dernières peuvent être récupérées en fonction de ces étiquettes. Le principal avantage d'une telle méthode est que l'utilisateur puisse récupérer l'image de la même manière qu'il récupère un document texte. Cette méthode d'affectation d'étiquettes est appelée annotation d'images.

L'annotation en général vise donc à assigner à des images et des vidéos un ensemble de labels qui décrivent leurs contenus aux niveaux syntaxiques et sémantiques. A l'aide de ces labels, la gestion et la manipulation des images et des vidéos peut être considérablement facilitées. En particulier, l'annotation d'images par mots-clés constitue une manière possible d'associer de la « sémantique » à une image. En effet, elle consiste à assigner à chaque image, un mot-clé ou un ensemble de mots-clés, destiné(s) à décrire le contenu sémantique de l'image. Ainsi, cette opération peut être vue comme une fonction permettant d'associer de l'information visuelle, représentée par les caractéristiques de bas niveau (forme, couleur, texture, ...) de l'image, à de l'information sémantique, représentée par ses mots-clés. Le défi majeur dans l'annotation d'images est de combler le fossé sémantique [SWS+00] entre les caractéristiques calculables de bas niveau et l'interprétation humaine des images. L'interprétation comprend des concepts sur différents niveaux d'abstraction qui ne peuvent pas être simplement mis en correspondance avec des fonctionnalités, mais nécessitent un raisonnement supplémentaire avec les connaissances générales et spécifiques à un domaine. Par exemple, les images (a) et (b) dans la figure 1.1 ont la même signification sémantique mais des apparences totalement différentes et les images (b) et (c) sont similaires visuellement mais n'ont pas la même signification sémantique.

(a) pêche (b) pêche (c) soleil

FIGURE 1.1 – Exemple du fossé sémantique

L'approche naïve consiste à annoter manuellement les images [DGL+11] par un ensemble d'utilisateurs, avec des mots-clés souvent issus d'un ensemble de mots-clés prédéfini appelé « vocabulaire ». Cette indexation manuelle, bien qu'elle soit performante pour rechercher/classifier des images est très coûteuse et se révèle pratiquement inapplicable aux grandes bases d'images. Par ailleurs, elle peut générer des ambigüités car différents utilisateurs peuvent utiliser différents mots-clés pour décrire la même image. Par conséquent, des approches utilisant des thésaurus [THA12] ont été proposées afin de réduire ces ambiguïtés potentielles.

Les approches viables dans le cadre de grands corpus concernent des techniques d'annotation automatique [ZIL12, Wan11] ou semi-automatique [Han08, WDS+01, ZLX10]. Les approches semi-automatiques nécessitent l'intervention de l'utilisateur pour valider des annotations automatiques, dans un système de retour de pertinence, par exemple. La plupart des techniques d'annotation automatique visent à apprendre, grâce à des méthodes d'apprentissage automatique et à partir d'exemples d'images annotées, des relations entre caractéristiques visuelles et/ou mots-clés. Globalement, cela revient à modéliser des similarités d'image à image [GMVS09], d'image à mots-clés [CCMV07] ou une combinaison des deux [VJ12]. Les relations apprises sont ensuite utilisées afin d'attribuer des mots-clés à des images non annotées. Le problème majeur de la plupart des méthodes existantes réside dans le fait qu'un mot-clé soit choisi pour annoter une image sans prendre en compte les autres mots-clés éventuels de cette image.

De manière plus générale, l'avancée des techniques de numérisation, la démocratisation du matériel multimédia et l'essor des services multimédia ont considérablement augmenté le nombre et la taille des bases de données images et ont engendré un besoin accru en indexation et en recherche d'images. L'indexation d'images consiste à extraire des caractéristiques à partir du contenu des images, et à les organiser de façon à permettre, ensuite, à l'utilisateur, de procéder à une recherche rapide et efficace des images. Les caractéristiques peuvent être liées au texte représentées par des mots-clés ou à des indices de bas niveau représentés par des vecteurs de forme, couleur ou texture. Dans le premier cas on parle d'indexation textuelle et dans l'autre cas d'indexation visuelle. Même si l'on peut reprocher à l'utilisation du mot-clé la lourdeur due à la nécessité d'une indexation préalable de la base d'images, il reste néanmoins un des meilleurs médiateurs entre la machine et l'homme car contrairement aux indices visuels, le texte véhicule des concepts sémantiques.

Dans l'objectif de satisfaire ces différents besoins, cette thèse va s'articuler autour des thématiques d'annotation d'images, de retour utilisateur et de hiérarchie sémantique. Dans cette perspective, la figure 1.2 illustre le contexte et les contribution de ce travail. Un utilisateur qui veut chercher des images dans ces bases fournit un mot-clé (requête textuelle par exemple contenant le mot « animaux ») ou une image exemple (requête visuelle contenant une image de plusieurs animaux). La réponse à sa requête textuelle n'est pas toujours possible car les images ne sont pas accompagnées d'informations textuelles décrivant leurs contenus. La réponse à sa requête visuelle n'est pas seyante à cause du fossé sémantique des systèmes (CBIR). Pour résoudre ce problème, l'indexation textuelle des bases est nécessaire en annotant les images par des mots-clés. Ces derniers peuvent être obtenus en utilisant l'annotation manuelle par plusieurs annotateurs. Cette annotation manuelle bien qu'elle soit performante, est fastidieuse et coûteuse. La solution la plus adéquate est alors l'annotation automatique (chapitre 3). Bien que cette dernière soit rapide, elle est moins performante que l'annotation manuelle et nécessite une étape préliminaire pour créer un ensemble manuel d'apprentissage. Pour améliorer les perfor-

mances de cette méthode et réduire l'effort manuel initial, nous pouvons introduire l'utilisateur dans le processus (chapitre 4). Ce pendant, l'utilisateur qui a lancé une recherche n'as pas reçu de réponse parce que le vocabulaire utilisé par les annotateurs ne contient pas le mot « animaux ». Pour enrichir ce vocabulaire et réduire le problème du fossé sémantique, nous intégrons une hiérarchie sémantique qui permet d'exploiter les relations sémantiques entre les mots-clés du vocabulaire (chapitre 5). Après l'utilisation d'une hiérarchie sémantique, l'annotation est automatiquement enrichie par le mot « animaux » à toutes les images déjà annotées par un mot de la famille des animaux (« bear », « horse », « tiger »,...). L'utilisateur reçoit donc une réponse à sa requête initiale.

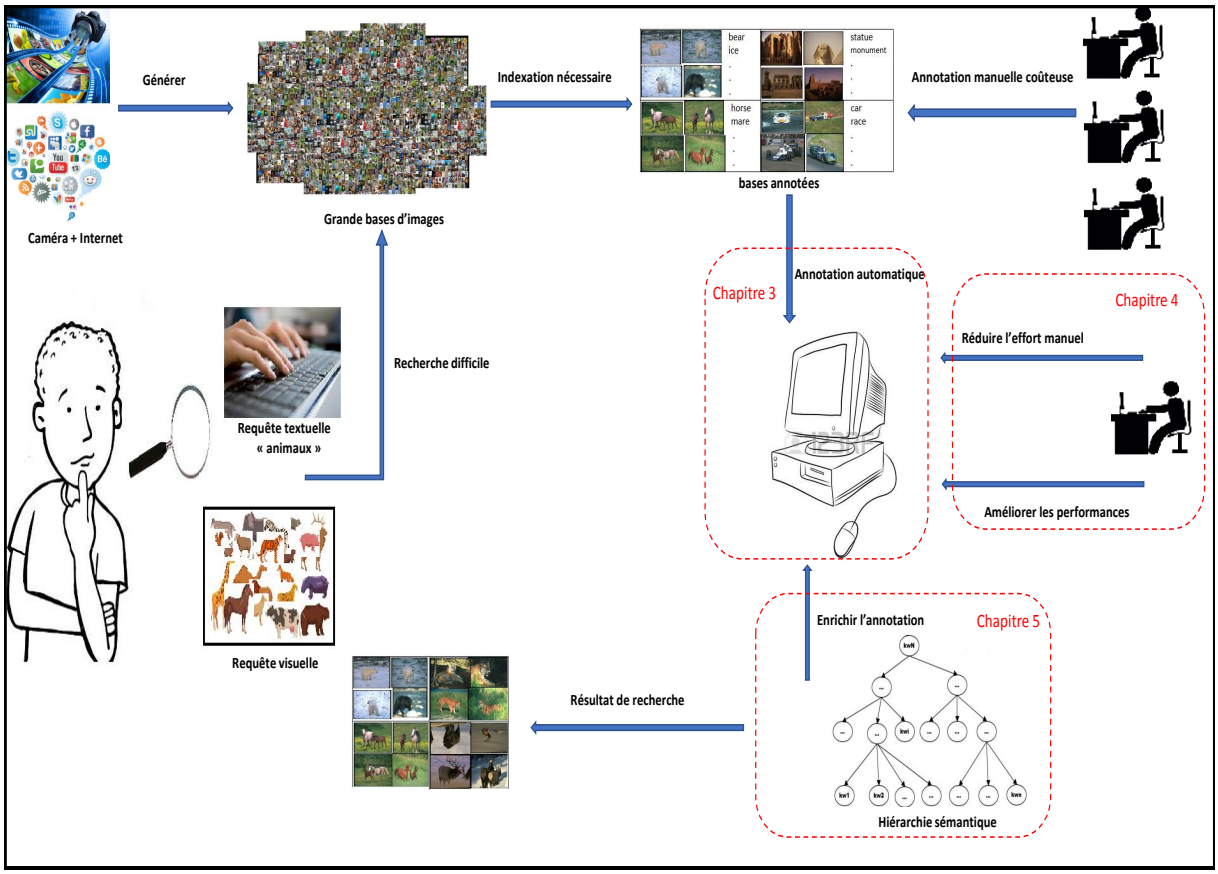


FIGURE 1.2 – Schéma général

1.2 Contributions

Partant du constat que le texte conduit à de meilleurs résultats qu'une indexation visuelle et automatique, mais qu'il est coûteux pour l'utilisateur un bon compromis apparaît : à partir de bases d'images partiellement annotées manuellement, nous souhaitons, grâce à l'annotation automatique, compléter les annotations de ces bases, pour pouvoir rendre plus efficaces les méthodes de recherche et/ou classification d'images. Pour mieux réduire le coût de l'effort manuel, nous souhaitons aussi intégrer l'utilisateur dans le processus d'annotation à travers des

retours utilisateur. L'utilisation d'algorithmes d'apprentissage seuls est insuffisante pour combler le problème du fossé sémantique [SWS+00], et donc pour produire des systèmes efficaces pour l'annotation automatique d'images, nous souhaitons utiliser une hiérarchie sémantique prenant en compte les relations sémantiques entre les mots-clés du vocabulaire d'annotation et permettant aussi d'enrichir l'annotation avec de nouveaux mots-clés.

Les contributions de ce travail peuvent donc être résumées en trois points :

i) Proposition d'un modèle graphique probabiliste pour l'extension d'annotation et la classification d'images. Ce modèle est un mélange de distributions multinomiales et de mélanges de Gaussiennes qui combinent des caractéristiques visuelles et textuelles.

ii) Intégration de trois niveaux de retours utilisateurs dans le modèle proposé. Ces retours utilisateurs permettent de réduire l'effort fastidieux de l'annotation manuelle et améliorer les performances du modèle. La combinaison de ces retours permet de mieux atteindre l'objectif souhaité.

iii) Proposition d'une méthode semi-automatique de construction d'une hiérarchie sémantique à partir d'un ensemble de mots-clés. L'intégration de la hiérarchie sémantique construite dans le modèle augmente considérablement les performances d'annotation. En outre, elle permet d'enrichir l'annotation d'images en utilisant de nouveaux mots-clés qui n'appartenaient pas au vocabulaire d'annotation initial.

1.3 Organisation du mémoire

La suite de ce mémoire est organisée comme suit :

Dans le **chapitre 2**, nous présentons une étude bibliographique approfondie sur différentes méthodes d'annotations proposées dans la littérature. Ensuite, nous présentons les mesures d'évaluation des systèmes d'annotation et les bases de données utilisées dans cette thèse. La dernière partie de ce chapitre est consacrée à une synthèse des méthodes de l'état de l'art tout en se concentrant sur les avantages et les inconvénients de chacune pour motiver nos choix.

Dans le **chapitre 3**, nous présentons le modèle que nous avons proposé pour l'extension d'annotation d'images. Nous commençons par un rappel sur le principe des modèles graphiques probabilistes et en particulier les réseaux bayésiens. Nous décrivons ensuite les caractéristiques visuelles utilisées dans cette thèse. Ensuite, nous présentons le modèle que nous avons proposé. Nous présentons en détail la structure, les paramètres et comment utiliser l'inférence dans notre modèle pour la classification et l'extension d'annotation d'images. La dernière partie de ce chapitre est consacrée à l'évaluation du modèle sur trois bases d'images et une base de documents.

Dans le **chapitre 4**, nous nous intéressons à trois niveaux de retours utilisateurs afin d'améliorer les performances de notre modèle et réduire le laborieux effort manuel d'annotation. D'abord, nous présentons le principe de fonctionnement du premier retour utilisateur à base d'apprentissage dans l'apprentissage ainsi que son évaluation sur les bases d'images. Ensuite, nous présentons le principe et l'évaluation du deuxième retour utilisateur à base d'apprentissage incrémental. Puis, nous décrivons le principe et l'évaluation du troisième retour utilisateur à base d'apprentissage actif. Enfin, nous montrons comment combiner les trois retours utilisateur dans un seul système.

Dans le **chapitre 5**, nous décrivons notre approche d'utilisation d'une hiérarchie sémantique pour exploiter les relations entre les mots-clés et enrichir le vocabulaire d'annotation par de nouveaux mots. D'abord, nous présentons les informations utilisées pour construire la hiérarchie. Ensuite, nous présentons l'algorithme de construction de la hiérarchie ainsi que son intégration dans le modèle proposé. Enfin, nous présentons les résultats expérimentaux après l'intégration de la hiérarchie construite.

Nous terminerons ce mémoire par une conclusion sur les travaux accomplis et présenterons quelques perspectives dans le **chapitre 6**.

Chapitre 2

Annotation d'images : état de l'art

Sommaire

2.1 Introduction

Dans ce chapitre, nous nous intéressons à la problématique de l'annotation d'images. En effet, la croissance rapide des archives de contenus visuels disponibles, par exemple les sites internet de partage de photos ou vidéos, a engendré un besoin en techniques d'indexation et de recherche d'information multimédia, et plus particulièrement d'images. L'annotation d'images permet l'indexation et la recherche dans des grandes collections d'images d'une façon facile et rapide. La première partie de ce chapitre est consacrée à une étude détaillée des méthodes d'annotation d'images. Nous faisons la distinction entre annotation manuelle, automatique et semi-automatique tout en se concentrant sur les méthodes d'annotation automatique présentées dans la littérature car il s'agit du contexte de recherche de cette thèse. La deuxième partie est consacrée à la présentation des mesures d'évaluation des systèmes d'annotation et à la description des bases de données utilisées dans cette thèse. Nous terminons ce chapitre par une synthèse des méthodes de l'état de l'art pour motiver nos choix dans cette thèse.

2.2 Méthodes d'annotation d'images

L'annotation d'images par mots-clés constitue une manière d'associer de la « sémantique » à une image. En effet, elle consiste à assigner à chaque image, un mot-clé ou un ensemble de mots-clés, destiné(s) à décrire le contenu sémantique de l'image. On distingue trois types d'annotation : annotation manuelle, annotation semi-automatique et annotation automatique.

2.2.1 Annotation manuelle

L'annotation manuelle consiste à assigner manuellement un ensemble de mots-clés à une image par un utilisateur. Elle est nécessaire pour créer des bases d'images qui sont utilisées comme des vérités-terrain. Ces dernières sont utilisées pour valider les méthodes d'annotation automatiques.

Bien que l'annotation manuelle soit efficace puisqu'elle permet d'obtenir une description plus appropriée du contenu d'une image, elle est très longue et coûteuse. Un autre problème de l'annotation manuelle est la subjectivité des annotations obtenues suivant la personne qui annote l'image. En effet, selon Sunderland [Sut82], une même image peut avoir plusieurs significations. Pour prouver cette hypothèse, il a mené l'expérience suivante. Il a demandé à trois personnes de donner une description de la peinture "Le Christ dans la maison de ses parents" de John Everett Millais (figure 2.1).

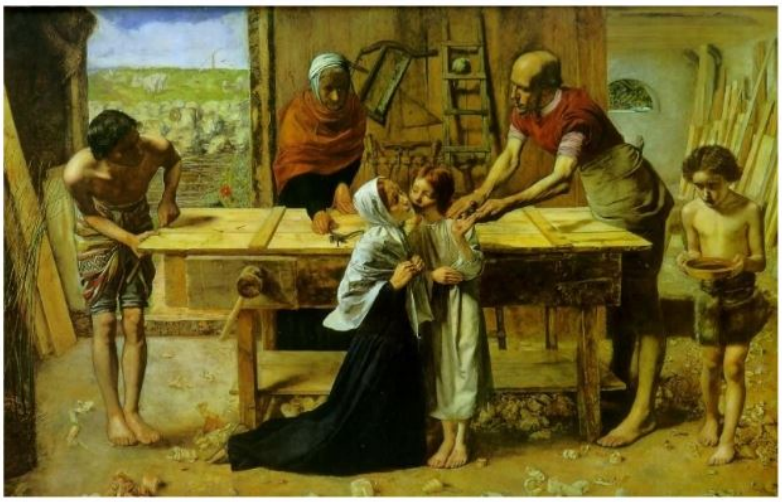


FIGURE 2.1 – Peinture de John Everett Millais 1850

- La première personne est un enfant de 12 ans qui a décrit le contenu de l'image : trois hommes, deux enfants, une femme à genou, un plancher, copeaux de bois, etc.
- La deuxième personne est un historien d'art. Ce denier a décrit les caractéristiques de la peinture : son nom, le musée où elle se trouve, son style et sa symbolique, le peintre, la date de mise en œuvre.
- La troisième personne est un profane. Il a indiqué qu'il s'agit d'une image religieuse qui représente la famille Sainte dans un atelier. Il indique aussi l'existence des sentiments de bonheur et d'amour entre l'enfant et ses parents.

Cette expérience nous montre que l'annotation d'une image est différente d'une personne à une autre. Cela dépend de ses études, sa culture, son origine, etc. Une solution à ce problème est d'annoter la même image par plusieurs personnes et de ne conserver que les annotations en commun pour obtenir une « opinion publique » sur la description de l'image. Cette solution nécessite la disponibilité de plusieurs personnes. Grâce au développement d'Internet et des sites collaboratifs, il est possible de trouver ces personnes pour annoter les images.

FIGURE 2.2 – Exemple d'annotation d'une image de flickr

Prenons l'exemple de Flickr[3] et ESP Game[4] [VAD04]. Dans le premier site, comme le montre la figure 2.2, lorsqu'une image est ajoutée sur le site, les internautes peuvent l'annoter (les mots-clés en gris). Le site aussi propose des mots-clés (mots-clés en blanc) pour cette image et qui peuvent être validés par les internautes. Ces mots-clés sont proposés en utilisant les similarités visuelles de l'image ajoutée et celles déjà annotées par les utilisateurs. L'annotation d'images dans le deuxième site repose sur un jeu en ligne. Le site propose une image à deux joueurs. Ces derniers attribuent des mots-clés à l'image et s'ils proposent le même mot-clé, elle sera validée par le système comme nouvelle annotation de l'image. Les mots-clés attribués par les deux joueurs ne doivent pas figurer dans une liste de mots tabous affichés par le site. Cette liste contient les annotations déjà validées par le système. Un exemple est représenté dans la figure 2.3. Le premier joueur a proposé les mots clés « park » et « bench ». Le système affiche le message "Matched on : bench", cela veut dire que le deuxième joueur a aussi proposé le mot-clé « bench ». La liste de mots tabous, affichée à gauche de l'image, contient les deux mots-clés « road » et « path ». Le nouveau mot-clé « bench » est ajouté à cette liste.

Plusieurs outils ont été développés ces dernières années pour l'annotation manuelle d'images comme VGG Image Annotator[5], RectLabel[6], LEAR Image Annotation Tool[7] et LabelMe[8]. Une étude détaillée de ces outils est présentée dans [DGL+11].

3. http ://www.flickr.com
4. http ://www.espgame.org/
5. http ://www.robots.ox.ac.uk/ vgg/software/via/
6. https ://rectlabel.com/
7. https ://lear.inrialpes.fr/people/klaeser/software_image_annotation
8. http ://labelme.csail.mit.edu/Release3.0/

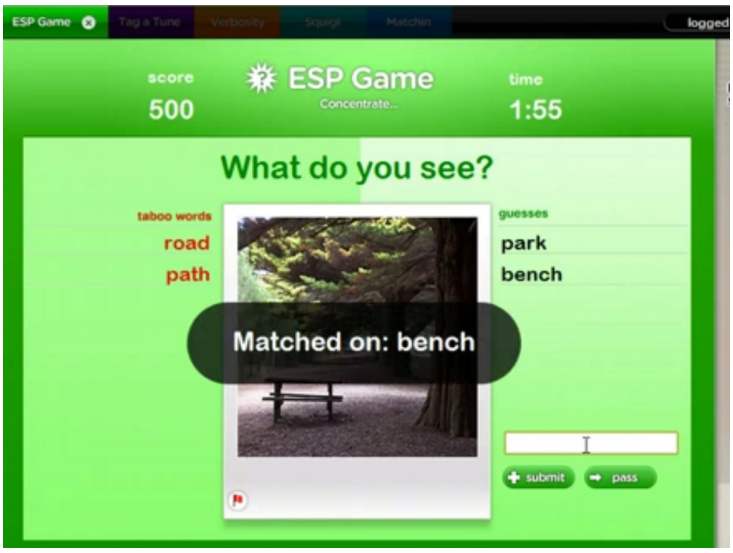


FIGURE 2.3 – Exemple d'annotation d'une image de ESP Game

2.2.2 Annotation automatique

L'objectif de l'annotation automatique d'images est d'attribuer automatiquement une collection de mots-clés d'un dictionnaire donné à une image donnée. Autrement dit, l'entrée est l'image cible et la sortie est un ensemble de mots-clés qui décrivent cette image en la meilleure façon possible.

Un ordinateur peut calculer facilement des fonctionnalités de bas niveau à partir de la couleur, la texture et la forme mais ne peut pas donner une interprétation sémantique de ces fonctionnalités contrairement à une personne qui peut déduire facilement une sémantique à partir d'une image. Le défi majeur dans l'annotation automatique d'images est donc de combler le fossé sémantique [FZQ12, THA12] entre les fonctionnalités de bas niveau calculables et l'interprétation humaine des images.

Le problème d'annotation automatique d'images a été largement étudié durant ces dernières années, et de nombreuses approches ont été proposées pour résoudre ce problème. Ces approches peuvent être regroupées en plusieurs modèles. Les principaux modèles d'annotation automatique sont les modèles graphiques, les modèles génératifs et les modèles discriminatifs.

2.2.2.1 Modèles graphiques

Un graphe est une structure qui permet de représenter un ensemble d'objets (nœuds) dans lesquels certaines paires d'objets sont liées par des arcs. C'est une manière intuitive de représenter et de visualiser des relations entre plusieurs variables. Dans les modèles graphiques, les nœuds représentent généralement des variables aléatoires modélisant le problème à résoudre, et les arcs représentent des dépendances conditionnelles entre ces variables. Ces modèles ont réussi à résoudre plusieurs problèmes d'apprentissage automatique, pour cela ils ont été utilisés pour l'annotation d'images [RLL+07, LLL+09, ZTH+14, LYRZ10, THY+11, TYH+09, JZHJ14].

Pour résoudre le problème d'annotation d'images, ces méthodes se basent sur la construction

d'un graphe de voisinage. Des caractéristiques de bas niveau sont calculées sur chaque image pour obtenir un vecteur de dimension n. Ce vecteur sert à construire le graphe de voisinage. Pour annoter une nouvelle image, elle est insérée dans le graphe et hérite des annotations de ses voisins dans le graphe. Un exemple d'un graphe de voisinage est présenté dans la figure 2.4 où l'image au centre est connectée à toutes les autres images.

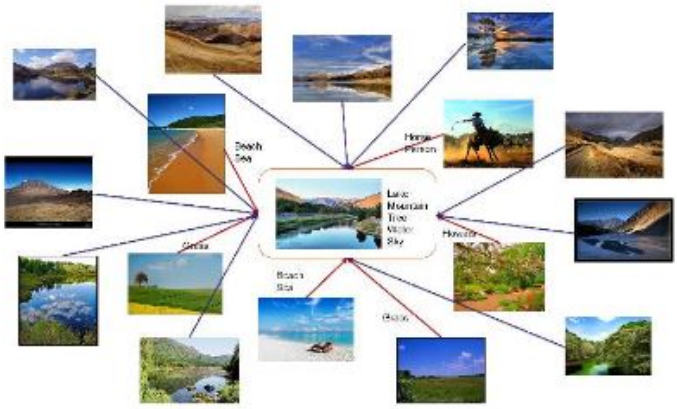


FIGURE 2.4 – Exemple d'un graphe de voisinage (source : [THY+11]

Rui et al. [RLL+07] ont proposé un modèle (BGRM) pour l'annotation d'images Web. Pour une image Web donnée, un ensemble d'annotations est extrait du texte qui l'entoure et d'autres informations textuelles dans la page Web. Cet ensemble est souvent incomplet, les annotations sont complétées en recherchant dans une base d'images à grande échelle. Les annotations sont modélisées par un graphe biparti. Un algorithme de renforcement est utilisé pour ranger les annotations. Celles avec les plus grand scores sont retenues comme annotations finales.

Liu et al. [LLL+09] ont formulé le problème d'annotation d'images dans un cadre d'apprentissage graphique. Ils ont utilisé deux types d'apprentissage : l'apprentissage graphique à base d'images dont les nœuds sont les images et les arcs sont les relations entre les images ; et l'apprentissage graphique à base de mots où les nœuds sont des mots et les arcs sont des relations entre les mots. L'apprentissage graphique basé sur l'image est d'abord effectué pour apprendre les relations entre les images et les mots, c'est à dire, pour obtenir les annotations candidates pour chaque image. Ensuite, l'apprentissage graphique à base de mots est utilisé pour affiner les relations entre les images et les mots pour obtenir les annotations finales pour chaque image.

Par ailleurs, Zhang et al. [ZTH+14] ont construit un graphe appelé ObjectPatchNet à partir des images Internet faiblement annotées. Les sommets de ce graphe (PatchNode) représentent des blocs d'images visuellement semblables. Les arcs du graphe modélisent la relation de co-occurrence entre différents objets de la même image. Ce graphe est adapté à une tâche d'annotation d'images évolutive et peut être utilisé comme un vocabulaire visuel avec des étiquettes sémantiques pour faciliter la récupération d'images.

Pan et al. [PYFD04] ont proposé une approche graphique pour affiner l'annotation d'images (GCap). Ils ont d'abord représenté une image sous la forme d'un ensemble de régions, dont chacune est décrite par un vecteur de fonctionnalités visuelles. Un graphe est construit sur l'ensemble des données d'apprentissage. Ensuite, ils ont défini trois types de nœuds dans ce graphe :

un nœud d'image représentant une image, un nœud de région représentant une région d'image et un nœud de mot représentant un mot-clé. Les liens entre les nœuds représentent la relation entre une image et ses régions et les mots-clés. Enfin, pour annoter une nouvelle image, une corrélation entre ses régions et les autres régions du graphe est recherchée et les mots-clés représentant ces dernières régions sont attribués à la nouvelle image. Un exemple de la méthode GCap est illustré dans la figure 2.5, dans laquelle trois images, leurs légendes, leurs régions et le graphe représentatif sont représentés respectivement.

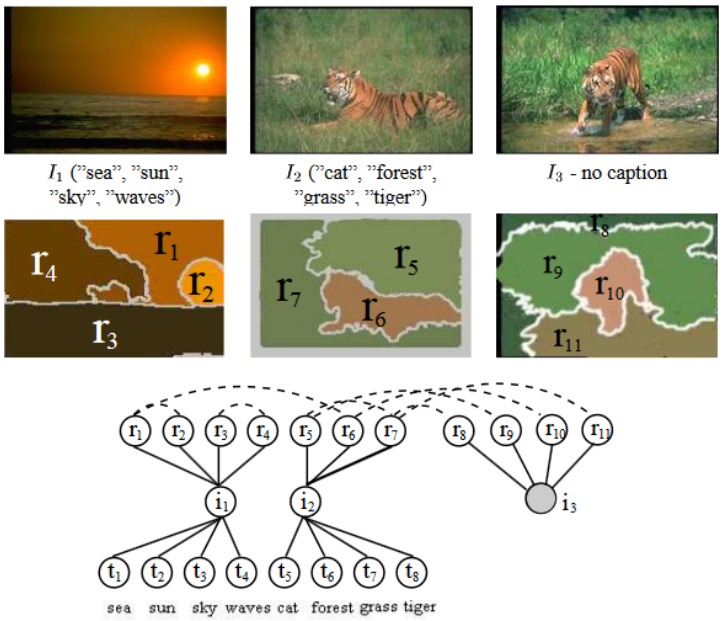


FIGURE 2.5 – Exemple illustratif de la méthode GCap (source : [PYFD04])

Dans [LYRZ10], un graphe multi-arcs est proposé pour propager l'annotation du niveau image au niveau région. Chaque sommet du graphe est d'abord caractérisé par une image unique. Ensuite, chaque image est codée en tant que sac de régions avec plusieurs segmentations d'images, et la similarité entre les régions sert à construire des multiples arêtes entre chaque paire de sommets. La stratégie de propagation d'étiquettes proposée intègre à la fois l'étiquetage automatique, le raffinement des étiquettes et l'attribution des étiquettes aux régions.

Tang et al. [THY+11] ont proposé une approche d'apprentissage semi-supervisée à base de graphes creux pour l'annotation d'images à grande échelle. Ils exploitent simultanément les données étiquetées et non étiquetées. Le graphe creux peut éliminer la plupart des arcs entre les données qui ne sont pas sémantiquement liées ; et Il est donc plus résistant et plus discriminant que les graphes classiques. Ils ont utilisé les k plus proches voisins pour accélérer la construction du graphe creux sans perdre son efficacité pour améliorer la performance d'annotation. Un travail similaire est présenté dans [TYH+09]. Un exemple de graphe creux correspondant au graphe de voisinage de la figure 2.4 est présenté dans la figure 2.6. Par rapport au graphe de voisinage, les nœuds (les images) du graphe creux sont moins connectés entre eux.

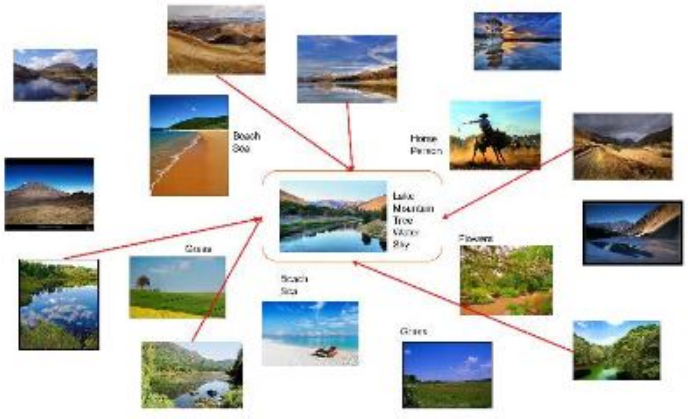


FIGURE 2.6 – Exemple d'un graphe creux (source : [THY+11])

2.2.2.2 Modèles génératifs

Les modèles génératifs construisent une distribution conjointe des descripteurs et des mots-clés d'une image pour trouver une correspondance entre les descripteurs d'une image et les mots-clés d'annotation. Ces modèles génératifs visent à apprendre un modèle simple pour tous les termes du vocabulaire. Ils se concentrent donc sur l'apprentissage des corrélations entre les caractéristiques visuelles et les concepts sémantiques. Parmi ces modèles, on trouve les modèles de traduction, les modèles de pertinence et les modèles probabilistes.

Les modèles de traduction [HTO99, DBdFF02] traitent le problème d'annotation d'images comme un problème de traduction automatique, où une région d'image doit être traduite par une étiquette sémantique. Duygulu et al. [DBdFF02] ont utilisé un modèle de traduction automatique pour associer des mots-clés et des régions d'images. Dans ce modèle, les images sont segmentées en régions. Ces dernières sont classées en utilisant une variété de descripteurs. Un apprentissage des correspondances entre les types de régions et les mots-clés fournis avec les images est effectué pour prédire des mots-clés à une nouvelle image.

Dans les modèles de pertinence tels que CMRM [JLM03], CRM [LMJ04], DCMRM [LWL+07] et MBRM [FML04], l'annotation d'images a été conçue comme un problème de calcul de la probabilité conjointe des régions des images et des étiquettes.

Jeon et al. [JLM03] ont proposé le modèle de pertinence CMRM pour estimer la probabilité conjointe des mots-clés et des images en utilisant des blobs [9] discrets pour représenter les régions. Ils ont supposé que les régions dans une image peuvent être décrites à l'aide d'un petit vocabulaire de blobs. Ces derniers sont générés à partir des descripteurs d'images en utilisant un regroupement. Ce modèle a ensuite été amélioré par le modèle de pertinence continue (CRM) [LMJ04] et le modèle de pertinence de Bernoulli (MBRM) [FML04]. Un exemple illustratif du modèle MBRM est présenté dans la figure 2.7. Dans ce modèle, la variable W représente les mots-clés d'annotation, c'est un vecteur binaire échantillonné à partir du modèle de Bernoulli. L'image est produite par un premier échantillonnage d'un ensemble de vecteurs de caractéristiques visuelles $\{g_1 ... g_n\}$, puis par la génération des régions $\{r_1 ... r_n\}$ à partir des caractéristiques visuelles. Les régions sont concaténées pour former l'image.

9. Un blob est une zone dense et connectée dans une image.

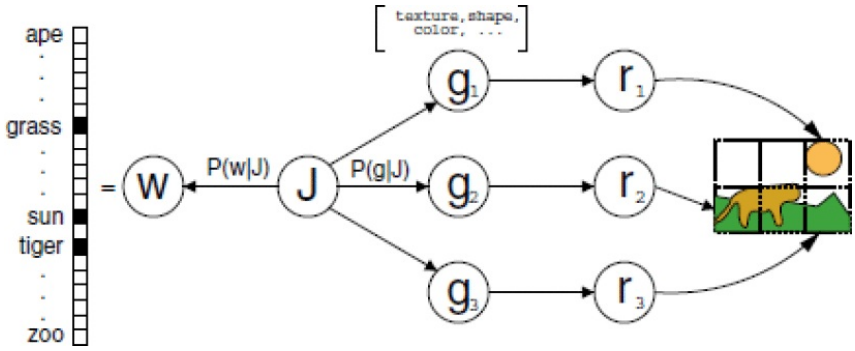


FIGURE 2.7 – Exemple d'annotation d'image avec MBRM (source : [FML04])

Liu et al. [LWL+07] ont proposé le modèle DCMRM qui estime la probabilité conjointe par l'espérance des mots dans un vocabulaire prédéfini. Ce modèle implique deux types de relations dans l'annotation d'images : la relation mot-image et la relation mot-mot. Ces deux relations peuvent être estimées en utilisant les techniques de recherche des données Web ou à partir des données d'apprentissage disponibles.

Les modèles probabilistes utilisent généralement les modèle thématiques *(topic model)* pour représenter la distribution conjointe des caractéristiques visuelles et textuelles. Les modèles thématiques [MGP04, BDF+03, NHGT14, YH08] ont été développés au début pour découvrir les structures sémantiques d'un corps de texte. Ils ont été étendu pour d'autres applications telles que l'imagerie et, en particulier, l'annotation d'images. Ils comprennent : l'analyse sémantique latente (LSA), l'analyse sémantique latente probabiliste (PLSA) et l'allocation latente de Dirichlet (LDA).

Barnard et al. [BDF+03] ont proposé une nouvelle approche pour la modélisation des ensembles de données multimodales, en mettant l'accent sur le cas spécifique des images segmentées avec le texte associé. L'apprentissage de la distribution conjointe des régions d'images et des mots d'images comporte de nombreuses applications. Les auteurs considèrent en détail la prédiction des mots associés à des images entières et correspondant à des régions d'images particulières.

Putthividhya et al. [PAN10] ont proposé un nouveau modèle (tr-mmLDA) pour la tâche d'annotation d'images et de vidéo. L'idée principale de ce modèle réside dans le fait qu'une nouvelle approche de régression est proposée pour capturer les corrélations entre les caractéristiques d'images ou de vidéos et les textes d'annotations.

Dans [CBL09], un modèle graphique probabiliste a été présenté pour la classification et l'annotation d'images. Ce modèle introduit une variable aléatoire latente pour créer le lien entre les caractéristiques visuelles et les mots-clés. Ce modèle présente l'inconvénient de nécessiter une segmentation des images sans avoir à annoter les régions. Un autre modèle graphique probabiliste pour l'annotation et la classification d'images a été proposé dans [BT10]. Dans ce modèle, l'échantillon des caractéristiques visuelles (variables continues) suit une loi dont la fonction de densité est une densité de mélange de Gaussiennes et les variables discrètes (mots-clés) suivent une distribution multinomiale.

Bien que les modèles génératifs, tels que les modèles de traduction et les modèles probabilistes, aient montré une performance très compétitive même avec un large vocabulaire d'anno-

tation, leur capacité d'apprentissage reste limitée et ils sont couteux pour l'utilisateur. La faible capacité d'apprentissage est principalement due au nombre important de paramètres à estimer qui nécessitent des hypothèses simplificatrices qui peuvent dégrader la performance souhaitée. Un autre inconvénient des modèles génératifs est qu'ils sont perturbés par des images d'apprentissage ayant des similitudes visuelles élevées, mais ayant différentes sémantiques d'où la nécessité d'un grand nombre d'images pour faire la phase d'apprentissage d'une façon plus appropriée.

2.2.2.3 Modèles discriminatifs

Les modèles discriminatifs modélisent la dépendance d'une variable non observée y sur une variable observée x. Cela veut dire modéliser la distribution de probabilité conditionnelle $P(y/x)$. Ils permettent de convertir le problème d'annotation d'images en un problème de classification. Les premiers travaux se sont concentrés sur la classification d'images en deux catégories. Dans cette approche d'annotation, les descripteurs de bas niveau sont calculés sur les images, et injectés ensuite dans un classificateur binaire qui donne une décision sur l'appartenance ou non d'une image à une classe qui représente un concept particulier. Cependant, une image est généralement annotée par plus de deux mots. Par conséquent, le problème d'annotation automatique d'images doit être considéré comme un problème de classification multi-classes. À la différence des méthodes d'annotation basées sur la classification binaire, les méthodes d'étiquetages multiples annotent une image par plusieurs concepts sémantiques. Dans ces méthodes, une image est représentée par un ensemble de descripteurs ou de régions. Une image sera annotée par une étiquette si une de ses régions est associée à un concept. Bien que les modèles discriminatifs ont une forte puissance de classification et de discrimination, ils sont inefficaces avec un grand nombre de classes. Plusieurs classificateurs ont été utilisés pour l'annotation comme les séparateurs à vaste marge (SVM) [CCS04, CHV99, SFCL04, ALLQ13], les K plus proches voisins (KPPV) [VJ12, MPK10, GMVS09], les arbres de décision [MDGW07, MGPW05, LZL08, FZQ12, JHCZ09, FM13] et les réseaux de neurones artificiels (RNA) [PLK04, KPK04].

2.2.2.3.1 SVM

SVM est une méthode de classification par apprentissage supervisé. Elle permet de séparer linéairement les exemples positifs des exemples négatifs en utilisant une fonction (noyau). Il s'agit de trouver un hyperplan pour séparer les données avec une marge maximale entre le plus proche des exemples positifs et des exemples négatifs.

Grâce à leurs bonnes performances en classification, les SVM ont été largement utilisés pour l'annotation d'images. Certains travaux de recherche [CCS04, CHV99, SFCL04] utilisent un SVM pour chaque concept. Chaque classificateur produit une décision probabiliste et la classe avec la probabilité maximale est choisie comme concept de l'image à classifier. D'autres travaux [GCL05, QH07] utilisent des ensembles multiples de classificateurs SVM. Chaque ensemble classifie indépendamment une image donnée et la décision finale est la combinaison des sorties de tous les ensembles.

Goh et al. [GCL05] ont utilisé des SVM à une classe, deux classes et multi-classe pour annoter automatiquement des images. Ils ont proposé un ensemble dynamique à base de confiance

(CDE) pour améliorer le classificateur en améliorant de façon adaptative l'ensemble des étiquettes, des caractéristiques de bas niveau et les données d'apprentissage. Ce facteur de confiance utilise un schéma de classification à trois niveaux. D'abord, CDE utilise des SVM à une classe pour caractériser un facteur de confiance pour vérifier l'exactitude d'une annotation réalisée par un SVM binaire. Le facteur de confiance est ensuite propagé aux SVM multi-classes aux niveaux ultérieurs. CDE utilise le facteur de confiance pour apporter des ajustements dynamiques à ses classificateurs membres afin d'améliorer la précision de la prédiction, de s'adapter à de nouvelles étiquettes et d'aider à la découverte de caractéristiques de bas niveau utiles.

Dans [QH07], le modèle proposé utilise une méthode automatique d'estimation du poids pour fusionner les résultats de deux ensembles complémentaires de SVM pour l'annotation automatique. Un ensemble de SVM est obtenu en formant les fonctionnalités de sacs de blocs d'images en utilisant la technique d'apprentissage par instance multiple (MIL). Les images sont divisées en blocs non superposés et la technique MIL est utilisée sur les caractéristiques de couleur et de texture en bloc pour extraire les caractéristiques du sac. Ces caractéristiques extraites sont ensuite introduites dans un ensemble de SVM pour trouver les hyperplans optimaux pour annoter chaque image d'apprentissage. Un autre ensemble de SVM est obtenu en formant les caractéristiques globales de la couleur et de la texture pour trouver les hyperplans optimaux pour annoter chaque image d'apprentissage. Une fois que ces deux ensembles complémentaires de SVM sont configurés, chaque image de test est annotée en construisant à la fois les fonctions de sacs de blocs d'image et les caractéristiques globales de couleur et de texture et en affectant ces deux fonctionnalités à leurs ensembles de SVM respectifs.

Dans [MCM14], les auteurs ont présenté un modèle hybride pour l'annotation d'images. Plus précisément, ce modèle est basé sur un SVM utilisé comme modèle discriminatif pour résoudre le problème des faibles annotations (les images ne sont pas annotées avec tous les mots-clés pertinents). Un modèle DMBRM [FML04] est utilisé comme modèle génératif pour résoudre le problème des données déséquilibrées (échantillons d'apprentissage insuffisants par étiquette).

Les SVM sont performants même en présence d'un grand nombre de variables, ils peuvent fonctionner avec un faible nombre d'exemples d'apprentissage. Cependant, ils présentent l'inconvénient d'être mal adaptés lorsque les données sont manquantes.

2.2.2.3.2 RNA

Les RNA sont un système de traitement de l'information inspiré de la capacité du cerveau humain à apprendre de l'observation. RNA est un réseau d'apprentissage qui se compose de plusieurs couches de nœuds interconnectés, qui sont également appelés neurones ou perceptrons.

Kazuhiro et al. [KH02] ont proposé un nouveau système de recherche et d'annotation d'images de paysages qui utilise des noms d'objets spécifiques en plus des mots d'impression. Dans ce système, une image est divisée en certaines régions et chaque région est classée dans les catégories ciel, terre et eau à l'aide d'un réseau neuronal. Ensuite, les caractéristiques de l'image sont extraites de chaque catégorie, et les mots d'impression sont donnés automatiquement à l'image. Après un classement initial des régions dans trois catégories, elles sont classées dans des objets plus détaillés, comme la montagne ou le ciel nuageux. Par la reconnaissance hiérarchique, les

noms d'objets spécifiques sont donnés automatiquement à une image.

Chen et al. [CFCF10, CFCF12] ont proposé un modèle de réseau neuronal avec une structure adaptative pour l'annotation d'images. Dans ce modèle, la structure adaptative se base sur des fonctionnalités visuelles globales et locales. Ce modèle est divisé en deux réseaux : réseau de reconnaissance et réseau de corrélation. Le premier réseau est constitué d'un ensemble de sous-réseaux dont le nombre est déterminé par le nombre de régions segmentées d'une image. Il produit un vecteur de mots-clés pour l'image. Le réseau de corrélation augmente les performances d'annotation en utilisant les informations de corrélation entre les mots-clés. Dans le processus d'annotation, l'image est d'abord segmentée en plusieurs régions qui créent un vecteur de caractéristiques visuelles pour chaque région. La couche d'entrée du réseau de reconnaissance reçoit le vecteur des caractéristiques d'une région et la couche de sortie génère un vecteur de mots-clés. Ce vecteur indique quel mot-clé doit être sélectionné pour étiqueter la région d'entrée. Enfin, le réseau de corrélation reçoit un vecteur de mots-clés du réseau de reconnaissance et affine l'annotation en utilisant des informations de corrélation entre les mots-clés.

Pankaj et al. [SPS13] ont proposé une nouvelle technique d'annotation d'images utilisant un réseau de neurones. Dans ce travail, ils utilisent un réseau avec 2 couches cachées, chacune possédant 25 neurones. Ils ont utilisé 50 neurones dans la couche d'entrée et 2 neurones dans la couche de sortie. D'autres travaux d'annotation d'images à base de RNN ont été proposés dans [OMF12, ZZZP08, DFPSS07, CLBS11].

Au cours des dernières années, les réseaux de neurones convolutionnels profonds (CNN) ont gagné une popularité considérable pour l'apprentissage des représentations d'image. Divers systèmes utilisant des CNN profonds et des réseaux neuronaux récurrents (RNN) ont été proposés pour l'annotation automatique d'images.

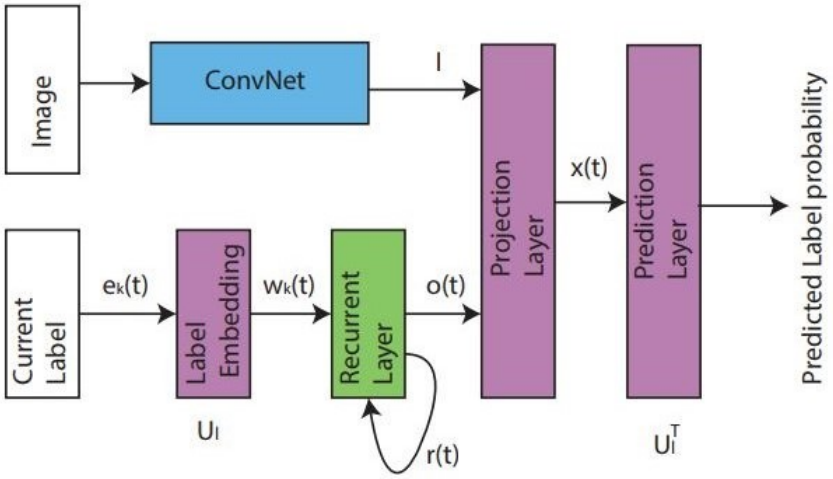


FIGURE 2.8 – Schéma général de la méthode [WYM+16]

Les réseaux de neurones récurrents (RNN) en combinaison avec les réseaux de neurones convolutionnels profonds (CNN) ont été utilisés dans un cadre unifié dans [WYM+16] pour résoudre le problème des dépendances entre les étiquettes dans une image. Le cadre CNN-RNN proposé apprend une conjointe relation étiquette-image pour caractériser les dépendances entre

les étiquettes sémantiques ainsi que la pertinence image-étiquette. Les vecteurs caractéristiques d'une image sont générés par les CNN. Les RNN sont utilisés pour calculer la probabilité d'une prédiction multi-étiquettes séquentiellement comme un chemin de prédiction ordonné, où la probabilité a priori d'une étiquette à chaque pas de temps peut être calculée en fonction du vecteur des caractéristiques de l'image et de la sortie des neurones récurrents. Le schéma général de cette méthode est présenté dans la figure 2.8.

Dans [MMM15], les auteurs ont utilisé les fonctionnalités CNN et les vecteurs de représentation de mots pour la tâche d'annotation d'images. Le modèle proposé est basé sur la structure d'analyse de corrélation canonique (CCA) qui aide à modéliser à la fois les fonctions visuelles et les fonctions textuelles. Les fonctions visuelles d'une image sont obtenues en utilisant les réseaux récurrents de neurones. Les fonctions textuelles sont extraites à l'aide d'une architecture déjà apprise de word2vec [MCCD13].

Dans [MSCM16], les auteurs ont exploité le comportement à plusieurs échelles (en anglais scale-space) dans un cadre de diffusion de chaleur hypergraphique pour la tâche d'annotation d'images automatique. Cette nouvelle technique permet de modéliser la relation d'ordre supérieur entre les images dans l'espace des caractéristiques et fournit un mécanisme de diffusion d'étiquettes multi-échelle pour résoudre le problème de déséquilibre de classes dans les données d'apprentissage. Les fonctionnalités de réseaux neuronaux convolutionnels (CNN) ont été utilisées pour modéliser la similarité des images.

Les réseaux de neurones artificiels ont la capacité d'apprendre et de modéliser des relations non linéaires et complexes. Cependant, ils présentent un problème de choix du nombre de couches cachées et du nombre de neurones dans chaque couche. Ils fonctionnent aussi comme une boite noire. En effet, la relation exacte entre l'entrée et la sortie n'est pas transparente et est parfois difficile à expliquer.

2.2.2.3.3 Arbres de décision

Un arbre de décision est un outil d'aide à la décision où un ensemble de tests sont appliqués les uns après les autres pour séparer différentes classes. La forêt aléatoire est une méthode de classification basée sur un ensemble d'arbres de décision. Elle génère un ensemble d'arbres de décision dans un espace aléatoire qui sélectionne des parties des caractéristiques de façon aléatoire pour chaque arbre de décision. Le résultat de la prédiction pour une cible inconnue est déterminé en agrégant les sorties de tous les arbres de décision.

Dans [JHCZ09], les auteurs ont présenté une méthode d'annotation automatique d'images à base d'un apprentissage bayésien sur un arbre de décision. Dans cette méthode, les caractéristiques de bas niveau sont extraites des régions segmentées des images. Les caractéristiques de haut niveau sont obtenues à partir des caractéristiques de bas niveau en utilisant un algorithme d'apprentissage à base d'un apprentissage bayésien.

Fukui et al. [FKQ11] ont été les premiers à utiliser les forêts aléatoires [Bre01] pour l'annotation automatique d'images. Ils ont proposé une amélioration de l'algorithme SML [CCMV07] en utilisant des forêts aléatoires au lieu d'un modèle de mélanges de gaussiennes. Ils ont construit un modèle de relation probabiliste entre une caractéristique locale et une étiquette sémantique.

Fu et al. [FZQ12] ont présenté une méthode d'annotation d'images en utilisant aussi les

forêts aléatoires. Ils utilisent les annotations contenues dans les images d'apprentissage en tant qu'informations de contrôle afin de guider la génération des arbres aléatoires, permettant ainsi de récupérer les voisins les plus proches non seulement au niveau visuel mais aussi au niveau sémantique. Cette méthode considère les forêts aléatoires dans son ensemble et présente deux nouveaux concepts : les plus proches voisins sémantiques(SSN) et la mesure de similarité sémantique (SSM). La figure 2.9 illustre ces deux concepts. Une image requête passe par tous les arbres aléatoires. Les images d'apprentissage stockées dans les nœuds feuilles sur lesquels la requête tombe forment un groupe de voisins sémantiques. Ensuite, la mesure de similarité sémantique entre la requête et une image d'apprentissage donnée est calculée en fonction du nombre de fois que cette image apparaît dans le SSN. Une SSM plus importante indique une similarité plus élevée.

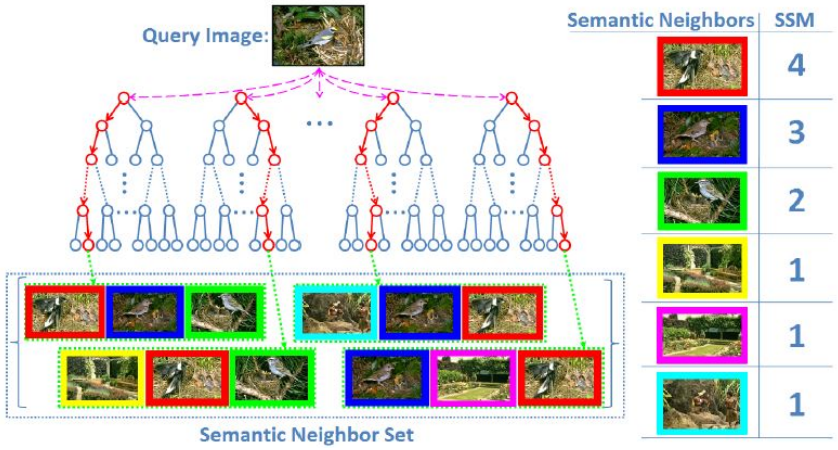


FIGURE 2.9 – Exemple d'annotation d'une image avec une forêt aléatoire (source : [FZQ12])

Les arbres de décision et les forêts aléatoires sont robustes et efficaces pour apprendre de grandes quantités de données d'entraînement. En outre, le temps de formation d'un modèle à base d'arbres de décision est habituellement plus court que la plupart des méthodes traditionnelles d'apprentissage automatique pour les ensembles de données à haute dimension. Cependant, ils présentent le problème du choix du nombre d'arbres de décision et la profondeur de chaque arbre. Ils nécessitent aussi des données discrétisées pour leur fonctionnement.

2.2.2.3.4 KPPV

Dans ces classificateurs, un nouvel élément est classifié dans la classe la plus représentée parmi ses k plus proches voisins majoritaires. Les KPPV utilisent un simple principe « Des images similaires partagent des étiquettes communes » [MPK10]. Une nouvelle image est annotée en fonction des images qui sont similaires dans l'ensemble d'apprentissage. Ensuite, les étiquettes sont classées en fonction de leurs fréquences dans les images similaires. L'annotation automatique d'images est ainsi réalisée en transférant les étiquettes les plus fréquentes sur l'image de test. Grâce à leur simplicité, les KPPV ont été largement utilisés pour l'annotation d'images.

Makadia et al. [MPK10] ont proposé la méthode JEC basée sur les voisins les plus proches. Cette méthode traite le problème d'annotation d'images comme celui de recherche d'images. Ils ont utilisé des caractéristiques visuelles globales à haute dimension plutôt que des caractéristiques visuelles locales simples. Bien que JEC est conceptuellement simple, cette méthode a permis d'obtenir les meilleurs résultats sur les jeux de données d'annotation de référence.

Inspiré par le succès de JEC, Guillaumin et al. [GMVS09] ont proposé l'algorithme TagProp basé sur la méthode KPPV. Le vote par voisinage et l'apprentissage de distance sont utilisés dans cette méthode. Pour prédire des annotations pour une image de test, ils calculent la probabilité d'utiliser des images voisines en fonction de leur rang ou de leur poids calculés à partir de la distance entre l'image de test et ses voisins. Ils ont également proposé une approche d'apprentissage métrique pour l'apprentissage de poids qui combine de multiples distances calculées en utilisant différentes caractéristiques visuelles.

Dans [VJ12], l'objectif est d'annoter automatiquement des images en utilisant la méthode 2PKNN, variante de la méthode classique KPPV, et de proposer un apprentissage métrique des poids et des distances. Pour une image non annotée, ses voisins sémantiques les plus proches correspondant à toutes les étiquettes sont identifiés. Ensuite, les étiquettes correspondant à cette image sont trouvées à partir des échantillons sélectionnés.

Li et al. [LSW09] ont proposé une mesure pour calculer la pertinence d'une étiquette pour une image donnée en prenant en compte sa fréquence dans les échantillons voisins de cette image et tout l'ensemble d'apprentissage. D'autres approches à base de KPPV ont été proposées dans [ZHH$^+$10, KIS14, LLLC12, THY$^+$11, UBBDB13, BBUDB15].

Les modèles à base de KPPV présentent l'avantage de ne pas nécessiter de phase d'apprentissage dans leur fonctionnement. Ils ont aussi donné de bons résultats pour l'annotation automatique d'images. Cependant, plusieurs problèmes peuvent survenir dans une approche à base de KPPV. L'ensemble des images voisines récupérées peut contenir de nombreuses étiquettes incorrectes à cause de l'écart sémantique [FZQ12, THA12]. Cela se produit car les fonctions visuelles peuvent ne pas être assez puissantes pour abstraire le contenu visuel de l'image. Ainsi, les algorithmes proposés ont tendance à récupérer uniquement les images dont les fonctionnalités sont très proches dans l'espace visuel, mais le contenu sémantique n'est pas bien conservé. Les méthodes à base de KPPV présentent également des inconvénients lorsque les images ne sont pas associées à suffisamment d'informations étiquetées. Cela s'explique principalement par le fait que les fréquences d'étiquettes sont généralement déséquilibrées.

2.2.2.4 Autres méthodes

Dans cette section, nous présentons d'autres méthodes d'annotation automatique d'images qui s'inscrivent dans le cadre des modèles proposés (dans les sections précédentes) mais qui méritent un focus particulier.

2.2.2.4.1 Retour de pertinence

Les chercheurs ont étudié l'utilisation de la rétroaction des utilisateurs pour améliorer l'annotation d'images, car il est difficile de construire un classificateur approprié pour annoter automatiquement les images avec des étiquettes textuelles. Plusieurs modèles de retour de pertinence ont été intégrés dans les modèles d'annotation d'images.

Dans [Chi13], Chiang a présenté un outil interactif, appelé IGAnn, sous la forme d'une procédure semi-automatique d'annotation d'images pour les utilisateurs. L'approche d'annotation dans IGAnn est basée sur le retour de pertinence. L'utilisateur identifie plusieurs images comme étant pertinentes ou non pour une étiquette donnée afin d'améliorer l'apprentissage du classificateur qui détermine lesquelles des images sont les plus susceptibles d'être associées à cette étiquette. L'apprentissage du classificateur hiérarchique est effectué par une approche semi-supervisée qui implique à la fois des images étiquetées et non étiquetées pour obtenir un bon résultat d'annotation.

Ning et al. [YHC12] ont proposé une technique de recherche multidirectionnelle (MDS) pour la propagation d'annotation d'images. Cette approche analyse dynamiquement le retour de pertinence de l'utilisateur et considère des voisinages multiples en décomposant la requête initiale en sous-requêtes séparées. À la différence des techniques KPPV basées sur un retour de pertinence qui recherchent dans un seul groupe de voisins, MDS donne un remède à cette faiblesse puisqu'elle recouvre des groupes d'images distincts. Elle peut donc annoter simultanément des images avec des caractéristiques visuelles différentes, mais avec le même concept sémantique.

Dans [ZLX10], le système présenté choisit des mots appropriés d'un vocabulaire comme des étiquettes pour une image donnée, et affine les étiquettes à l'aide du retour utilisateur. Pour une image donnée, le retour utilisateur sur les étiquettes corrige les sorties des classificateurs auxiliaires et le système recommande des étiquettes plus appropriées pour les prochaines itérations.

Mensink et al. [MVC11] ont proposé un modèle structuré pour l'étiquetage d'images qui prend en compte les dépendances entre les étiquettes des images explicitement. Ce modèle est interactif où l'utilisateur fournit la valeur de certaines étiquettes, ce qui conduit à des prévisions plus précises.

Les méthodes d'annotation utilisant un retour de pertinence améliorent significativement les résultats d'annotation automatique d'images. Cependant, elles sont coûteuses en temps pour l'utilisateur.

2.2.2.4.2 Apprentissage semi-supervisé inter-domaines

Les méthodes d'apprentissage inter-domaines [Csu17] ont été récemment proposées pour l'annotation. L'idée de base de ces méthodes est d'utiliser les données étiquetées de domaines auxiliaires. En effet, les données d'apprentissage étiquetées sont souvent fixes et assez petites, mais la quantité des données non étiquetées est vaste et en croissance rapide. Il convient donc d'utiliser à la fois les données non étiquetées dans le domaine cible et les données étiquetées dans un domaine auxiliaire pour augmenter les performances d'annotation d'images.

Wang et al.[WHS$^+$06] ont utilisé un nouvel algorithme (SSLKDE) d'apprentissage semi-supervisé pour l'annotation automatique des vidéos. Cet algorithme utilise des données étiquetées et non étiquetées pour estimer la densité de probabilité conditionnelle d'une classe.

Tang et al. [TYH$^+$09] ont proposé une nouvelle méthode d'annotation d'images en utilisant la déduction des concepts sémantiques à partir des images communautaires et leurs étiquettes bruitées associées. Pour déduire les concepts avec plus de précision, ils proposent une approche d'apprentissage semi-supervisée à base d'un graphe creux pour exploiter les données étiquetées

et non étiquetées simultanément.

Yuan et al. [YWSZ13] ont proposé une méthode d'apprentissage semi supervisée avec des groupes creux pour l'annotation d'images nommée S2CLGS. Cette méthode utilise à la fois les données d'apprentissage étiquetées et les données non étiquetées avec leurs informations structurelles diversifiées dans le domaine cible pour faire l'apprentissage semi-supervisé. La performance de S2CLGS pour l'annotation d'images est renforcée à l'aide des données étiquetées des domaines auxiliaires.

D'autres approches utilisant l'apprentissage semi-supervisé inter-domaines ont été proposées dans [DTXM09, DTXC09]. Ces approches permettent de résoudre le problème de la disponibilité de grandes collections de données étiquetées et permettent donc d'établir des modèles plus précis que ceux réalisés par des méthodes d'apprentissage purement supervisées. Cependant, ces approches nécessitent de réduire la différence de distribution des données entre les domaines auxiliaires et les domaines cibles, ce qui n'est pas toujours simple.

2.2.2.4.3 Utilisation des méta-données

Les images Web sont généralement entourées de plusieurs informations (description textuelle, URL, date, code HTML,...). Nous pouvons donc nous servir des ces informations pour annoter des images. Plusieurs techniques [LCZ+06, WZL+10, WJZZ07, JKWA05, CK12] ont été développées pour l'annotation d'images Web, la plupart de ces techniques intègre des caractéristiques visuelles et des méta-données.

Wang et al. [WZL+10] ont proposé un système automatique qui annote les images en utilisant à la fois les descriptions Web et les caractéristiques visuelles. Le système a besoin d'au moins un mot-clé initial correct et une image exemple pour lancer le processus. Le mot-clé est utilisé pour une recherche sur le Web pour trouver des images et leurs descriptions Web. Ensuite, les caractéristiques visuelles sont utilisées pour sélectionner un certain nombre d'images similaires à l'image exemple. Les descriptions Web des images sélectionnées sont regroupées en utilisant un algorithme spécial de regroupement de textes. Les mots se trouvant dans les clusters sont utilisés pour les annotations. Une méthode d'annotation similaire est présentée dans [LCZ+06, WZJM06].

Julian et al. [ML12] ont proposé une méthode pour la prédiction des étiquettes, tags et des groupes pour des images à partir des méta-données récupérées du réseau social Flickr. Les étiquettes sont fournies par des annotateurs humains en dehors de Flickr, qui fournissent des annotations basées uniquement sur le contenu de l'image. Les tags peuvent être fournis par un certain nombre de commentateurs, et peuvent inclure des informations difficiles à détecter à partir du contenu visuel seul, comme la marque de l'appareil photo et l'emplacement de l'image. Les groupes sont similaires aux étiquettes, avec la différence que les groupes auxquels une image est attribuée sont choisis uniquement par l'auteur de l'image.

Dongjin et al. [CK12] ont analysé les images et leurs données textuelles co-occurrentes dans des articles de presse pour générer des annotations possibles à ces images. Les données textuelles dans les documents d'information décrivent le point central des reportages en fonction des images et des titres donnés. Ils ont utilisé les données textuelles en tant que ressource pour affecter des annotations aux images en utilisant la valeur TF (Term Frequency) et les similarités WUP [WP94] de WordNet. La méthode proposée a montré que l'analyse de texte est une autre

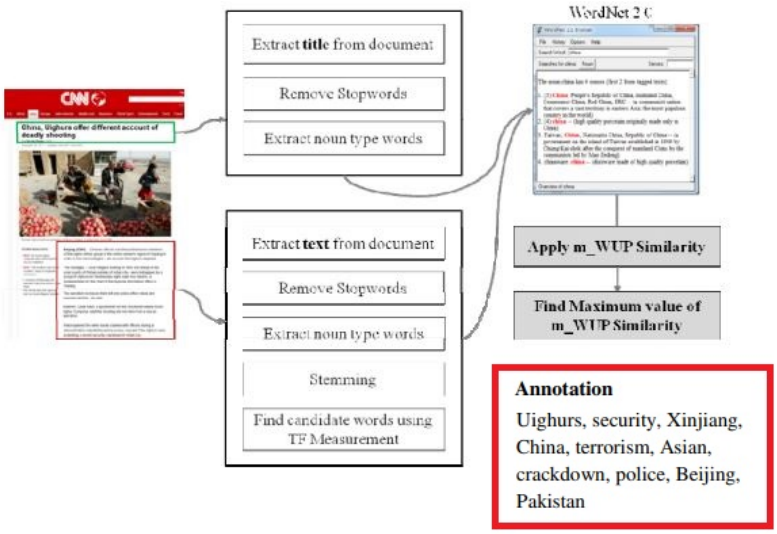


FIGURE 2.10 – Exemple d'annotation d'une image avec la méthode [CK12]

technique possible pour annoter automatiquement des images. Un exemple d'annotation d'une image en analysant le texte qui l'entoure est représenté dans la figure 2.10.

2.2.2.4.4 Hiérarchies sémantiques

Dans la plupart des approches d'annotation d'images, la sémantique est limitée à sa manifestation perceptuelle à travers l'apprentissage d'une fonction de correspondance associant les caractéristiques de bas niveau à des concepts visuels de plus haut niveau sémantique. Les performances de ces approches varient en fonction du nombre de concepts et de la nature des données ciblées. Récemment, plusieurs travaux se sont intéressés à l'utilisation des hiérarchies sémantiques pour surmonter ces problèmes.

Les hiérarchies sémantiques peuvent être groupées en quatre classes [Gar04, THA12] : taxonomie, thésaurus, ontologie et réseau sémantique. La taxonomie est utilisée pour décrire n'importe quel genre de hiérarchie entre des objets. Elle est employée pour des relations de type hyperonymie / hyponymie. Un thésaurus est une liste organisée de termes contrôlés et normalisés pour représenter les concepts d'un domaine. Il peut être considéré comme une prolongation d'une taxonomie. Une ontologie est un modèle pour la description formelle des concepts. Elle est définie par un ensemble de concepts, de propriétés et de types de relations entre les concepts. Le modèle formel devrait être compréhensible par une machine. Les réseaux sémantiques peuvent être définis [Gar04] comme intermédiaires entre les thésaurus et les ontologies. Ils décrivent plus de relations que les thésaurus mais moins formel que les ontologies.

Plusieurs approches se sont basées sur la base lexicale WordNet [Mil95] pour la construction des hiérarchies sémantiques [DDS+09, MS07, ATY+97, DGLW07, LS06, LSFF09, WML04]. L'utilisation des hiérarchies sémantiques est de plus en plus intéressante pour la tâche d'annotation d'images. Elles permettent une meilleure performance lorsque le nombre de catégories augmente. Cependant, la construction d'une hiérarchie sémantique à partir d'un vocabulaire donné est complexe.

Nous reviendrons de manière plus détaillée sur les hiérarchies sémantiques dans le chapitre 5 de ce manuscrit où nous présenterons quelques travaux de l'état de l'art ainsi que notre méthode d'annotation d'images à partir d'une hiérarchie sémantique.

2.2.2.4.5 Modèles de codage creux

Le codage creux *(sparse coding)* est le processus d'apprentissage d'une représentation éparse d'un signal en termes de coefficients d'un ensemble de signaux de base ou de variables prédictives [TF17].

Récemment, motivé par le succès du codage creux pour la classification d'images [CYC+11, GTCZ10, GCT11, HWWT14, KL11], les méthodes basées sur le codage creux ont été appliquées pour résoudre les problèmes d'annotation d'images.

Le système d'annotation proposé dans [TF17] utilise deux couches de traitement de codage creux pour traiter les images d'apprentissage comme des prédicteurs et les images de test comme le signal cible. Les images d'apprentissage sont groupées dans des thèmes en se basant sur la similitude textuelle et visuelle. La première couche profite de cette structure de groupe et identifie des thèmes pertinents pour les images de test. La couche suivante de codage creux traite l'ensemble réduit d'apprentissage et de vocabulaire, contenant des images et des mots appartenant aux thèmes pertinents seulement.

Dans [JWL+16], les auteurs ont proposé une nouvelle approche d'annotation d'images (MLDL). Cette méthode se base sur l'apprentissage du dictionnaire multi-étiquettes avec la régularisation de la cohérence des étiquettes et l'intégration d'étiquettes partiellement identiques. L'apprentissage du dictionnaire vise à apprendre un dictionnaire visuel complet à partir de l'espace des images d'apprentissage.

Cao et al. [CZG+15] ont présenté la méthode SLED (Semantic Label Embedding Dictionary) à base de codage creux pour annoter automatiquement des images. Dans cette méthode, les données d'apprentissage sont divisées au début en un ensemble de groupes superposés. Ensuite, plusieurs dictionnaires spécifiques à chaque étiquette sont entraînés afin de décorréler explicitement la représentation de chaque étiquette. Pour découvrir l'information de contexte cachée dans les étiquettes de co-occurrence, la relation sémantique entre les mots visuels dans les dictionnaires et les étiquettes est explorée dans une méthode d'apprentissage multitâche par rapport aux coefficients de reconstruction des données d'apprentissage. Enfin, une image de test est annotée à travers une méthode robuste de propagation d'étiquettes basée sur RankSVM [STL+14].

D'autres approches d'annotation d'images utilisant le codage creux ont été proposées dans [GCTR14, LTCT14, LHWW15, WLW12].

2.2.3 Annotation semi-automatique

L'annotation semi-automatique est une solution intermédiaire entre l'annotation manuelle et l'annotation automatique. Elle nécessite l'intervention d'un utilisateur pour annoter des images ou pour raffiner les résultats d'une annotation automatique. Elle est utilisée généralement dans un processus de recherche d'images. Quand un utilisateur tape un mot-clé pour chercher des images, le système de récupération propose le résultat de la recherche. L'utilisateur effectue un

retour utilisateur en choisissant les plus pertinentes. Le mot-clé de la requête sera automatiquement ajouté comme une nouvelle annotation aux images qui ont reçu une réponse positive de l'utilisateur.

Plusieurs systèmes proposent ce types d'annotation d'images. Citons l'exemple du site ALIPR (Automatic Linguistic Indexing of Pictures in Real-Time) [10] [LW03]. Lorsque une image est ajoutée sur ce site, une liste de mots-clés est affichée. L'utilisateur coche un ensemble de mots-clés de cette liste ou ajoute ses propres mots-clés qu'il juge pertinents pour l'image. Un exemple est présenté dans la figure 2.11.

L'annotation semi-automatique permet d'obtenir des meilleurs résultats en terme de précision que l'annotation purement automatique puisqu'il y a un utilisateur qui valide les annotations proposées. Elle est plus rapide que l'annotation manuelle puisque l'utilisateur choisit des mots-clés dans une liste proposée. Cependant, puisque c'est une combinaison de deux types d'annotation, elle hérite des mêmes inconvénients que l'annotation manuelle et automatique.

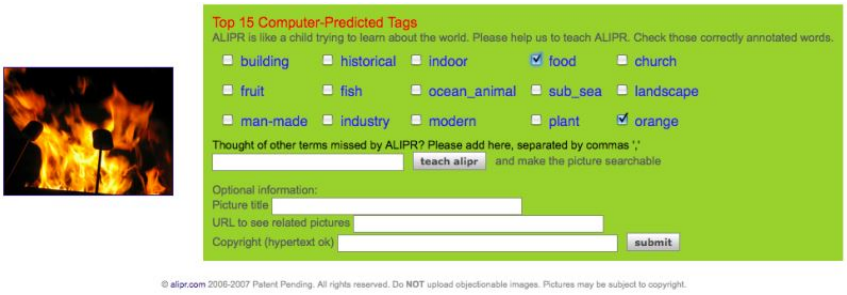


FIGURE 2.11 – Exemple d'annotation d'une image de ALIPR

2.3 Critères d'évaluation des systèmes d'annotation

Plusieurs mesures de qualité pour les systèmes d'annotation d'images sont utilisées dans la littérature. Elles peuvent être divisées, selon Kwasnicka [KP06], en deux catégories principales : mesures par annotation et mesures par mot. Dans les sections suivantes, nous détaillons ces deux catégories et les mesures retenues dans le cadre de cette thèse.

2.3.1 Mesures par annotation

Les mesures par annotation se concentrent sur le résultat d'annotation d'image par image. D'abord, les mesures sont calculées après l'annotation de chaque image. Ensuite, les mesures moyennes sont calculées pour toutes les images dans l'ensemble de test.

2.3.1.1 Taux d'annotation

Une des mesures de qualité de base pour les méthodes d'annotation automatique est le taux d'annotation. Il donne le nombre de mots correctement prédits dans l'annotation résultante. La

10. http ://http ://www.alipr.com//

mesure a la valeur 1 si tous les mots sont correctement prédits et la valeur 0 signifie qu'aucun des mots n'est correctement prédit. Le taux moyen d'annotation est une moyenne arithmétique sur toutes les images dans l'ensemble de test.

$$\tau_i = \frac{c_i}{l_i}, \qquad \tau = \frac{1}{I}\sum_{i=1}^{I} \tau_i \tag{2.1}$$

où τ_i représente le taux d'annotation de l'image i, τ représente le taux moyen d'annotation de l'ensemble de test, c_i représente le nombre de mots correctement prédits dans l'annotation de l'image i, l_i représente la longueur de l'annotation construite, et I représente la taille de l'ensemble de test.

2.3.1.2 Score normalisé

La deuxième mesure dans cette catégorie est le score normalisé. On le note par NS. Il est semblable au taux d'annotation, mais introduit en plus une pénalité pour tous les mots mal interprétés.

$$NS_i = \frac{c_i}{l_i} - \frac{d_i}{V - l_i}, \qquad NS = \frac{1}{I}\sum_{i=1}^{I} NS_i \tag{2.2}$$

où V représente la taille du vocabulaire et d_i représente le nombre de mots prédits incorrectement. Le score normalisé moyen est calculé sur toutes les annotations dans l'ensemble de test.

2.3.2 Mesures par mot

Les mesures par mot peuvent être calculées lorsque tout l'ensemble de test est annoté. Elles rassemblent les informations de chaque image annotée avec des mots. Les moyennes sont calculées pour tous les mots dans le vocabulaire.

2.3.2.1 Precision et rappel

Supposons qu'une étiquette e est présente m_1 fois dans les images de la vérité-terrain, et apparaisse dans m_2 images lors des tests à partir desquels m_3 prédictions sont correctes.

Précision : rapport entre les images correctement annotées par un mot-clé et toutes les images annotées par ce mot-clé par le modèle

$$P_e = \frac{m_3}{m_2} \tag{2.3}$$

Rappel : rapport entre les images correctement annotées par un mot-clé et toutes les images annotées par ce mot-clé dans les images de vérité-terrain

$$R_e = \frac{m_3}{m_1} \tag{2.4}$$

Pour obtenir un aperçu des performances d'un système d'annotation, on calcule les précisions et rappels moyen sur l'ensemble du vocabulaire de taille V :

$$P = \frac{1}{V}\sum_{e=1}^{V} P_e \tag{2.5}$$

$$R = \frac{1}{V}\sum_{e=1}^{V} R_e \tag{2.6}$$

2.3.2.2 Score E

Le score E est une tentative de combiner les deux mesures rappel et précision en une mesure de qualité synthétique facile à comparer :

$$E = 1 - \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{2.7}$$

2.3.2.3 F-mesure

Les F-mesure sont des moyennes harmoniques pondérées entre rappel et précision :

$$F_\alpha = \frac{(1+\alpha^2)(PR)}{\alpha^2 P + R} \text{ avec } \alpha >= 0 \tag{2.8}$$

Le paramètre α permet d'attribuer plus ou moins de poids à la précision. Quand $\alpha = 1$, rappel et précision ont un poids identique. Dans ce cas, la mesure F peut être représentée à l'aide du score E comme indiqué dans l'équation suivante :

$$F = \frac{2PR}{P+R} = 1 - E \tag{2.9}$$

2.3.2.4 N+

Une autre mesure utilisée dans les systèmes d'annotation est la mesure $N+$ qui représente le nombre de mots qui sont correctement affectés à au moins une image de test (c.à.d., nombre de mots avec rappel strictement positif). $N+$ est une mesure qui représente la quantité de mots utilisés lors du processus d'annotation. Cela peut être considéré comme une mesure de la quantité du vocabulaire qui a été couverte par la méthode.

2.3.3 Critères d'évaluation retenues

Dans ce document, nous avons choisi, comme la majorité des travaux de l'état de l'art, les mesures par mot. Nous avons utilisé les quatre critères d'évaluation : rappel, précision, F-mesure et $N+$. Lorsque le calcul de ces mesures n'est pas possible, nous avons utilisé le taux d'annotation qui fait partie des mesures par annotation.

2.4 Bases de données

Nous avons effectué nos expériences sur les trois bases d'images les plus populaires en annotation d'images : Corel-5K [DBdFF02], ESP-Game [VAD04] et IAPRTC-12 [EHG+10]. Nous avons aussi testé notre modèle sur une base de documents.

– **Corel-5K :** cette base est la plus utilisée pour l'annotation et la recherche d'images. Elle est divisée en 4500 images pour l'apprentissage et 500 images pour les tests avec un vocabulaire de 260 mots-clés. Les images sont regroupées en 50 catégories, chacune d'elles contient 100 images. Chaque image est annotée manuellement avec 1 à 5 mots-clés avec une moyenne de 3.4 mots-clés par image.

– **ESP-Game :** cette base est obtenue à partir d'un jeu en ligne. Nous avons utilisé un sous-ensemble de 20770 images utilisé dans les travaux de la littérature. Ce sous-ensemble est divisé en 18689 images pour l'apprentissage et 2081 images pour les tests avec un vocabulaire de 268 mots-clés. Chaque image est annotée avec une moyenne de 4.7 mots-clés par image.

– **IAPRTC-12 :** cette base est une collection d'environ 20000 images naturelles. Elle est composée de 17665 images pour l'apprentissage et 1962 images pour les tests avec un vocabulaire de 291 mots-clés. Chaque image est annotée avec une moyenne de 5.7 mots-clés par image.

– **Base de documents :** Cette base a été constituée par nous même. Elle contient 100 documents pour l'apprentissage et 50 documents pour les tests. Les documents sont regroupés en 10 catégories (carte vitale, carte d'identité, passeport, fiche de salaire, facture de téléphone, facture d'électricité, permis de conduire, *curriculum vitæ*, articles scientifiques et articles de presse). Les documents d'apprentissage et de test ont été annotés manuellement par 1 à 5 mots-clés en utilisant un vocabulaire de 46 mots.

Dans le tableau 2.1, nous présentons les informations détaillées de chaque base de données.

TABLE 2.1 – Détails des bases de test

Base	nombre d'images	taille du vocabulaire	taille d'apprentissage	taille de test	mots par image	images par mot
Corel-5K	5 000	260	4 500	500	3.4	58.6
ESP-Game	20 770	268	18 689	2 081	4.7	362.7
IAPRTC-12	19 627	291	17 665	1 962	5.7	347.7
Documents	150	46	100	50	3.7	12.09

La figure 2.12 représente quelques images exemples avec leurs annotations issues des quatre bases de données utilisées pour la partie expérimentation. Par exemple, la deuxième image de la base Corel-5k représente une image annotée par les mots-clés : "sky", "jet" et "plane". La première image de la base de documents représente un document (permis de conduire). Elle est annotée par les mots-clés : "nom", "adresse" et "catégorie".

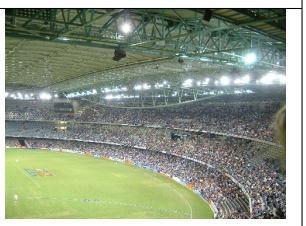
Corel-5k	 sky, sun, clouds, tree	 sky, jet, plane	 bear, polar, snow, tundra
IAPRTC-12	 grandstand, lawn, player, roof, round, spectator, stadium	 grey, helmet, jean, lamp, man, sweater	 bush, city, house, roof, sky, valley, view
ESP-GAME	 round, stone, green, sky, grass, man, bunker, building, concrete	 barn, stone, white, cow, dirt, tail, bull, grass, farm, eye, animal, meat	 pink, silver
Base de documents	 nom, adresse, catégorie	 journal, article	 nom, numéro, sexe, taille

FIGURE 2.12 – Exemples d'images issues des bases de test

2.5 Discussion

Nous clôturons ce chapitre par une discussion sur les différentes méthodes d'annotation d'images.

Bien que l'annotation manuelle d'images soit meilleure en matière de précision pour la récupération d'images, car les mots-clés sont choisis par l'utilisateur pour décrire le contenu sémantique des images, c'est un travail intensif et un processus fastidieux. Les méthodes d'annotation automatique nécessitent un grand nombre d'exemples pour effectuer l'apprentissage. Cependant, dans le monde réel, les images annotées disponibles ne sont pas toujours suffisantes. Un bon compromis serait donc de choisir une méthode d'annotation automatique et d'essayer de réduire l'ensemble d'apprentissage. Dans ce mémoire, nous nous plaçons dans l'optique d'annotation automatique.

TABLE 2.2: Synthèse des méthodes d'annotation d'images

Méthodes d'annotation	**Avantages**	**Inconvénients**
Modèles graphiques [RLL+07, LLL+09, ZTH+14, LYRZ10, THY+11, TYH+09, JZHJ14]	– bonne présentation des variables – traitent des caractéristiques de diverses natures – traitent le problème des données manquantes – paramètres faciles à régler	– complexité en temps et en espace
Modèles génératifs [HTO99, DBdFF02, LMJ04, LWL+07, FML04, JLM03, MGP04, BDF+03, NHGT14, YH08]	– établissent la relation entre caractéristiques visuelles et textuelles – efficaces avec un grand vocabulaire	– coûteux – nécessitent des hypothèses simplificatrices – nombre important de paramètres à estimer
Modèles discriminatifs [ALLQ13, MCM14, WYM+16, MSCM16, FZQ12, FM13, VJ12, KIS14, BBUDB15, TF17, JWL+16]	– forte puissance de classification et de discrimination	– perturbation avec un grand nombre de classes

Méthodes d'annotation	**Avantages**	**Inconvénients**
SVM [CCS04, CHV99, SFCL04, ALLQ13, GCL05, QH07, MCM14]	– performants même en présence d'un grand nombre de variables – fonctionnent avec un faible nombre d'exemples d'apprentissage	– mal adaptés aux problèmes avec données manquantes
RNA [PLK04, KPK04, KH02, CFCF10, CFCF12, SPS13, OMF12, ZZZP08, DFPSS07, CLBS11, WYM$^+$16, MMM15, MSCM16]	– peuvent apprendre des classes multiples à la fois – bonnes performances avec l'apprentissage profond	– fonctionnent comme une boite noire – choix du nombre de couches – choix du nombre de neurones
AD [MDGW07, MGPW05, LZL08, FZQ12, JHCZ09, FM13, FKQ11]	– robustes et efficaces pour apprendre de grandes quantités de données d'entraînement – temps de formation d'un modèle court	– choix du nombre d'arbres de décision – profondeur de chaque arbre – nécessitent des données discrétisées
KPPV [VJ12, MPK10, GMVS09, LSW09, ZHH$^+$10, KIS14, LLLC12, THY$^+$11, UBBDB13, BBUDB15]	– ne nécessitent pas une phase d'apprentissage – bonnes performances	– écart sémantique – fréquences des étiquettes déséquilibrées
Retour de pertinence [Chi13, YHC12, ZLX10, MVC11]	– améliorent significativement les résultats	– coûteux pour l'utilisateur

Méthodes d'annotation	**Avantages**	**Inconvénients**
Métadonnées [LCZ+06, WZL+10, WJZZ07, JKWA05, WZJM06, ML12]	– utilisent des fonctionnalités textuelles et visuelles	– difficulté de relier les caractéristiques visuelles aux caractéristiques textuelles – difficulté d'extraction des caractéristiques textuelles
Apprentissage semi-supervisé [Csu17, WHS+06, TYH+09, YWSZ13, DTXM09, DTXC09]	– résout le problème de la disponibilité de grandes collections de données étiquetées – établit des modèles plus précis que ceux réalisés par des méthodes d'apprentissage purement supervisées	– réduit la différence de distribution des données entre les domaines auxiliaires et les domaines cibles
Hiérarchies sémantiques [DDS+09, MS07, ATY+97, DGLW07, LS06, LSFF09, WML04, NST+06, KR14]	– utiles lorsque le nombre de catégories augmente – résolvent le problème d'écart sémantique	– difficulté de construction d'une hiérarchie sémantique
Codage creux [TF17, JWL+16, CZG+15, GCTR14, LTCT14, LHWW15]	– bonne représentation des caractéristiques visuelles	– nécessitent les étiquettes des classes

Dans le tableau 2.2, nous avons tenté de faire une synthèse de la majorité des techniques d'annotation automatique d'images existantes dans la littérature. Nous nous sommes concentrés sur les avantages et les inconvénients de chacune de ces techniques. Dans les travaux de l'état de l'art, il existe plusieurs techniques d'annotation automatique, y compris les modèles génératifs, les modèles discriminatifs et les modèles graphiques.

Dans l'ensemble, l'annotation automatique d'images est un domaine de recherche difficile où plusieurs verrous majeurs sont à lever. Le premier est l'extraction de caractéristiques dimensionnelles élevées. La majorité des fonctionnalités existantes ont des limites à la description des images et aucune d'elles n'est suffisamment puissante pour représenter la grande variété

d'images dans la nature. La pratique courante consiste à combiner plusieurs types de fonctionnalités pour représenter autant d'images que possible. Cependant, le traitement et l'analyse des fonctions d'images à haute dimension constitue un problème très complexe. La performance des classificateurs se dégrade considérablement lorsque la dimension des caractéristiques est trop élevée. Pour résoudre ce problème, les réseaux de neurones convolutionnels ont fait leur apparition. En effet, il y a eu une explosion médiatique pour les architectures de réseaux à neurones profonds pour la vision par ordinateur ces dernières années. L'idée de base est de prendre une image brute et effectuer une série de transformations sur l'image. Sur chaque transformation, on apprend une représentation plus dense de l'image. Il s'avère qu'en utilisant ce principe, on apprend de plus en plus de fonctionnalités abstraites qui permettent de mieux représenter l'image originale.

Le deuxième verrou est de savoir comment créer un modèle d'annotation efficace. La plupart des modèles d'annotation automatique d'images existants se basent sur des caractéristiques d'images de bas niveau. Cependant, en raison du grand nombre d'images requises pour construire un modèle d'annotation efficace, le nombre d'images n'est pas assez grand pour former un modèle précis. Par conséquent, des informations textuelles ou des métadonnées devraient être utilisées pour améliorer la précision de l'annotation. Cependant, très souvent, les métadonnées ne sont ni précises ni adéquates. La façon d'intégrer l'information visuelle de faible niveau et l'information textuelle de haut niveau dans un modèle d'annotation cohérent est une question difficile. L'utilisation des hiérarchies sémantiques peut résoudre ce problème.

Le troisième verrou est le manque de vocabulaire standard et de taxonomie pour l'annotation automatiques d'images. Une modélisation hiérarchique du vocabulaire est nécessaire pour annoter correctement les images. Une taxonomie hiérarchique standardise non seulement le vocabulaire d'annotation mais permet aussi une annotation étape par étape qui est plus précise.

Pour le choix d'une méthode d'annotation automatique, nous avons choisi les modèles graphiques probabilistes et en particulier les réseaux bayésiens (chapitre 3). En effet, ces derniers sont une union entre la théorie des probabilités et la théorie des graphes. Ils bénéficient donc des avantages de ces deux théories. Ils utilisent une structure graphique pour fournir une représentation explicite des informations et un calcul probabiliste pour représenter l'incertitude. En outre, un avantage des modèles graphiques probabilistes est de traiter le problème des données manquantes et d'offrir la possibilité de combiner plusieurs sources d'informations. Par contre, les réseaux bayésiens, comme tous les modèles génératifs, nécessitent une grande quantité d'apprentissage. Un autre inconvénient est que les relations entre les termes du vocabulaire d'annotation ne sont pas exploitées.

Pour réduire l'ensemble d'apprentissage et améliorer la qualité d'annotation, nous avons intégré l'utilisateur dans le processus d'annotation (chapitre 4). L'intervention de l'utilisateur est effectué à trois niveaux d'apprentissage. Dans le premier, l'utilisateur améliore la qualité de l'ensemble d'apprentissage. Plus précisément, les images du jeu d'apprentissage sont initialement annotées par un petit ensemble de mots-clés. Ensuite, le modèle est utilisé pour ajouter automatiquement d'autres annotations aux images. L'utilisateur valide les nouvelles étiquettes correctes et corrige certaines fausses étiquettes. Les images avec leurs nouvelles étiquettes sont utilisées pour réapprendre le modèle. Ce processus peut être répété plusieurs fois. De cette façon, nous réduisons l'abondance de l'annotation manuelle. Dans le deuxième niveau, au cours d'une étape de recherche d'images, les données sont introduites progressivement dans le système et appris progressivement. Le système doit pouvoir modifier et ajuster ses paramètres après

l'observation de chaque exemple. Dans le troisième niveau, nous proposons un apprentissage actif de notre modèle pour réduire le nombre d'instances en choisissant l'exemple conduisant à la plus grande réduction du taux d'erreur.

Pour exploiter les relations entre les mots-clés et enrichir le vocabulaire d'annotation par de nouveaux mots, nous avons intégré une taxonomie dans notre modèle (chapitre 5). La construction de la taxonomie se base sur la combinaison de trois caractéristiques. La première est une similarité visuelle qui représente la correspondance visuelle entre les mots-clés du vocabulaire. La deuxième est une similarité conceptuelle qui définit un degré de similarité entre les mots-clés en se basant sur word2vec. La troisième caractéristique est une similarité contextuelle qui mesure la dépendance entre chaque paire de mots-clés dans le vocabulaire. Ensuite, ces caractéristiques sont utilisées avec la méthode de regroupement Kmeans pour rassembler les mots-clés les plus proches sémantiquement dans un seul groupe. Un nouveau mot-clé qui n'appartient pas au vocabulaire est ajouté comme parent de ce groupe. Les nouveaux concepts parents ajoutés sont à leurs tours regroupés jusqu'à atteindre la racine de la taxonomie.

2.6 Conclusion

Dans ce chapitre, nous avons tenté de présenter une étude détaillée de l'état de l'art concernant l'annotation d'images. Nous avons regroupé les méthodes citées en plusieurs catégories tels que les modèles graphiques probabilistes, les modèles génératifs et les modèles discriminatifs. Après cette étude, nous avons présenté les critères d'évaluation des systèmes d'annotation. Ensuite, nous avons présenté les bases de données utilisées dans ce mémoire. Nous avons terminé ce chapitre par une synthèse de toutes les méthodes d'annotation automatique d'images en présentant les avantages et les inconvénients de chacune. Cette synthèse nous a permis d'argumenter notre choix pour les modèles graphiques probabilistes comme méthode d'annotation et l'importance d'intégrer un retour utilisateur et une hiérarchie sémantique dans notre modèle d'annotation automatique. Le chapitre suivant sera consacré à la description de notre modèle d'annotation automatique d'images.

Chapitre 3

Modèle graphique probabiliste pour l'annotation d'images

Sommaire

3.1 Introduction

La première partie de ce chapitre est consacrée aux modèles graphiques probabilistes et en particulier les réseaux bayésiens. La deuxième partie est dédiée aux caractéristiques visuelles

utilisées dans cette thèse. Ensuite, nous présentons le modèle graphique probabiliste que nous avons proposé pour l'annotation d'images. Nous détaillons la structure, les paramètres et comment utiliser l'inférence dans notre modèle pour la classification et l'extension d'annotation d'images. Nous présentons nos expérimentations sur trois bases d'images. Enfin, la dernière partie de ce chapitre est consacrée à l'évaluation d'annotation sur une base de documents.

3.2 Modèles graphiques et réseaux bayésiens

Les modèles graphiques probabilistes utilisent les graphes pour représenter facilement l'information et modéliser des problèmes complexes. Ils possèdent un raisonnement probabiliste pour représenter l'incertitude. Les réseaux bayésiens sont un cas particulier des modèles graphiques probabilistes utilisés pour la modélisation de la distribution de probabilité conjointe des variables du problème à résoudre.

3.2.1 Modèles graphiques probabilistes

Les modèles graphiques probabilistes [Pea14] sont une union entre la théorie des probabilités et la théorie des graphes. Ce sont des modèles probabilistes pour la représentation des connaissances fondés sur une description graphique des variables aléatoires représentées par les nœuds du graphe et des dépendances entre les variables représentées par les arcs entre les nœuds du graphe. Dans un modèle graphique [Jen96], on trouve généralement deux composantes. Une composante graphique sous la forme d'un graphe orienté acyclique (DAG) ou non orienté. Elle met en valeur les variables du problème à modéliser et les relations d'influence (dépendance conditionnelle, causalité, etc.). Une composante numérique qui sert à quantifier les relations d'influence dans le modèle. Ces modèles décrivent graphiquement les dépendances conditionnelles entre les variables.

3.2.2 Réseaux bayésiens

Un réseau bayésien [NWL$^+$11] est un modèle graphique probabiliste. Il permet de représenter un ensemble de variables aléatoires nœuds liés par des arcs orientés. Ces derniers représentent les dépendances conditionnelles entre les variables. Un réseau bayésien se base sur les dépendances conditionnelles pour la simplification du théorème de Bayes généralisé. Il est défini par une description graphique des variables et par une description probabiliste des dépendances conditionnelles entre ces variables.

- **Description graphique :** sous forme d'un graphe acyclique dirigé $G(V, E)$ où V est l'ensemble des nœuds représentant les variables aléatoires à modéliser, et E représente l'ensemble des arcs entre ces nœuds. Un arc entre deux nœuds signifie qu'il y a une relation de dépendance entre les deux variables représentées par ces deux nœuds. La structure du graphe peut être donnée directement par un expert ou par un apprentissage automatique.
- **Description probabiliste :** ensemble de tables de probabilités de chaque variable. Ces tables peuvent être données par un expert du domaine ou par un apprentissage automatique.

Formellement, on peut définir un réseau bayésien par un :
- graphe acyclique orienté $G = (V, E)$, où V représente l'ensemble des nœuds du graphe G, et E représente l'ensemble des arcs entre les nœuds ;
- espace probabiliste fini (Ω, Z, p) ;
- ensemble de variables aléatoires modélisées par les nœuds du graphe et définies sur (Ω, Z, p), tel que :

$$p(V_1, V_2, ..., V_n) = \prod_{i=1}^{n} p(V_i | C(V_i)) \tag{3.1}$$

$C(V_i)$ est l'ensemble des causes (parents) de la variable V_i dans le graphe G.

À travers sa représentation graphique des connaissances d'un système [Fra06], un réseau bayésien permet de :
- prévoir, contrôler et simuler un comportement,
- aider à l'analyse des données,
- prendre des décisions.

De ce fait, les réseaux bayésiens ont été employés dans plusieurs applications du monde réel. Pour définir un réseau bayésien, il faut définir sa structure et ses paramètres.

3.2.3 Apprentissage de structure

Le nombre de toutes les structures possibles à partir de n nœuds est super-exponentiel [Rob77]. La formule suivante permet de calculer ce nombre :

$$NS(n) = \begin{cases} 1 & \text{, si } n = 0 \ \ ou \ \ 1 \\ \sum_{i=1} n(-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i) & \text{, sinon} \end{cases} \tag{3.2}$$

(avec 5 nœuds, on obtient $NS(5) = 29281$ et avec 10 nœuds, on obtient $NS(10) = 4.2 \times 10^{18}$).

Le parcours de toutes ces structures est impossible. Pour cela, plusieurs algorithmes d'apprentissage de structures ont été proposés. On distingue les algorithmes d'apprentissage de la causalité et les algorithmes à base de score.

3.2.3.1 Apprentissage de la causalité

Les étapes des algorithmes d'apprentissage de la causalité sont :
- construire un graphe non dirigé représentant les relations entre les variables par ajout ou suppression d'arêtes ;
- détecter les structures de forme V dans ce graphe ;
- orienter les arcs entre les nœuds.

Les algorithmes d'apprentissage de la causalité se basent sur les données pour obtenir les relations de dépendance conditionnelle entre les variables. Les algorithmes IC [Gil01], PC [SGS00] et BN-PC [CBL97] sont des exemples de cette famille.

3.2.3.2 Utilisation d'un score

Les algorithmes à base de score cherchent la structure maximisant une fonction de score. Pour faciliter l'application de ces approches à base de score, il faut utiliser un score décomposable, c'est-à-dire qui peut s'écrire sous forme d'une somme de scores locaux au niveau de

chaque nœud. Les algorithmes GS [CGH95], K2 [CH92] et SEM [Fri98] sont des exemples de cette famille.

3.2.4 Apprentissage des paramètres

Après la définition de la structure du réseau bayésien, il faut trouver les tables de probabilités conditionnelles associées aux variables. Ces tables peuvent être données directement par un expert ou extraites de données. Dans ce dernier cas, deux possibilités se présentent : les données sont complètes ou incomplètes.

3.2.4.1 Données complètes

L'objectif de l'apprentissage des paramètres avec des données complètes est d'estimer les distributions de probabilités (ou les paramètres des lois correspondantes) à partir de données disponibles. On distingue principalement deux méthodes : la méthode statistique et la méthode bayésienne.

3.2.4.1.1 Apprentissage statistique

La méthode la plus utilisée dans le cas des données complètes est l'estimation statistique qui consiste à utiliser la fréquence d'apparition d'un événement pour estimer sa probabilité. Cette méthode est appelée maximum de vraisemblance. Elle permet de calculer la probabilité suivante :

$$P(X_i = X_k | pa(X_i) = X_j) = \theta_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \tag{3.3}$$

Où $N_{i,j,k}$ est le nombre d'échantillons d'apprentissage pour lesquels la valeur de X_i est X_k et la valeur de ses parents est X_j.

3.2.4.1.2 Apprentissage bayésien

L'estimation bayésienne consiste à chercher les paramètres θ les plus probables sachant l'observation des données. Elle permet de calculer la probabilité *a posteriori* en fonction de la probabilité *a priori*. L'apprentissage bayésien est donné par :

$$P(\theta|D) \propto P(D|\theta).P(\theta) \tag{3.4}$$

La fréquence d'une variable sachant ses parents peut être nulle. La probabilité *a priori* est donc multinomiale. La distribution conjuguée qui permet de résoudre le problème de division par zéro est la distribution de Dirichlet :

$$P(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{i,j,k}^{\alpha_{i,j,k}-1} \tag{3.5}$$

$\alpha_{i,j,k}$: coefficients de Dirichlet.
n : nombre de variables utilisées dans le réseau.
r_i : les états possibles d'un nœud.
q_i : les configurations possibles des parents d'un nœud.

La distribution de Dirichlet permet de déterminer la loi des paramètres $P(\theta|D)$:

$$P(\theta|D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{i,j,k}^{N_{i,j,k}+\alpha_{i,j,k}-1} \tag{3.6}$$

L'approche bayésienne MAP (Maximum à posteriori) donne alors :

$$P(X_i = X_k|pa(X_i) = X_j) = \theta_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k} + \alpha_{i,j,k} - 1)} \tag{3.7}$$

Une autre estimation bayésienne EAP (espérance a posteriori) calcule l'espérance des paramètres $\theta_{i,j,k}$.

$$P(X_i = X_k|pa(X_i) = X_j) = \theta_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k}}{\sum_k (N_{i,j,k} + \alpha_{i,j,k})} \tag{3.8}$$

Ces approches ne sont utilisées que si les données sont complètes, ce qui n'est pas toujours vrai dans les applications du monde réel, d'où la nécessité de traiter le cas des données manquantes.

3.2.4.2 Données manquantes

Si les données d'apprentissage sont complètes, l'apprentissage des paramètres d'un réseau bayésien n'est pas difficile. Cependant, dans le monde réel, les données d'apprentissage peuvent être incomplètes pour diverses raisons. Plusieurs méthodes d'apprentissage des paramètres dans les réseaux bayésiens traitant le cas des données manquantes ont été proposées. La méthode la plus utilisée est l'algorithme Estimation-Maximisation (EM) [DLR77].

3.2.4.2.1 L'algorithme EM

EM est une méthode qui permet de déterminer les paramètres d'un réseau bayésien à partir de données manquantes. Elle est divisée en deux étapes. Dans la première étape d'estimation (E), on calcule l'espérance de la vraisemblance (ou log de la vraisemblance) des données sachant les dernières variables observées. Dans la deuxième étape de maximisation (M), on calcule le maximum de vraisemblance des paramètres en maximisant la valeur trouvée à l'étape E. Les paramètres trouvés en M sont utilisés de nouveau pour une nouvelle phase d'évaluation de l'espérance. La performance de cet algorithme dépend de la taille des données d'apprentissage et du nombre de données manquantes. Le problème majeur de la méthode EM est les maximums locaux. Pour cela, d'autres extensions de cette méthode ont été proposées comme les méthodes GEM [DLR77], CEM [CG92] et SEM [CD85]. L'algorithme 1 décrit les itérations de EM.

Algorithm 1 : Algorithme EM

Entrées : Graphe G, ensemble de données D
Sorties : Paramètres θ

1 : Initialiser les paramètres θ_0 aléatoirement ou manuellement ;
Répéter
2 : E : calculer l'espérance de la log-vraisemblance

$$Q(\theta;\tilde{\theta}) = E_{\tilde{\theta}}(logP(D|\theta)); \tag{3.9}$$

3 : M : choisir les meilleures valeurs des paramètres en maximisant Q

$$\tilde{\theta} = \arg\max_{\theta} \ Q(\theta;\tilde{\theta}); \tag{3.10}$$

Jusqu'à convergence
La converge peut être vérifiée si les paramètres ne sont pas modifiés ou si le nombre maximum d'itérations est atteint.

3.2.5 Inférence

L'utilisation essentielle des réseaux bayésiens est le calcul des probabilités conditionnelles des événements. Cette utilisation s'appelle inférence. Vu la correspondance entre la structure graphique et la structure probabiliste, les problèmes de l'inférence sont ramenés à des problèmes de la théorie des graphes. On l'utilise pour calculer des probabilités marginales sur les nœuds en l'absence ou la présence de variables observées (évidence). D'un point de vue mathématique, il s'agit de calculer $P(X/e)$. X désigne l'ensemble de variables aléatoires et e une instanciation possible d'une partie ou de tout l'ensemble de X. Plusieurs approches ont été développées pour résoudre le problème de l'inférence et on distingue deux classes d'algorithmes d'inférence : les algorithmes exacts et les algorithmes approchés. Les méthodes d'inférence exactes exploitent les indépendances conditionnelles contenues dans les réseaux et donnent à chaque inférence les probabilités a posteriori exactes. Si la dimension des réseaux est grande, le temps de calcul est de plus en plus important. Dans ce cas, on utilise des approches approximatives.

L'algorithme d'inférence le plus connu et le plus utilisé dans les réseaux bayésiens est l'arbre de jonction.

3.2.5.1 Arbre de jonction

La méthode d'arbre de jonction (clique-tree propagation algorithm) a été proposée par Lauritzen [LS88, LJ97]. Elle est utilisable pour toutes les types de structures. Les étapes de cette méthode sont les suivantes :

- moralisation : construire un graphe non dirigé (graphe moral),
- triangulation : ajouter des arcs au graphe moral précédent pour obtenir un graphe triangulé,
- construction de l'arbre de jonction : connecter les cliques [11] du graphe triangulé précédent.

11. **Définition :** une clique est un sous graphe complet maximal : – *complet :* chaque sommet est adjacent à tous les autres. – *maximal :* non contenu dans un sous graphe complet.

Toutes les cliques sur le chemin entre X et Y doivent contenir $X \cap Y$,
- propagation : injecter les informations dans le réseau pour calculer l'inférence,
- mise à jour des distributions de probabilité.

La complexité de l'inférence est NP-difficile [CH92]. Elle conduit à un temps de calcul énorme dans le cas des réseaux bayésiens plus complexe c'est pourquoi on utilise des approches approximatives.

3.3 Caractéristiques visuelles

L'extraction des caractéristiques visuelles (niveau visuel ou physique) est une transformation mathématique calculée sur les pixels d'une image numérique. Les caractéristiques visuelles permettent généralement de trouver certaines propriétés visuelles de l'image, on parle de descripteurs de bas-niveau. On peut faire une description locale ou globale d'une image.

3.3.1 Histogramme de couleurs RVB

Un descripteur de couleur permet de donner une description des couleurs d'une image dans un espace de couleur (RGB, HSV, LAB). Il faut tout d'abord choisir l'espace de couleurs et ensuite choisir la manière de décrire les couleurs. Un histogramme de couleurs [SB91] donne le nombre de pixels de chaque couleur contenu dans l'image. On quantifie d'abord l'espace de couleurs en n couleurs possibles, puis on normalise l'histogramme par la taille de l'image. Si une image I est de taille $M \times N$, son histogramme H est donné par l'équation suivante :

$$H(c) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} c(I(i,j)) \tag{3.11}$$

Dans la figure 3.1, on représente trois histogrammes d'une image de la base Corel5k. Le premier histogramme (b) représente la répartition des couleurs de la composante rouge de l'image. Le deuxième histogramme (c) représente la répartition des couleurs de la composante verte de l'image. Le troisième histogramme (d) représente la répartition des couleurs de la composante bleu de l'image.

3.3.2 SIFT

La méthode SIFT [Low99] permet de détecter et identifier des éléments similaires entre différentes images (paysages, objets, personnes,..). Les descripteurs SIFT sont des informations numériques qui permettent de caractériser le contenu visuel de cette image indépendamment de l'échelle, de l'angle d'observation et de la luminosité. Pour calculer ces descripteurs, il faut :
- détecter les points-clé (points d'intérêts) ;
- considérer une région autour du point-clé de 16×16 pixels ;
- subdiviser cette région en 16 zones de 4×4 pixels ;
- calculer l'histogramme des orientations sur chaque zone ;
- fournir le descripteur SIFT de dimension 128 pour chaque point-clé.

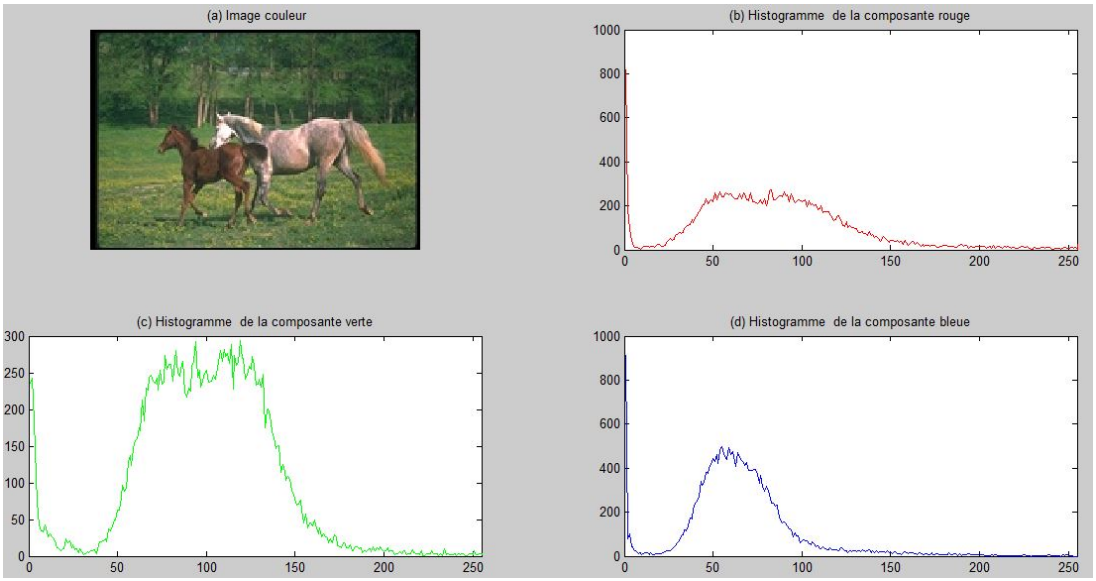


FIGURE 3.1 – Histogramme des 3 composantes d'une image couleur

Dans la figure 3.2, on représente la détection de quelques points d'intérêt avec la méthode SIFT d'une image [12].

FIGURE 3.2 – Détection des points d'intérêt avec SIFT

12. Source : https ://fr.wikipedia.org/wiki/Scale-invariant_feature_transform

3.3.3 GIST

Le descripteur GIST a été proposé par Oliva et Torralba [OT01] qui basent leurs travaux sur la manière dont la vision humaine perçoit la structure générale d'une scène. GIST permet de capturer de manière précise la forme globale d'une image en caractérisant l'orientation des différents contours qui y apparaissent. Il permet de récupérer les fréquences et orientations principales contenues dans une image en utilisant plusieurs filtres de Gabor. Les étapes de calcul de ce descripteur sont :

- réduire la taille de l'image en une image de taille fixe de m x m pixels ;
- diviser l'image obtenue en 16 zones ;
- calculer un descripteur basé sur les histogrammes d'orientation du gradient sur chaque zone.

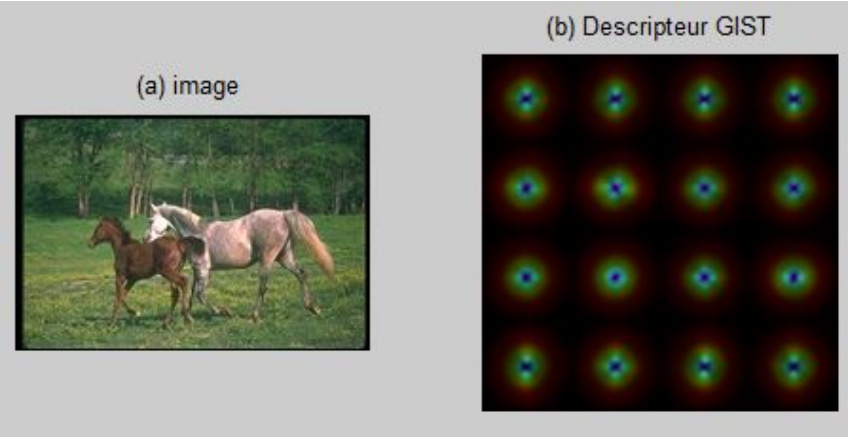


FIGURE 3.3 – Visualisation du descripteur GIST

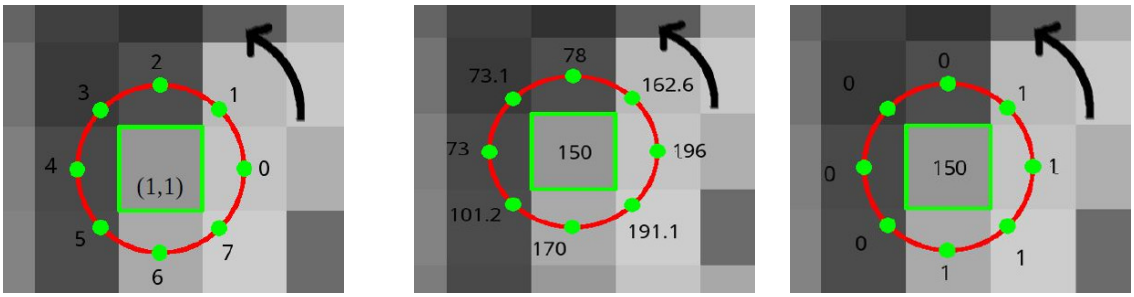
Dans la figure 3.3, on représente l'illustration du descripteur GIST sur une image de la base Corel-5K. L'image de gauche est divisée en 16 zones. Le spectre d'énergie locale est calculé sur chaque zone. Ces spectres sont représentés dans l'image de droite.

3.3.4 LBP

Le descripteur LBP (Local Binary Pattern) [OPH96] est utilisé en reconnaissance de formes pour reconnaître des textures. Il a été utilisé en 1993 pour mesurer le contraste local d'une image puis en 1995 pour analyser les textures.

Dans l'exemple de la figure 3.4, le LBP est calculé pour le pixel de coordonnées (1,1). Une orientation anti-horaire et un point de départ en 0 (image (a) de la figure 3.4) sont choisis. L'image sur laquelle est calculée le LBP doit être en niveaux de gris. Pour chaque pixel de l'image, on définit un voisinage de rayon r (en rouge). Sur ce voisinage, on choisit p points (ici 8). Pour chaque point, on récupère la couleur du pixel (image (b) de la figure 3.4). Si les coordonnées du point ne correspondent pas à un pixel de l'image, on récupère la couleur à ces coordonnées par interpolation bilinéaire des couleurs. Ensuite, pour chaque point, on compare la couleur de ce point avec le point central. Si la couleur au point est supérieure ou égale à la couleur au centre (150 sur l'exemple de la figure 3.4) alors la valeur 1 sera assignée à ce point

et 0 sinon. On convertit le nombre motif binaire 11000011 obtenu (image (c) de la figure 3.4) en décimale et on obtient 195.



(a) Orientation du calcul de LBP (b) Couleurs interpolées (c) Motif binaire 11000011

FIGURE 3.4 – Calcul du descripteur LBP

3.4 Modèle proposé

Comme mentionné dans la section 2.5 du chapitre 2, nous avons fait notre choix pour les modèles graphiques probabilistes et plus précisément les réseaux bayésiens. En effet, ces derniers sont des graphes acycliques dirigés où les nœuds représentent des variables aléatoires et les nœuds correspondent aux relations entre les variables aléatoires. Les données quantitatives, les variables latentes et les paramètres inconnus peuvent être inclus dans le même modèle. Les réseaux bayésiens offrent donc un avantage important pour la construction de modèles en utilisant des bases théoriques solides adaptés à un raisonnement incertain. Notre modèle est un mélange de distributions multinomiales et de mélanges de Gaussiennes. Nous supposons que les caractéristiques visuelles sont considérées comme des variables continues. Elles suivent une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Les caractéristiques textuelles (mots-clés) sont considérées comme des variables discrètes. Elles suivent une distribution multinomiale. Nous détaillons dans les sections qui suivent la structure, les paramètres et l'utilisation de notre modèle pour l'annotation automatique d'images.

Définition : Une variable aléatoire X suit une distribution Gaussienne (ou normale) avec une moyenne μ et une variance σ^2, si sa fonction de densité de probabilité (fdp) est :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.12}$$

La distribution Gaussienne de X est habituellement représentée par : $X \sim \mathcal{N}(\mu, \sigma^2)$. La courbe représentative (figure 3.5) de la densité Gaussienne est appelée courbe en cloche. Un modèle de mélanges de Gaussiennes (GMM) est une distribution de probabilité qui consiste en un mélange d'un ou plusieurs composantes de distributions Gaussiennes. Toute densité continue peut être approchée arbitrairement par un GMM en utilisant un nombre suffisant de Gaussiennes, et en ajustant leurs moyennes, leurs covariances et les poids. Un mélange de Gaussiennes peut être écrit comme suit :

$$f_X(x) = \sum_{j=1}^{K} w_j \mathcal{N}(x|\mu_j, \sum_j) \tag{3.13}$$

Où w_j est le poids de la composante Gaussienne j. μ_j et $\sum_j$ représentent respectivement la moyenne et la covariance de la composante Gaussienne j.

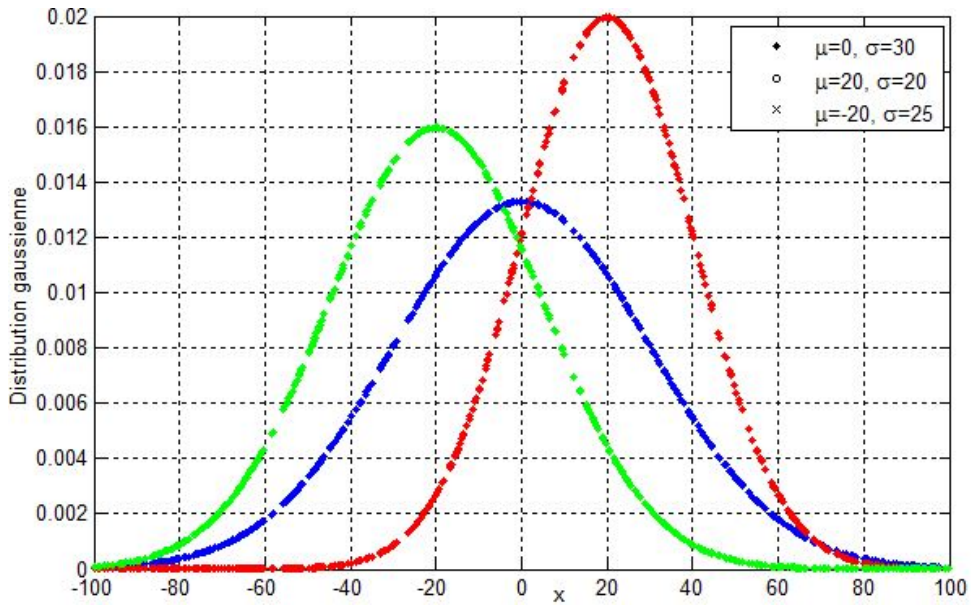


FIGURE 3.5 – Exemple de trois distributions Gaussiennes

3.4.1 Structure

La structure de notre modèle, comme illustré dans la figure 3.6, est divisée en trois parties :

1. La première partie est constituée d'un seul nœud racine *Classe* qui permet de représenter la classe d'une image donnée. Ce nœud est modélisé par une variable aléatoire discrète qui peut prendre k valeurs correspondant aux classes prédéfinies $C_1, ..., C_k$.

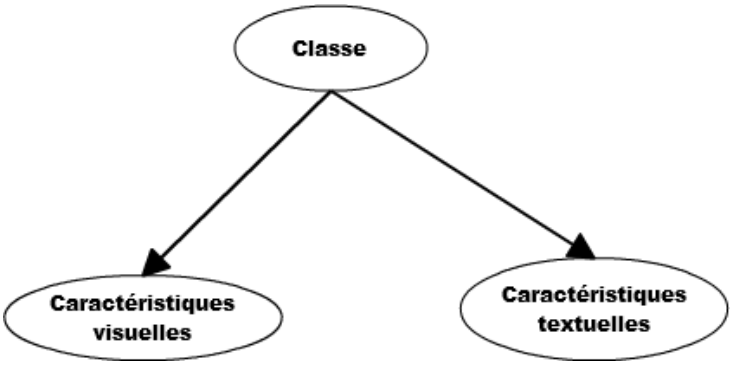


FIGURE 3.6 – Structure globale du modèle

2. La deuxième partie (figure 3.7) est utilisée pour représenter les caractéristiques visuelles d'une image. Nous supposons que ces caractéristiques sont considérées comme des variables continues. Elles suivent une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Les caractéristiques visuelles sont modélisées par deux nœuds :
   – Le nœud *Gaussienne* est modélisé par une variable aléatoire continue qui est utilisée pour représenter les descripteurs calculés sur l'image.
   – Le nœud *Composante* est modélisé par une variable aléatoire cachée qui est utilisée pour représenter le poids des Gaussiennes utilisées. Il peut prendre g valeurs correspondant au nombre de Gaussiennes utilisées dans le mélange.

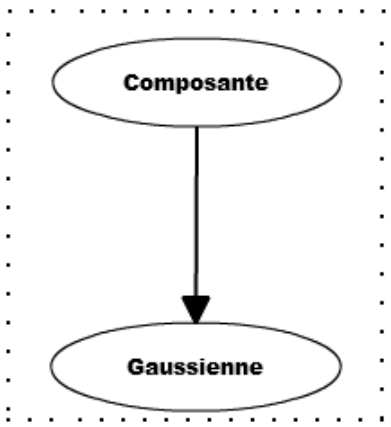


FIGURE 3.7 – Caractéristiques visuelles

3. La troisième partie (figure 3.8) est utilisée pour représenter les caractéristiques textuelles. Elle est modélisée par N nœuds discrets, où N est le nombre maximum de mots-clés utilisés pour annoter une image. Des arcs sont ajoutés entre les N nœuds pour représenter les dépendances conditionnelles entre les mots-clés.

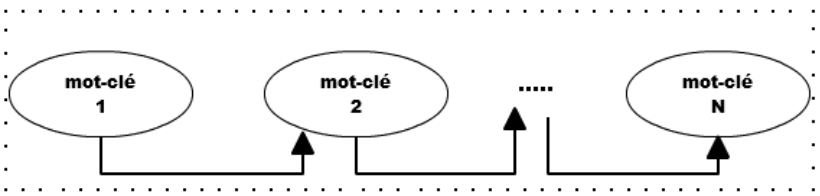


FIGURE 3.8 – Caractéristiques textuelles

La concaténation des trois sous-structures nous donne la structure détaillée de notre modèle illustrée dans la figure 3.9.

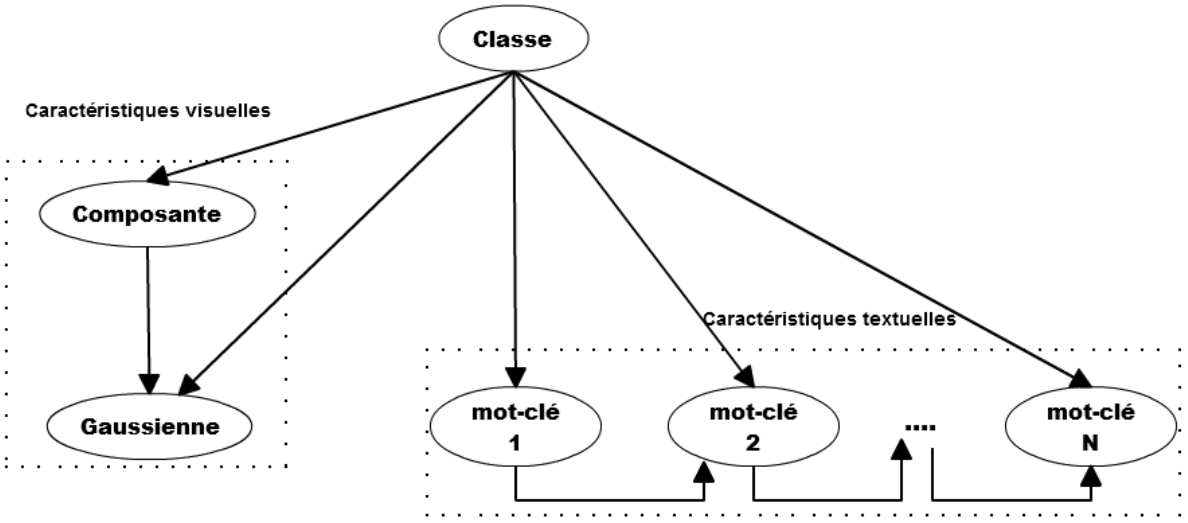


FIGURE 3.9 – Structure détaillée du modèle

3.4.2 Paramètres

Pour le réseau décrit dans la section précédente, on peut écrire la probabilité jointe :

$$P(C, TC, VC) = P(C) \prod_{i=1}^{M} P(VC_i|C) \prod_{i=1}^{N} P(TC_i|C) \tag{3.14}$$

où TC (T pour Textual et C pour Characteristics) représente les caractéristiques textuelles (les mots-clés $Kw_1, ..., Kw_N$) et VC (V pour Visual et C pour Characteristics) représente les caractéristiques visuelles (les caractéristiques de bas niveau $VC_1, ..., VC_M$).

Soit $Kw_1, ..., Kw_N$ l'ensemble des mots-clés dans une image. Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ peut être représentée par un espace vectoriel booléen des mots du vocabulaire :

$Kw_j = \{m_1, ..., m_n\}$, où $m_i = 0$ ou $1, \forall i \in \{1, ..., n\}$ et $\sum_{i=1}^{n} m_i = k$.

Chaque variable $Kw_j, \forall j \in \{1, ..., N\}$ suit une distribution multinomiale avec les paramètres $\Phi_{TC} = (k, p_1, ..., p_n)$, où p_i est la probabilité associée à chaque valeur m :

$$p(m_1 = p_1, ..., m_n = p_n) = \frac{k!}{m_1!m_2!...m_n!} p_1^{m_1} p_2^{m_2} ... p_n^{m_n} \tag{3.15}$$

Soit I un ensemble de m images $(Im_1, ..., Im_m)$ et g groupes $(G_1, ..., G_g)$ dont chacun a une densité Gaussienne avec une moyenne $\mu_l, \forall l \in \{1, ..., g\}$ et une matrice de covariance $\sum_l$.

Soit $\pi_1, ..., \pi_g$ les proportions des différents groupes, on note par $\theta_k = (\mu_k, \sum_k)$ le paramètre de chaque Gaussienne, et $\Phi_{VC} = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ le paramètre global du mélange. Alors, la densité de probabilité de I conditionnellement à la classe $c_i, \forall i \in \{1, ..., k\}$ est définie par :

$$P(im, \Phi_{VC}) = \sum_{l=1}^{g} \pi_l p(im, \theta_l) \tag{3.16}$$

où $p(im, \theta_l)$ est la Gaussienne multivariée définie par le paramètre θ_l. On note par Φ le paramètre global de ce modèle :

$$\Phi = (\Phi_{VC}, \Phi_{TC}) = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g, k, p_1, ..., p_n) \tag{3.17}$$

L'équation 3.14 peut être réécrite :

$$P(C, TC, VC) = P(C) f(\pi_1, .., \pi_g, \theta_1, .., \theta_g, k, p_1, .., p_n) \tag{3.18}$$

Les paramètres du modèle peuvent être appris d'un ensemble d'images d'apprentissage pour estimer la probabilité jointe de chaque image et chaque classe. Nous devons maximiser la log-vraisemblance L_I de I :

$$\begin{aligned} L_I &= log(P(C, TC, VC) \\ &= \sum logP(C) + \sum logf(\pi_1, ..., \pi_g, \theta_1, ..., \theta_g) + \sum logf(k, p_1, ..., p_n) \end{aligned} \tag{3.19}$$

3.4.3 Annotation et classification

Pour annoter ou classifier une image requête, l'algorithme d'inférence arbre de jonction [LS88] est utilisé. Cet algorithme nous permet de calculer les probabilités conditionnelles des mots-clés du vocabulaire sachant les caractéristiques visuelles et textuelles de l'image requête. Pour une image Im_i partiellement annotée, représentée par ses caractéristiques visuelles $VC_1, ..., VC_M$ et ses mots-clés existants $Kw_1, ..., Kw_n$, nous pouvons utiliser l'inférence bayésienne pour étendre l'annotation de cette image avec d'autres mots-clés. Les caractéristiques observées de l'image sont considérées comme une «évidence» définie par :

$$P(I_i) = P(VC_1, ..., VC_M, Kw_1, ..., Kw_n) = 1 \tag{3.20}$$

Nous pouvons donc calculer la probabilité *a posteriori*

$$P(Kw_i|I_i) = P(Kw_i|VC_1, ..., VC_M, Kw_1, ..., Kw_n) \; \forall i \in \{1, ..., N\} \tag{3.21}$$

où N est la taille du vocabulaire utilisé. Le mot-clé ayant la probabilité maximale sera retenu comme une nouvelle annotation de l'image. Nous pouvons également calculer la probabilité *a posteriori*

$$P(C_i|I_i) = P(C_i|VC_1, ..., VC_M, Kw_1, ..., Kw_n) \tag{3.22}$$

dans le but d'identifier la classe de l'image. L'image requête est affectée à la classe C_i maximisant la probabilité définie dans l'équation 3.22.

3.5 Évaluations sur les bases d'images

Dans cette section, nous proposons d'évaluer notre modèle d'annotations sur trois bases d'images. Pour évaluer notre modèle d'annotation d'images, nous utilisons les quatre mesures d'évaluation standard utilisées dans l'annotation d'images. Nous annotons automatiquement

par notre modèle chaque image dans la base de test par 5 mots-clés et nous calculons le rappel (R), la précision (P), la mesure F1 et la mesure N+. Concernant les caractéristiques visuelles, nous avons utilisé :

– L'histograme des couleurs RGB ;
– Le descripteur LBP ;
– Le descripteur GIST ;
– Le descripteur SIFT.

Le tableau 3.1 montre les performances de différents systèmes d'annotation d'images y compris le notre sur les trois ensembles de données Corel-5k, ESP-Game et IAPRTC-12. Les lignes de ce tableau sont regroupées en fonction de l'approche utilisée par les méthodes. Le premier groupe de lignes contient des méthodes basées sur des modèles de pertinence. Le deuxième groupe contient des méthodes utilisant des algorithmes à base des plus proches voisins. Le troisième groupe contient des systèmes utilisant des représentations profondes d'images basées sur les CNN. Le quatrième groupe montre les performances de certaines méthodes à base du codage creux. Le cinquième groupe contient une variété d'approches comme les forêts aléatoires. Le dernier groupe montre les performances des méthodes proches de notre modèle et qui utilisent des modèles graphiques probabilistes. La dernière ligne montre les résultats de notre méthode. Dans le tableau 3.1, P représente la précision moyenne, R le rappel moyen, F1 la moyenne harmonique entre le rappel et la précision, et N+ le nombre de mots-clés qui ont été correctement assignés aux images de test.

Amara et Hassan ont montré dans [TF17] que pour un système d'annotation donné :

– La précision pour un mot-clé donné est directement proportionnelle à sa fréquence, c'est-à-dire que les mots-clés très fréquents ont tendance à obtenir de meilleurs scores de précision.
– Le rappel pour un mot-clé donné est inversement proportionnel à la taille du vocabulaire d'annotation, c'est-à-dire qu'un grand nombre de mots-clés disponibles entraîne des scores de rappel inférieurs.

Ceci explique le comportement des différents systèmes d'annotation présentés dans le tableau 3.1. La plupart de ces systèmes atteignent un score de rappel beaucoup plus faible que le score de précision pour les deux bases ESP-Game et IAPRTC-12 et l'inverse sur la base Corel-5K. Notre méthode fournit des résultats concurrentiels pour les trois ensembles de données, tout en maintenant à la fois une précision élevée et un rappel élevé de manière équilibrée. Cela indique que notre système maintient ses caractéristiques conçues pour une grande variété d'ensembles de données. Notre méthode présente de bons résultats pour le critère N+, ce qui signifie que la quantité du vocabulaire qui a été couvert par la méthode est importante. Nous pouvons donc conclure qu'elle ne pénalise pas les étiquettes qui ne sont présentes que dans quelques images d'apprentissage.

Notre méthode d'annotation surpasse toutes les méthodes du premier et cinquième groupe à part la dernière méthode [MCM14] qui est une combinaison d'un modèle de retour de pertinence et d'une méthode utilisant le classificateur SVM, ce qui la rend complexe. Elle est compétitive par rapport aux méthodes du deuxième groupe qui utilisent les KPPV. Ces méthodes présentent l'inconvénient du grand temps d'annotation. En effet, chaque image à annoter doit être comparée à toutes les images de la base. Au contraire, pour notre méthode, l'apprentissage est effectué une fois pour toutes, et pour annoter une image, nous calculons seulement la probabilité présentée dans l'équation 3.21. En outre, ces méthodes souffrent du problème du choix du nombre

TABLE 3.1 – Performance de notre modèle par rapport aux autres modèles d'annotation de l'état de l'art

Méthode	Corel-5K				ESP-Game				IAPRTC-12			
	P	R	$F1$	$N+$	P	R	$F1$	$N+$	P	R	$F1$	$N+$
CRM [LMJ04]	16	19	17	107	–	–	–	–	–	–	–	–
MBRM [FML04]	24	25	25	122	18	19	18	209	24	23	23	223
SKL-CRM [ML14]	39	46	42	184	41	26	32	248	47	32	38	274
SVM-DMBRM [MCM14]	36	48	41	197	55	25	34	**259**	**56**	29	38	**283**
JEC [MPK10]	27	32	29	139	22	25	23	224	28	29	28	250
TagProp [GMVS09]	33	42	37	160	39	27	32	239	46	35	40	266
NMF-KNN [KIS14]	38	**56**	45	150	33	26	29	238	–	–	–	–
2PKNN [VJ12]	44	46	45	191	53	27	36	252	54	37	44	278
CNN-R [MMM15]	32	41	37	166	44	28	34	248	49	31	38	272
HHD [MSCM16]	31	49	38	194	35	36	34	257	32	44	36	280
MultiSC-AIA [TF17]	–	–	–	–	39	**38**	39	237	41	42	42	250
GS [ZHH$^+$10]	30	33	31	146	–	–	–	–	32	29	30	252
MSC [WYZZ09]	25	32	28	136	35	23	28	236	36	31	33	262
MLDL [JWL$^+$16]	**45**	49	**47**	**198**	**56**	31	**40**	**259**	**56**	40	47	282
SLED [CZG$^+$15]	35	51	42	196	49	30	37	253	50	**47**	**48**	282
SML [CCMV07]	23	29	26	137	–	–	–	–	–	–	–	–
RF [FZQ12]	29	40	34	157	41	26	32	235	44	31	36	253
RFC-PSO[EBKHS14]	26	22	24	109	–	–	–	–	–	–	–	–
Fuzzy[MY16]	27	32	29	–	–	–	–	–	30	27	28	–
AP[RLF12]	–	–	–	–	24	24	24	–	28	26	27	–
Corr-LDA [CBL09]	21	36	27	131	24	23	23	219	27	25	26	223
GMM-Mult [BT08]	27	38	32	154	29	27	28	234	30	26	28	236
Notre méthode	34	45	39	175	31	36	33	251	34	39	36	274

de voisins et la distance à utiliser entre les caractéristiques visuelles. Bien que les méthodes du troisième groupes utilisant l'apprentissage profond offrent de bonnes performances et réduisent le besoin d'ingénierie des fonctionnalités, ces algorithmes nécessitent une grande quantité de données en phase d'apprentissage et nécessitent plus de puissance de calcul et de stockage.

Par rapport aux méthodes citées dans le tableau 3.1, à l'exception du dernier groupe, notre méthode présente l'avantage d'être utilisée pour les deux tâches d'annotation et de classification des images. Un autre avantage de notre modèle repose sur l'interprétation de la structure du réseau qui fournit des informations précieuses sur la dépendance conditionnelle entre les

variables.

Par rapport aux deux modèles [CBL09] et [BT08] du dernier groupe, nous observons que les performances de notre modèle sont meilleures que ces deux modèles. La supériorité par rapport au modèle Corr-LDA est justifiée par le fait que nous utilisons un mélange de Gaussiennes multivariées alors que ce modèle utilise une Gaussienne multivariée. La supériorité par rapport au modèle GMM-Mult est justifiée par l'addition des relations sémantiques entre les mots-clés et l'utilisation des caractéristiques visuelles plus pertinentes.

		
sky, sun, clouds, tree	sky, jet, plane	bear, polar, snow, tundra
sky, **sun**, **clouds**, **tree**, **palm**	**sky**, **jet**, **plane**, **f-16**, **kit**	**bear**, **polar**, **snow**, **ice**, **slope**
		
water, boats, bridge	tree, horses, mare, foals	sky, buildings, flag
water, **boats**, **bridge**, **arch**, **pyramid**	**tree**, **horses**, **mare**, **foals**, **field**	**sky**, **buildings**, **skyline**, **street**, **outside**
		
flowers, house, garden	plants	mountain, sky, water, tree
flowers, **house**, **garden**, **lawn**, **reptile**	**plants**, **leaf**, **stems**, **interior**, **zebra**	**mountain**, **sky**, **clouds**, **tree**, **whales**

FIGURE 3.10 – Exemples d'annotation d'images de la base Corel-5k

La figure 3.10 illustre l'annotation de quelques images de la base Corel-5k où les étiquettes de la vérité terrain sont données. Les mots-clés en gras sont trouvés automatiquement par notre modèle d'annotation. Nous constatons que dans la grande majorité des cas les mots-clés trouvés correspondent à la vérité terrain. Les mots-clés en gras et soulignés sont ajoutés automatiquement par notre extension d'annotation. Par exemple, la quatrième image est annotée manuellement par trois mots-clés, deux nouveaux mots-clés "arch" et "pyramid" sont automatiquement

ajoutés après l'extension automatique.

Pour améliorer la précision de la méthode proposée, nous pouvons définir un seuil λ sur la probabilité d'un mot-clé. Une image ne sera annotée par un mot-clé Kw seulement si :

$$P(KW|im_1, im_2, ..., im_m, Kw_1, Kw_2, ..., Kw_n) > \lambda \tag{3.23}$$

TABLE 3.2 – Précision de l'annotation avec différents seuils

Base	Corel-5K			ESP-Game			IAPRTC-12		
Seuil (λ)	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
Précision	39	46	53	39	42	51	37	41	49
Mots-clés par image	3.85	3.38	3.11	4.17	3.76	3.32	4.53	4.17	3.75

Le tableau 3.2 montre la précision de notre modèle avec différents seuils sur les trois ensembles de données. Nous notons clairement que l'utilisation d'un seuil améliore considérablement la précision du modèle au déterminent du nombre de mots-clés par image. L'utilisation du seuil pénalise les annotations qui ont une probabilité inférieure à λ. Par exemple, pour ESP-Game, un seuil de $0,1$ supprime en moyenne un mot-clé par image.

3.6 Application à l'annotation et la classification de documents

3.6.1 Introduction

La quantité de documents numériques stockée chaque jour par les entreprises et les archives personnelles sont en croissance permanente. En collaboration avec le Web, ces archives atteignent des tailles inimaginables. La popularité de ces grandes collections de documents numériques dépend de leur facilité d'utilisation. Cependant, toutes ces bases de données ne sont pas souvent équipées d'informations d'indexation adéquates. Cela rend beaucoup plus difficile l'accès à des informations intéressantes pour l'utilisateur. Ainsi, pour effectuer la recherche d'information sur les images de documents, un mécanisme approprié est nécessaire pour caractériser le contenu du document de manière significative. Pour les images de documents numérisés, l'accès au niveau contenu se faisait traditionnellement à l'aide d'outils de reconnaissance optique des caractères (OCR). Cependant, malgré des efforts considérables, les OCR robustes ne sont pas disponibles pour beaucoup de types de documents [Bor14]. Le texte obtenu est, par conséquent, mal adapté pour l'indexation et la reconnaissance. Les inconvénients des OCR peuvent être surmontés par une approche fondée sur l'annotation manuelle en attribuant des mots-clés pertinents à un document pour décrire son contenu sémantique. L'annotation manuelle étant performante mais très coûteuse pour l'humain, une solution pour ce problème peut être d'annoter partiellement des documents et d'étendre automatiquement les annotations aux autres documents.

La tâche d'annotation des documents est d'un grand intérêt pour les utilisateurs. Elle permet de mieux comprendre le contenu sémantique des documents et permet aux utilisateurs une recherche plus rapide et plus robuste. Contrairement aux pages Web, pour lesquels plusieurs outils

d'annotation sémantique ont été développés comme SMORE [KHPG06], Annotea [KKPS02] et Semtag [DEG+03], moins d'efforts ont été consacrés à l'annotation des images numérisées de documents par mots-clés.

Coüasnon et al. [CCL07] ont proposé deux types d'annotations pour les documents d'archives manuscrits : annotations textuelles et annotations géométriques. Les annotations textuelles représentent toute sorte d'informations sur lesquelles il est intéressant de faire une recherche (date, place, nom, *etc*). Les annotations géométriques représentent une position dans l'image comme une cellule, un champ, ou une zone, représentée par un rectangle ou un polygone. Pour produire automatiquement ces annotations, il est nécessaire, d'abord, de détecter les régions d'intérêt qui contiennent des informations intéressantes. Ensuite, un système de reconnaissance de l'écriture est appliqué. Li et al. [LMVG14] ont proposé un cadre d'apprentissage itératif pour l'étiquetage des symboles graphiques manuscrits. Un graphe relationnel entre les segments est construit. Les nœuds du graphe représentent les segments et les arcs représentent les relations spatiales entre ces segments. Ensuite, les segments sont regroupés pour construire un dictionnaire de codes visuels. Enfin, l'utilisateur donne des étiquettes à ces groupes, ce qui permet de diminuer l'effort manuel. Dans [DCCO14], les auteurs ont présenté une nouvelle approche d'annotation des documents administratifs. La méthode est basée sur l'annotation sémantique des documents suivant le texte contenu dans le document et/ou le logo qu'il contient. La première étape consiste à extraire les logos contenus dans le document. Dans la seconde étape, chaque logo fait l'objet d'une requête sous forme d'image dans un moteur de recherche Web (Google Images [13]) pour identifier le nom du logo. Un ensemble de documents Web ayant un contenu similaire au logo requête est récupéré pour construire un lexique de mots. Dans la dernière étape, le document de départ est annoté avec le lexique des mots construit selon les logos qu'il contient. Chakravarthy et al. [CCL06] ont proposé l'outil AKTiveMedia pour l'annotation de documents textes, images et html. Cet outil utilise à la fois des annotations à base d'ontologies et de texte libre. Les annotations libres sont faites par l'auteur et le lecteur du document. Ces annotations sont ensuite complétées par des ontologies.

Par ailleurs, le problème de la classification de documents [SG12, SM12, CB07] a été traité dans la majorité des travaux de recherche en utilisant l'information textuelle et/ou structurelle. Cependant, l'extraction de ces informations peut être compliquée ou irréalisable pour diverses raisons : documents de mauvaise qualité, en différentes langues, présentant peu d'information textuelle, *etc*. Dans ce cas, il est préférable d'utiliser des caractéristiques visuelles ou de combiner les caractéristiques. Malgré le grand effort dans les travaux de recherche concernant la classification des documents, peu de travaux utilisent une combinaison des caractéristiques visuelles et textuelles, contrairement au domaine des images, où les caractéristiques visuelles sont largement exploitées.

Dans [KYD13], les descripteurs SURF [BETVG08] sont utilisés pour calculer les caractéristiques visuelles d'un document. Le document est segmenté en plusieurs régions. Les relations spatiales entre ces régions sont représentées par un histogramme des mots de code visuel. Les caractéristiques visuelles et spatiales sont utilisées par un classificateur de forêt aléatoire pour la classification et la recherche des documents. Dans [CSB06], des descripteurs de couleur et de texture calculés sur des figures contenues dans les documents sont utilisés pour construire un dictionnaire de mots visuels. Chaque sous figure est représentée par un mot visuel. Ainsi,

13. https ://images.google.com/

le document peut être décrit comme une séquence de mots visuels. Ces derniers sont utilisés par un classificateur naïve Bayes pour la classification de documents. Rusinõl et al. [RoKBL12] ont présenté une nouvelle méthode de recherche des documents en utilisant des caractéristiques visuelles et textuelles. Les caractéristiques visuelles sont représentées par les descripteurs SIFT [Low99] et les caractéristiques textuelles sont représentées par un vecteur de sac de mots fournis par un OCR. Des documents similaires sont récupérés en utilisant la distance classique du cosinus.

Dans cette direction, nous avons appliqué notre modèle présenté dans les sections précédentes aux documents. Ce modèle permet de combiner des caractéristiques visuelles et textuelles afin d'étendre l'annotation aux documents partiellement annotés. Le modèle peut également être utilisé pour la tâche de classification des documents.

3.6.2 Évaluations du modèle sur la base de documents

Pour les caractéristiques visuelles des documents, nous avons utilisé les trois descripteurs LBP [OPH96], l'histogramme des longueurs des séquences [GPV13] et l'histogramme des couleurs RGB. L'histogramme des longueurs des séquences est un descripteur visuel où une séquence est une série de pixels de la même valeur. Cet histogramme est utilisé pour l'analyse et la classification des documents et il est rapide à calculer. Les deux descripteurs LBP et histogramme des couleurs ont été présentés dans la section 3.3. Nous avons expérimenté notre modèle en utilisant la base de documents décrite dans la section 2.4 du chapitre précédent. Nous avons annoté chaque document de test par 5 mot-clés et nous avons calculé le rappel, la précision et le score F1 pour la classification et le score N+ en plus de ces trois critères pour l'annotation.

TABLE 3.3 – Performances de la classification de documents

annotation manuelle (mots-clés)	P	R	F1
200 (2 par document)	78	72	75
300 (3 par document)	92	91	92
500 (5 par document)	97	96	96

TABLE 3.4 – Performances de l'annotation de documents

annotation manuelle (mots-clés)	P	R	F1	N+
200 (2 par document)	28	43	34	28
300 (3 par document)	35	59	44	35
500 (5 par document)	45	62	52	39

Les tableaux 3.3 et 3.4 représentent respectivement les résultats de la classification et de l'annotation (cf. section 3.4.3) des documents de test en fonction du taux d'annotation manuelle effectuée par l'utilisateur au début. Dans la première colonne de chaque tableau, nous présentons le nombre d'annotations dans les documents d'apprentissage. Dans les autres colonnes,

nous présentons les performances de l'annotation et de la classification des documents de test. A partir de ces deux tableaux, nous pouvons remarquer que les performances de classification et d'annotation sont améliorées lorsque le nombre d'annotations manuelles augmente. Par exemple, pour l'annotation, nous passons d'une précision de 28% et un rappel de 43% en utilisant 2 mots-clés par document d'apprentissage, à une précision de 45% et un rappel de 62% en utilisant 5 mots-clés. De même pour la classification, nous passons d'une précision de 78% et un rappel de 72% en utilisant 2 mots-clés par document d'apprentissage, à une précision de 97% et un rappel de 96% en utilisant 5 mots-clés.

TABLE 3.5 – Exemples d'annotation de documents

		
net à payer	formation, études	numéro, consommation, nom
net à payer, salaire, entreprise, date, numéro	**formation, études, compétences, e-mail, adresse**	**numéro, consommation, adresse, net à payer, date**
		
nom, adresse, catégorie	journal, article	nom, numéro, sexe, taille
nom, adresse, catégorie, signature, photo	**journal, article, auteur, paragraphe, publicité**	**nom, numéro, sexe, taille, signature**
		
titre, nom, résumé	nom, opérateur	nationalité
titre, résumé, introduction, travail, expérimentation	**nom, opérateur, e-mail, consommation, adresse**	**nationalité numéro, sexe, taille préfecture**

Le Tableau 3.5 illustre l'annotation de certains documents. Dans la deuxième, la cinquième et la huitième ligne, les étiquettes de la vérité terrain sont données. Dans les autres lignes, nous

trouvons les résultats de notre annotation. Par exemple, le deuxième document (*curriculum vitæ*) a été annoté par deux mots-clés au début, trois nouveaux mots clés "compétences", "e-mail" et "adresse" sont automatiquement ajoutés après l'extension d'annotation.

3.7 Conclusion

Dans ce chapitre, nous avons présenté un modèle graphique probabiliste pour l'extension d'annotation des images. Ce modèle a l'avantage d'être généralisé à différents type d'images. Il est défini par un mélange de distributions multinomiales et de mélanges de Gaussiennes pour lequel nous avons combiné des caractéristiques visuelles et textuelles. Les résultats expérimentaux sur Corel-5k, ESP-Game et IAPRTC-12 sont compétitifs, tout en maintenant à la fois une précision élevée et un rappel élevé de manière équilibrée. Le modèle a été utilisé aussi pour l'annotation et la classification des documents. Le chapitre suivant sera consacré à notre deuxième contribution dans cette thèse. Nous y intégrons les retours utilisateurs pour améliorer les performances et diminuer l'effort manuel d'annotation.

Chapitre 4

Annotation d'images basée sur des retours utilisateurs

Sommaire

4.1 Introduction

Le problème de l'annotation d'image a été largement étudié au cours des dernières années (voir chapitre 2 consacré à l'état de l'art sur les méthodes d'annotation d'images), et plusieurs approches ont été proposées pour résoudre ce problème. Parmi les approches proposées, certaines sont conçues pour réduire l'effort humain et améliorer la précision d'annotation en intégrant l'utilisateur dans le processus d'annotation.

Minerva et al. [NLS+02] ont présenté un outil d'annotation de séquences vidéos. Cet outil est semi-automatique où les étiquettes sont propagées automatiquement à partir d'une séquence

à une autre et l'utilisateur intervient pour confirmer ou rejeter les étiquettes propagées. L'apprentissage incrémental est utilisé dans cet outil via un SVM avec un noyau polynomial pour améliorer les performances de l'outil d'annotation. Une autre méthode similaire a été proposée dans [VR11] mais elle utilise l'apprentissage actif au lieu de l'apprentissage incrémental.

Dans [BJ15], un modèle d'annotation d'images à base d'apprentissage actif est proposé. Ce modèle combine l'approche du plus proche voisin ainsi que l'apprentissage actif pour annoter une nouvelle image. Les auteurs ont utilisé l'incertitude pour la sélection des échantillons à annoter manuellement. Ils ont traité le problème de déséquilibre de classes en sélectionnant un sous-ensemble d'exemples d'apprentissage avec une fréquence similaire d'étiquettes plutôt qu'un jeu complet d'apprentissage. Ils ont affiné l'algorithme 2-PKNN [VJ12] pour l'apprentissage actif en permettant à cet algorithme de prévoir un nombre variable de mots-clés pour chacune des images. Pour chaque image, ils ont conservé toutes les étiquettes qui satisfont un seuil minimum. Une image est alors sélectionnée pour l'apprentissage actif si son score de prédiction dépasse le seuil.

Zhang et al. [ZC03] ont présenté un système de recherche d'images qui utilise l'annotation cachée avec l'apprentissage actif pour améliorer les performances de recherche. Le système commence par initialiser la liste de probabilités d'avoir une annotation avec des probabilités a priori. Quelques images sont choisies et annotées au hasard. La liste de probabilités est recalculée sur la base des nouvelles images annotées au hasard. Le système sélectionne les images qui ont le gain maximum de connaissances, et demande à un utilisateur de les annoter. Encore une fois, la liste de probabilités de certaines images de la base de données est mise à jour parce que l'un de leurs voisins a été récemment annoté. Cette boucle est répétée jusqu'à ce que l'annotateur arrête ou la base de données soit entièrement annotée.

Une autre approche pour annoter les images semi-automatiquement et progressivement par des mots-clés est présentée dans [WDS+01]. Cette approche combine l'efficacité d'annotation automatique et l'exactitude d'annotation manuelle. Le processus d'annotation progressif est intégré au cours de la recherche d'images par mots-clés et par contenu. Lorsque l'utilisateur soumet une requête par mot-clé, puis fournit un retour de pertinence, les mots-clés de recherche sont automatiquement ajoutés aux images qui reçoivent une rétroaction positive et peuvent alors faciliter la recherche par mot-clé à l'avenir. Le processus d'annotation avec cette méthode est illustré dans la figure 4.1.

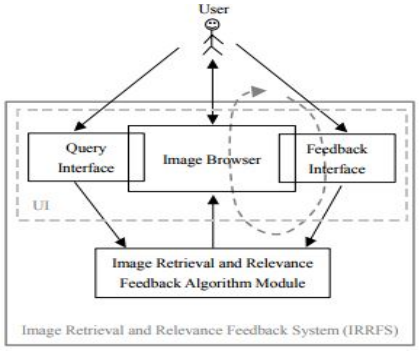


FIGURE 4.1 – Schéma général de la méthode [WDS+01]

Le problème d'annotation automatique d'images a aussi été traité dans [BL09] à l'aide d'apprentissage incrémental de classificateurs. Cet apprentissage commence avec une petite quantité de données labellisées pour créer un ensemble initial de classificateurs, et intègre progressivement des données non étiquetées dans le processus d'apprentissage incrémental pour améliorer les modèles. Les données non labellisées sont choisies en calculant une mesure de confiance et un score de discrimination de marge (margin-like discrimination score).

Afin de minimiser le laborieux effort manuel d'annotation, Gerard et al. [SCG02] ont utilisé un apprentissage actif. Lorsqu'un utilisateur tape un mot-clé pour rechercher une image, l'algorithme d'apprentissage actif sélectionne les images les plus ambigues par rapport à ce mot-clé afin de solliciter les commentaires des utilisateurs. L'utilisateur marque simplement une image comme « pertinente » ou « non pertinente » pour la requête.

Il ressort de cette étude bibliographique, l'importance du retour utilisateur dans un modèle d'annotation d'images. Par conséquent, afin d'améliorer la qualité de l'annotation et réduire l'effort manuel, dans ce chapitre, nous proposons d'intégrer des retours utilisateurs dans notre modèle décrit dans le chapitre précédent. Les retours utilisateurs sont à base de trois types d'apprentissages. Dans le premier, l'utilisateur améliore la qualité de l'ensemble d'apprentissage. Plus précisément, les images de l'ensemble d'apprentissage sont initialement annotées par un petit ensemble de mots-clés. Ensuite, le modèle est utilisé pour ajouter automatiquement d'autres annotations aux images. L'utilisateur valide les nouvelles étiquettes correctes et rectifie certaines fausses étiquettes. Les images avec leurs nouvelles étiquettes sont utilisées pour réapprendre le modèle. Ce processus peut être répété plusieurs fois. De cette façon, nous réduisons l'effort fastidieux de l'annotation manuelle. Au deuxième niveau, lors d'une étape de recherche d'images, les données sont progressivement introduites dans le système et apprises progressivement. Le système doit pouvoir changer et ajuster ses paramètres après l'observation de chaque exemple. Au troisième niveau, nous proposons un apprentissage actif de notre modèle pour réduire le nombre d'instances en choisissant l'exemple conduisant à la plus grande réduction du taux d'erreur. Enfin, nous combinons dans notre système ces trois modes d'apprentissage.

4.2 Retour utilisateur à base d'apprentissage dans l'apprentissage

4.2.1 Méthode proposée

Comme notre modèle est supervisé, nous avons besoin d'un ensemble de données pour l'apprentissage de ses paramètres. En comparant les performances d'annotation des mêmes images de test en utilisant un ensemble de données d'apprentissage annoté et un ensemble de données d'apprentissage partiellement annoté, nous avons remarqué que les résultats sont meilleurs en utilisant le premier ensemble de données (voir tableau 4.1). Pour obtenir un ensemble de données d'apprentissage bien annoté, un utilisateur doit le faire manuellement. L'annotation manuelle est un travail intensif et un processus fastidieux. Pour minimiser l'effort manuel et laborieux de l'annotation, nous voulons obtenir un jeu de données d'apprentissage annoté en utilisant un ensemble de données partiellement annoté et le retour des utilisateurs. Nous appelons le type d'apprentissage proposé "apprentissage dans l'apprentissage". La méthode proposée est présentée à travers un exemple illustratif dans la figure 4.2. À partir d'un petit ensemble d'anno-

tations (2 mots-clés par image), nous utilisons notre modèle pour ajouter un troisième mot-clé à chaque image de l'ensemble d'apprentissage par extension automatique d'annotation (étapes 1 et 2 de la figure 4.2). L'utilisateur intervient pour valider les annotations correctes et rectifier les fausses annotations (étape 3). Ce nouveau jeu de données obtenu permet de réapprendre le modèle et d'ajouter un quatrième mot-clé en inférant le modèle (étapes 4 et 5). L'utilisateur intervient de nouveau pour validation et correction (étape 6). Ce processus peut être répété plusieurs fois pour ajouter d'autres annotations. Les résultats expérimentaux présentés plus tard montrent un gain considérable par rapport à un effort manuel et que les résultats obtenus sont meilleurs qu'avec un jeu de données partiellement annotées.

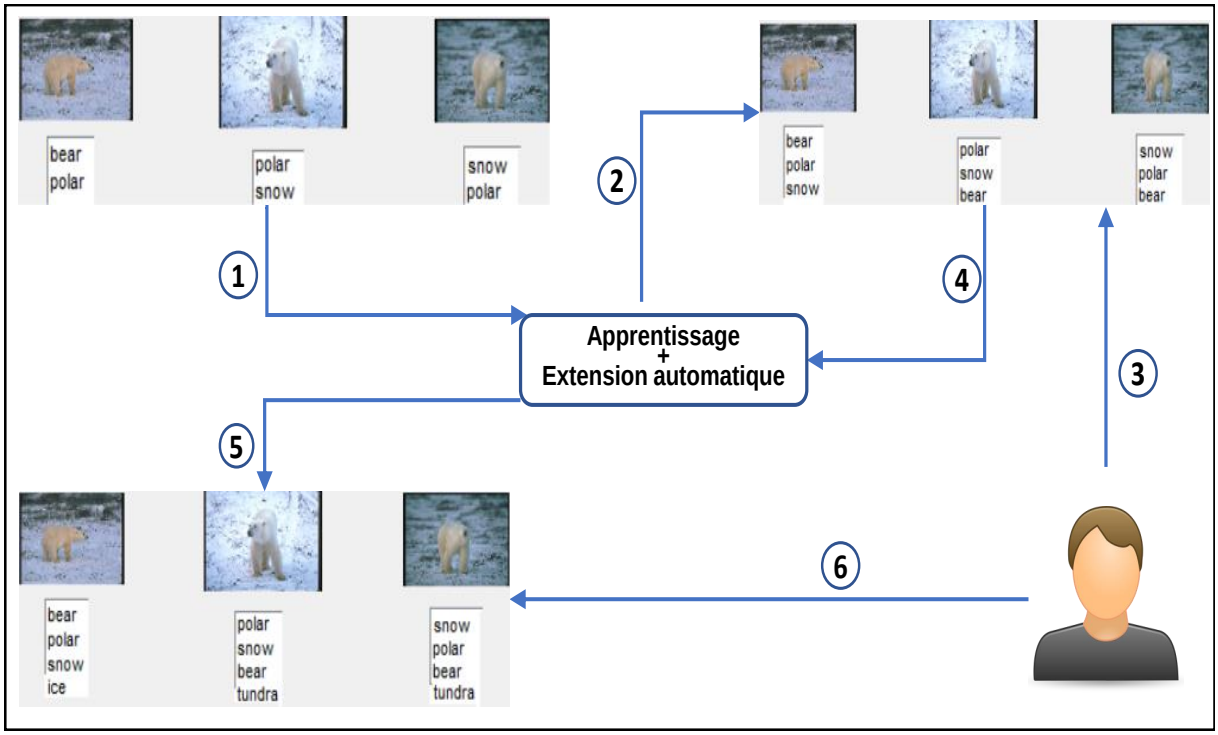


FIGURE 4.2 – Étapes du retour utilisateur à base de l'apprentissage dans l'apprentissage

Par exemple dans la figure 4.2, nous avons au début trois images (les trois images en haut à gauche de la figure) annotées chacune par deux mots-clés. Nous faisons l'apprentissage des paramètres de notre modèle avec ces images (étape 1). Après la phase d'apprentissage, nous inférons le modèle (étape 2) pour ajouter un troisième mot-clé à chacune des trois images (les trois images en haut à droite de la figure). Les trois mots-clés ajoutés sont « snow », « bear » et « bear ». L'utilisateur intervient pour corriger les mots-clés ajoutés (étape 3). Dans notre cas, les mots-clés ajoutés sont corrects donc l'utilisateur n'a rien à faire. Les trois images avec leurs nouvelles annotations sont utilisées pour réapprendre le modèle (étape 4). Après cette nouvelle phase d'apprentissage, nous inférons le modèle de nouveau (étape 5) pour ajouter un quatrième mot-clé à chacune des trois images (les trois images en bas à gauche de la figure). Les trois nouveaux mots-clés ajoutés sont « ice », « tundra » et « tundra ». L'utilisateur intervient de nouveau

pour corriger les mots-clés ajoutés (étape 6). Ainsi, partant de deux mots-clés par image, nous obtenons quatre mots-clés par image en utilisant des extensions automatique d'annotation et des retours utilisateurs. Cette méthode nous permet de générer automatiquement des mots-clés corrects ajoutés par le modèle qui auraient dû être ajoutés manuellement par l'utilisateur.

4.2.2 Évaluations

Le tableau 4.1 représente les résultats de l'annotation des images de test en fonction des annotations manuelles effectuées par l'utilisateur. Dans la première colonne, nous présentons le nombre d'annotations par image dans l'ensemble d'apprentissage. Dans les autres colonnes, nous montrons les performances d'annotation des images sur les trois ensembles de test décrits précédemment. À partir de ce tableau, nous pouvons remarquer que les performances sont améliorées de manière significative lorsque la longueur d'annotation passe de 2 à 4 mots-clés. En effet, lorsque plus de mots sont générés pour chaque image, son contenu sémantique est décrit d'une façon plus précise. Par conséquent, une longueur d'annotation plus longue des images d'apprentissage entraîne un rappel et une précision plus élevés pour les images de test. Par exemple, pour l'ensemble de données Corel-5K, nous passons d'une précision de 13% et d'un rappel de 18% (taux d'annotation manuel de 2 mots-clés par image), vers une précision de 34% et un rappel de 44% avec 4 mots clés.

TABLE 4.1 – Performances de l'apprentissage dans l'apprentissage

Longueur d'annotation	Corel-5K P R F1 N+	ESP-Game P R F1 N+	IAPRTC-12 P R F1 N+
2 mots-clés	13 18 15 97	16 23 19 206	13 18 15 175
3 mots-clés	24 33 28 132	21 28 24 227	19 26 22 232
4 mots-clés	34 44 38 174	25 32 28 241	25 31 28 243

TABLE 4.2 – Gain en effort manuel

	Corel-5K	ESP-Game	IAPRTC-12
3^{eme} mot-clé	512/3940	2508/15679	2088/16062
4^{eme} mot-clé	566/2353	2698/12848	2847/14279
Gain en effort manuel	17,13 %	18,24 %	16,26 %

Pour quantifier le gain en effort manuel, nous avons fait l'expérience suivante où les résultats sont présentés dans le tableau 4.2. Nous ne conservons que deux mots-clés par image. Nous essayons d'atteindre 4 mots-clés par image à l'aide d'une extension d'annotation itérative et des retours utilisateurs. Pour cette fin, un troisième mot-clé est automatiquement ajouté aux images d'apprentissage. Pour l'ensemble de données Corel-5K, sur 3940 mots-clés ajoutés (3940 images dans la vérité terrain sont annotées par 3 mots-clés), 512 annotations sont correctes, mais les autres 3428 ont besoin de la correction d'un utilisateur. Ce processus est

répété pour ajouter un quatrième mot-clé. Le nombre de corrections de l'utilisateur est de 1787 sur 2353 mots-clés ajoutés (2353 images dans la vérité terrain sont annotées par 4 mots-clés). Ainsi, nous gagnons $(512 + 566)/(3940 + 2353) = 17,13\%$ d'effort manuel. La même expérience est effectuée pour les deux autres bases de données ESP-Game et IAPRTC-12 où nous atteignons un gain de $18,24\%$ et de $16,26\%$ respectivement. Nous constatons aussi que le gain devient plus important au fur et à mesure que le nombre de mot-clé augmente.

4.3 Retour utilisateur à base d'apprentissage incrémental

4.3.1 Méthode proposée

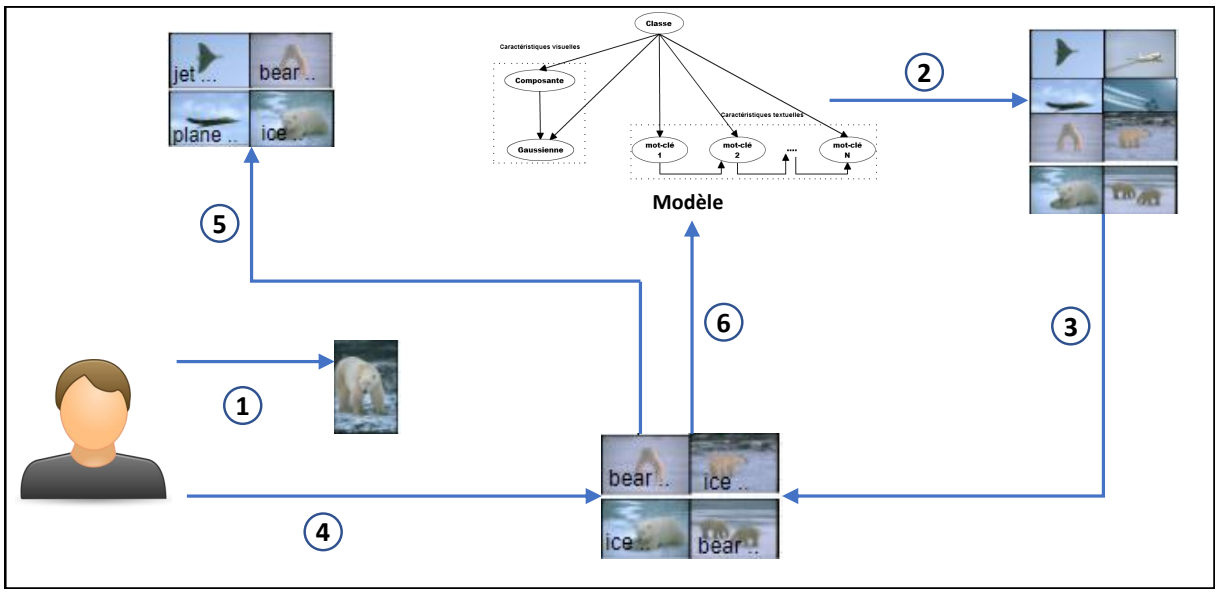


FIGURE 4.3 – Étapes du retour utilisateur à base de l'apprentissage incrémental

Le deuxième retour utilisateur est à base d'apprentissage incrémental. Dans un processus de recherche, l'utilisateur présente une image comme requête de recherche. Le système lui donne le résultat de recherche en affichant une liste d'images similaires à la requête. Les images récupérées sont accompagnées avec leurs annotations effectuées par le système. L'utilisateur intervient pour corriger les fausses annotations. Les images avec leurs nouvelles annotations après l'intervention de l'utilisateur sont utilisées pour enrichir l'ensemble d'apprentissage et la mise à jour des paramètres du modèle. Les étapes de ce retour utilisateur, comme illustré dans la figure 4.3, sont les suivantes :

1. L'utilisateur lance une recherche à travers une image requête ;
2. Le modèle propose une liste d'images à partir des images non annotées ;
3. Le modèle annote les images retournées par des mots-clés ;

4. L'utilisateur intervient pour corriger les fausses annotations ;
5. Les images avec leurs nouvelles annotations après l'intervention de l'utilisateur sont ajoutées à l'ensemble d'apprentissage ;
6. Les paramètres du modèle sont mis à jour avec les nouvelles images et leurs annotations.

La mise à jour des paramètres du modèle dans l'étape 6 est effectuée à travers un apprentissage incrémental. Dans cet apprentissage, les images de test sont annotées par le modèle. L'utilisateur corrige les fausses étiquettes. Les paramètres du modèle sont mis à jour de manière itérative en utilisant le modèle actuel et de nouveaux exemples.

Soit Φ le paramètre global du modèle et I_j l'image j après correction de l'annotation par l'utilisateur. I_j est représentée par ses caractéristiques visuelles $im_1, im_2, ..., im_m$ et ses mots clés existants $Kw_1, Kw_2, ..., Kw_n$.

$$\Phi^j \sim P(\Phi|\Phi^{j-1}, I_j) \tag{4.1}$$

Où Φ^{j-1} représente les paramètres du modèle appris à partir des $j-1$ images précédentes. Pour la mise à jour des paramètres, nous avons utilisé l'algorithme proposé dans [Die93] qui permet l'ajustement des paramètres d'un réseau bayésien après l'arrivée d'une nouvelle observation.

4.3.2 Évaluations

TABLE 4.3 – Taux d'annotation avec l'apprentissage incrémantal (Corel-5K)

Taille de l'ensemble d'apprentissage (images)	plane (%)	bear (%)	tiger (%)	horse (%)	sport (%)
1000	61,36	57,88	75,83	79,69	78,44
1000+20 (plane)	62,68	57,95	76,03	79,82	78,68
1020+20 (bear)	62,85	59,27	76,63	80,54	79,19
1040+20 (tiger)	63,13	60,43	78,94	80,92	79,66
1060+20 (horse)	63,47	61,12	79,86	82,80	80,17
1080+20 (sport)	64,02	61,31	80,05	82,98	84,61

Le tableau 4.3 montre les résultats d'apprentissage incrémental sur 5 classes de l'ensemble de données Corel-5K. L'ensemble d'apprentissage utilisé comprend 1000 images annotées et l'ensemble de test est composé de 100 images (20 par classe). Les images de test sont utilisées progressivement pour enrichir l'ensemble d'apprentissage et pour la mise à jour des paramètres du modèle. Tout d'abord, les images de test sont automatiquement annotées par le modèle. Le taux d'annotation de chaque classe est représenté dans la première ligne du tableau 4.3. Ensuite, l'utilisateur corrige les annotations des 20 images de la classe « plane ». Ces images avec leurs nouvelles étiquettes sont utilisées pour l'apprentissage incrémental. Le taux d'annotation après le retour utilisateur est présenté en deuxième ligne. Le processus est répété pour corriger les annotations de 20 images des autres classes. Nous remarquons que le taux d'annotation de chaque classe est amélioré lorsque l'ensemble d'apprentissage est enrichi par de nouveaux exemples de la même classe. Nous remarquons également que les nouveaux exemples de la classe « bear » (de même pour les classes « tiger » et « horse ») ont plus d'influence sur les classes « tiger »

et « horse » que sur « plane » et « sport ». La raison est que les trois classes « bear », « tiger » et « horse » sont plus proches sémantiquement entre eux que les deux autres classes, de sorte qu'elles peuvent partager plus d'annotations entre elles qu'avec les deux autres classes.

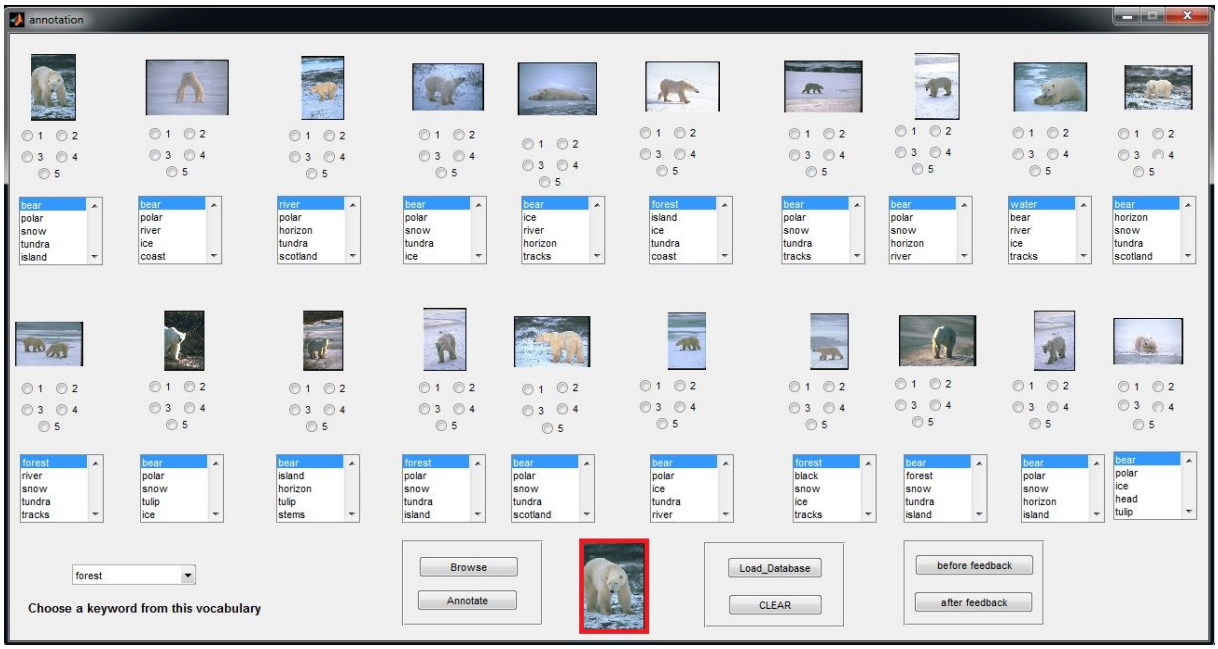


FIGURE 4.4 – Interface graphique de notre système d'annotation

Dans la figure 4.4, l'utilisateur lance une recherche en utilisant une image requête (encadrée en rouge). Des images similaires à la requête avec leurs annotations automatiques sont présentées par le modèle (en utilisant le bouton « Annotate »). L'utilisateur intervient pour corriger les fausses annotations de certaines images. Par exemple, s'il veut changer le cinquième mot « island » de la première image (en haut a gauche de la figure 4.4) par le mot-clé « snow », il choisit le mot-clé « snow » du vocabulaire d'annotation (la liste en bas à gauche de la figure 4.4) et clique sur le bouton radio numéro 5 sous la première image.

Une fois que les fausses étiquettes sont corrigées, l'utilisateur peut ajouter ces images à l'ensemble d'apprentissage pour la mise à jour des paramètres en utilisant le bouton « after feedback ». Ce bouton permet aussi d'afficher le résultats d'annotations de quelques images après le retour utilisateur. Pour voir l'annotation de ces images avant le retour utilisateur, le bouton « before feedback » est utilisé. Ainsi, on peut voir l'impact du retour utilisateur sur l'annotation des images. La figure 4.5 et 4.6 représentent respectivement un exemple de l'annotation de 20 images de la classe « bear » avant et après le retour utilisateur. Par exemple, l'image numéro 3 (image 3 de la première ligne de la figure 4.5) était annotée par les mots-clés « bear », « polar », « river », « flag » et « water ». Après le retour utilisateur, elle est annotée par les mots-clés « bear », « polar », « frost », « snow » et « ice ».

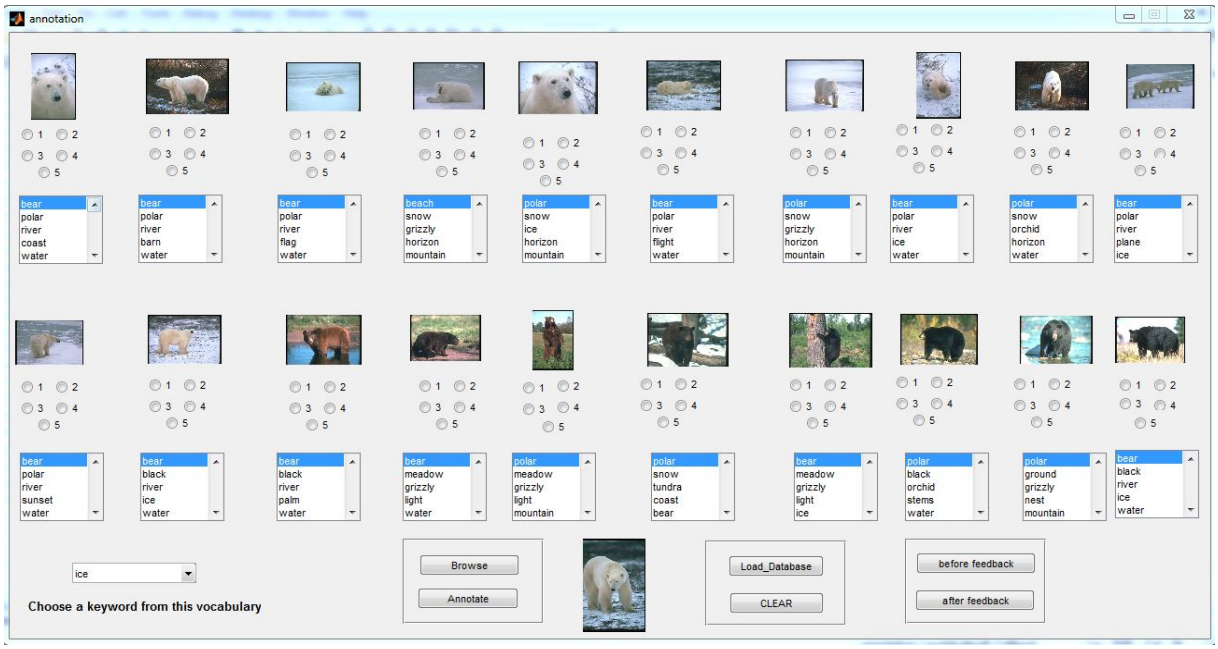


FIGURE 4.5 – Annotation d'images de la base Corel-5K avant le retour utilisateur

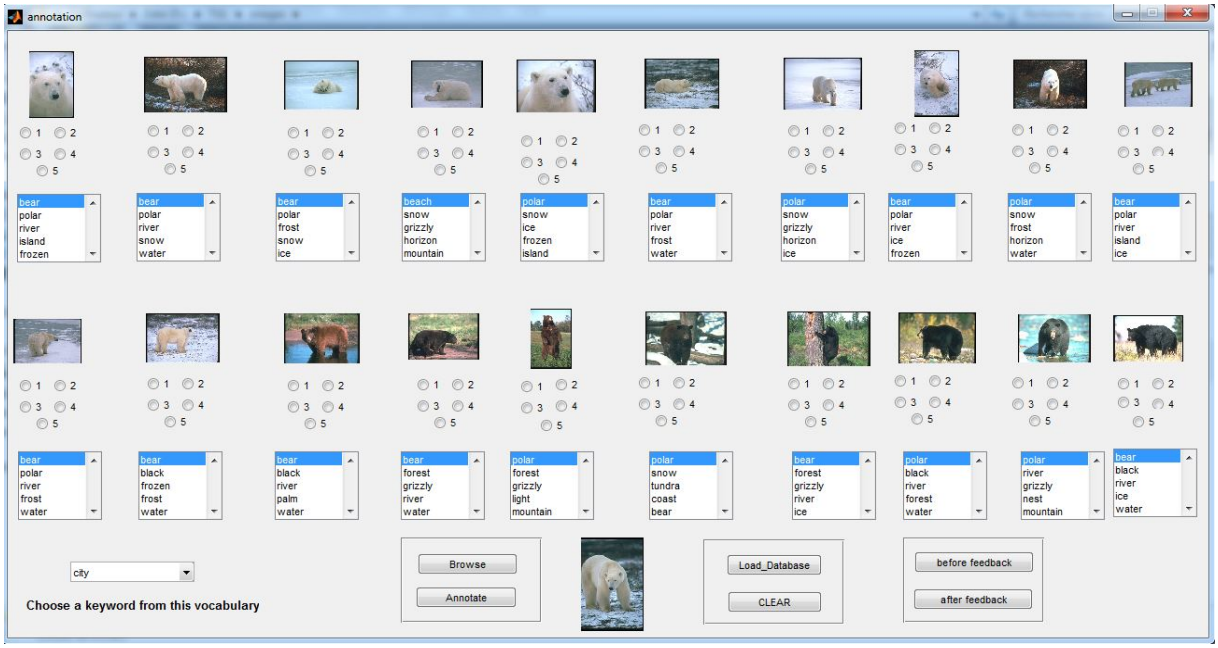


FIGURE 4.6 – Annotation d'images de la base Corel-5K après le retour utilisateur

4.4 Retour utilisateur à base d'apprentissage actif

Malgré de nombreuses recherches existantes qui ont été consacrées à l'annotation d'images, des études [LTS03] montrent qu'un grand ensemble d'apprentissage est souvent nécessaire afin d'obtenir des performances raisonnables en raison de la grande variété du contenu de l'image. Toutefois, l'annotation manuelle est laborieuse et prend du temps. Par exemple, des études existantes montrent qu'annoter typiquement 1 heure de vidéo avec 100 concepts peut prendre entre 8 à 15 heures [LTS03]. En outre, la performance de l'annotation est directement proportionnelle à la taille de l'ensemble d'apprentissage. Plus large est l'échantillon d'apprentissage, meilleure est la performance. Cependant, beaucoup de ces exemples d'apprentissage sont redondants et n'ajoutent pas de valeur à la tâche d'apprentissage. L'identification de ces exemples permet de réduire les efforts manuels. Par conséquent, plusieurs travaux utilisent l'apprentissage actif dans l'annotation d'images pour réduire les données d'apprentissage nécessaires.

4.4.1 Approches générales de l'apprentissage actif

L'apprentissage actif est une technique d'apprentissage automatique qui sélectionne les échantillons les plus informatifs pour l'étiquetage et les utilise comme données d'apprentissage. Dans l'apprentissage actif, les échantillons sont sélectionnés à partir d'un ensemble existant pour l'étiquetage en fonction de certains critères [WH11]. Un système d'apprentissage actif typique est composé de deux étapes : un moteur d'apprentissage et un moteur de sélection d'échantillons. À chaque étape, le moteur d'apprentissage apprend un modèle basé sur l'ensemble d'apprentissage actuel. Ensuite, le moteur de sélection d'échantillons sélectionne les échantillons non étiquetés les plus informatifs pour l'étiquetage manuel, et ces échantillons sont ajoutés à l'ensemble d'apprentissage. De cette façon, l'ensemble d'apprentissage obtenu est plus informatif que celui recueilli par l'échantillonnage aléatoire. Au cours des dernières décennies, plusieurs stratégies de sélection des échantillons ont été proposées dans les travaux de l'état de l'art. Nous présentons dans la suite de cette section les principales stratégies.

4.4.1.1 Échantillonnage par incertitude

Dans la stratégie de sélection par incertitude [LG94], l'exemple non étiqueté avec l'incertitude maximale est choisi pour l'annotation manuelle à chaque cycle d'apprentissage. L'incertitude maximale implique que l'apprenant actuel interroge l'exemple sur lequel il est le moins certain à propos de son étiquette. Dans cette stratégie, l'exemple non marqué avec une incertitude maximale est considéré comme l'exemple le plus instructif. Le point clé de l'échantillonnage par incertitude est de savoir comment mesurer l'incertitude d'un exemple non marqué. L'incertitude peut être mesurée en utilisant l'entropie [Sha48], la probabilité de prédiction [LG94] ou la marge entre deux classes [SDW01].

4.4.1.2 Échantillonnage par comité de modèles

L'échantillonnage par comité de modèles *(Query By Committee (QBC))* [SOS92] consiste à définir une comité de modèles à partir du modèle de base. Ces modèles sont entraînés sur le même ensemble d'exemples étiquetés. Chaque modèle du comité vote pour les étiquettes

des exemples non étiquetés. L'exemple le plus informatif qui sera sélectionné est l'exemple pour lequel les modèles définis sont le plus en désaccord. Pour mesurer le niveau de désaccord entre les modèles du comité pour un exemple donné, le vote par entropie [DE95], la divergence de Kullback-Leibler [MN98] et la divergence de Jensen–Shannon [MYSTM05] peuvent être utilisés.

4.4.1.3 Échantillonnage par changement de modèle prévu

La stratégie basée sur le changement de modèle prévu *(Expected Model Change)* [SCR08] sélectionne l'exemple qui donnerait le plus grand changement au modèle actuel si nous connaissions son étiquette. L'intuition derrière cette méthode est qu'elle préfère les exemples susceptibles d'influencer le modèle (c'est-à-dire avoir le plus d'impact sur ses paramètres), quelle que soit l'étiquette de requête qui en résulte.

4.4.1.4 Échantillonnage par réduction de l'erreur de généralisation

Roy et al. [RM01] ont proposé une stratégie de sélection d'exemples basée sur la réduction de l'erreur de généralisation *(expected error reduction (EER))*. Cette stratégie vise à mesurer non pas combien le modèle risque de changer, mais combien son erreur de généralisation risque de diminuer. Elle vise à réduire l'erreur de généralisation lors de l'étiquetage d'une nouvelle instance. Étant donné que nous n'avons pas accès aux données de test, Roy et al. ont suggéré de calculer la « future erreur » sur l'ensemble non étiqueté dans l'hypothèse que cet ensemble soit représentatif de la distribution de l'ensemble de test. En d'autres termes, l'ensemble non étiqueté peut être considéré comme un ensemble de validation. En outre, nous n'avons aucune connaissance des étiquettes vraies des échantillons non étiquetés. L'échantillonnage par réduction de l'erreur de généralisation estime alors le critère de cas moyen de perte potentielle.

4.4.1.5 Autres stratégies

D'autres stratégies d'échantillonnage ont été proposées dans les travaux de l'état de l'art telles que la réduction de la variance [CGJ96] et la pondération des instances [BBB16]. La réduction de la variance consiste à étiqueter les exemples qui minimiseront la variance de sortie. La pondération des instances consiste à associer un poids à un exemple permettant de déterminer son importance par rapport aux autres exemples.

4.4.2 Méthode proposée

Malgré de nombreuses recherches existantes qui ont été consacrées à l'annotation d'images, un grand ensemble d'apprentissage est souvent nécessaire afin d'obtenir des performances raisonnables en raison de la grande variété des contenus des images. Cependant, beaucoup de ces exemples d'apprentissage sont redondants et n'ajoutent pas de valeur à la tâche d'apprentissage. L'identification de ces exemples peut réduire les efforts manuels. En outre, la sélection des échantillons non étiquetés les plus informatifs donne lieu à un ensemble d'apprentissage plus informatif que celui recueilli par échantillonnage aléatoire. Par conséquent, nous propo-

sons d'utiliser l'apprentissage actif dans notre modèle d'annotation d'images pour réduire les données d'apprentissage nécessaires.

Plusieurs stratégies, comme présenté dans la section précédente, ont été développées pour la sélection d'échantillons dans l'apprentissage actif. Nous proposons d'utiliser la stratégie d'échantillonnage par réduction de l'erreur de généralisation [TK01] qui consiste à choisir une fonction de perte pour estimer le taux d'erreur futur sur tous les exemples et les étiquettes. L'exemple conduisant à la plus grande réduction du taux d'erreur est sélectionné. Cette approche sélectionne les exemples qui minimiseront l'erreur de généralisation du modèle prédictif, s'ils ont été annotés. L'erreur du modèle M avec les paramètres Φ est notée $Err_{\Phi}(M)$. Cette erreur est calculée à l'aide d'une fonction de perte, notée $Loss(M, \Phi)$, qui évalue l'erreur du modèle avec les paramètres Φ pour une image particulière I qui appartient à l'ensemble d'apprentissage T. Le calcul de l'erreur de généralisation intègre la fonction $Loss(M, \Phi)$, où $P(\Phi)$ est la distribution de probabilité des paramètres :

$$Err_{\Phi}(M) = \int_{T} Loss(M, \Phi) P(\Phi) d\Phi \tag{4.2}$$

Un exemple non annoté est ajouté à l'ensemble d'apprentissage et des hypothèses sont faites sur la valeur de l'étiquette manquante. Étant donné un exemple supplémentaire, les nouveaux paramètres du modèle seront notés $\tilde{\Phi}$. Le risque est estimé par une intégration sur tous les mots-clés du vocabulaire V.

$$Risk(M) = \int_{V} P(\tilde{\Phi}|\Phi) Err_{\Phi}(M) d\tilde{\Phi} \tag{4.3}$$

Il existe de nombreux choix possibles pour la fonction de perte telles que l'entropie ou la divergence de Kullback-Leibler (KL-divergence) [KL51]. Nous avons choisi cette dernière pour calculer la perte entre la distribution réelle et postérieure des paramètres.

Dans la théorie de l'information, la divergence de Kullback-Leibler est une mesure non symétrique de la différence entre deux distributions de probabilité P et Q. Spécifiquement, la divergence KL de Q à partir de P, définie par l'équation 4.4 , représente l'information perdue en utilisant la distribution Q pour approximer la distribution P. Généralement, P représente la distribution réelle que Q va approximer.

$$KL(P||Q) = \sum_{x} P(x) ln \frac{P(x)}{Q(x)} \tag{4.4}$$

L'algorithme 2 décrit les itérations de la méthode proposée. Plusieurs critères d'arrêt d'un apprentissage actif ont été proposés dans les travaux de l'état de l'art [ZWH08] tels que :

- l'ensemble des exemples non étiquetées vide ;
- les exemples non étiquetés ont tous une utilité nulle ou faible ;
- un nombre maximal n prédéfini d'exemples à étiqueter est atteint ;
- le classificateur actuel atteint l'efficacité maximale.

Dans notre algorithme, nous avons choisi le critère le plus simple qui consiste à définir un nombre de n exemples à annoter. Nous avons défini ce nombre à 80% de l'ensemble d'apprentissage.

Algorithm 2 : Algorithme d'apprentissage actif

Entrées : Sous-ensemble d'images annotées (T-sous), ensemble d'images non annotées (T), modèle M, n

Sorties : Images annotées

1 : Apprentissage du modèle avec T-sous ;
Répéter
pour Chaque image I dans T **faire**
pour Chaque mot-clé Kw dans le vocabulaire V **faire**
2 : Calculer l'erreur de généralisation avec l'équation (4.2)
fin pour
3 : Calculer le risque d'annoter l'image I avec l'équation (4.3)
fin pour
4 : chercher l'image $I_{min} = ArgMin_{I \in T} Risk(M)$;
5 : Demander à l'utilisateur d'annoter I_{min} ;
6 : T-sous = T-sous $\cup$ I_{min} ;
7 : $T = T - I_{min}$
Jusqu'à |T-sous| $> n$

4.4.3 Évaluations

Nous étudions la variation des performances d'annotation en fonction de la taille de l'ensemble d'apprentissage. Nous avons calculé le changement des performances en augmentant progressivement la taille des données d'apprentissage pour les deux algorithmes : avec simple apprentissage et avec apprentissage actif. Dans cette expérience, nous avons augmenté progressivement la quantité d'images d'apprentissage pour l'apprentissage actif. De plus, pour montrer que nous sélectionnons efficacement les échantillons les plus informatifs, nous échantillonnons aléatoirement la même quantité d'images et nous les ajoutons à l'ensemble d'apprentissage pour calculer les performances.

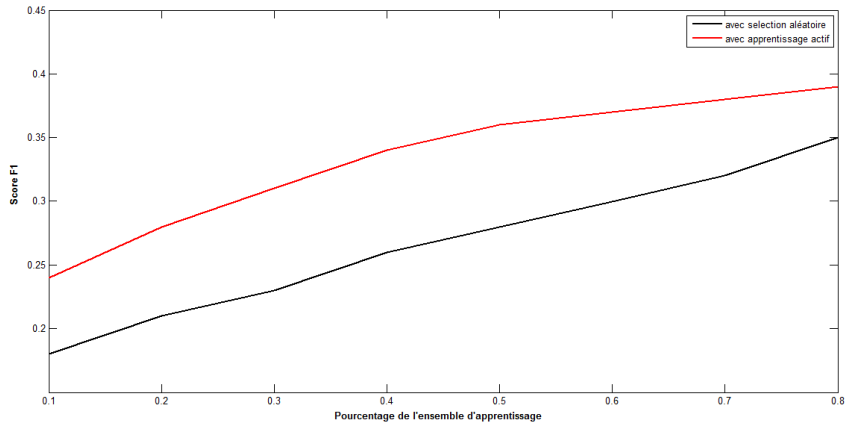


FIGURE 4.7 – Score F1 en fonction de l'ensemble d'apprentissage Corel-5K

Les trois figures 4.7, 4.8, et 4.9 représentent respectivement le score F1 des trois bases de test Corel5k, ESP-Game et IAPRTC12 en fonction du pourcentage de l'ensemble d'apprentis-

sage. Il ressort de ces trois figures que le taux d'annotation est relativement plus élevé pour l'apprentissage actif que l'apprentissage avec la sélection aléatoire des échantillons.

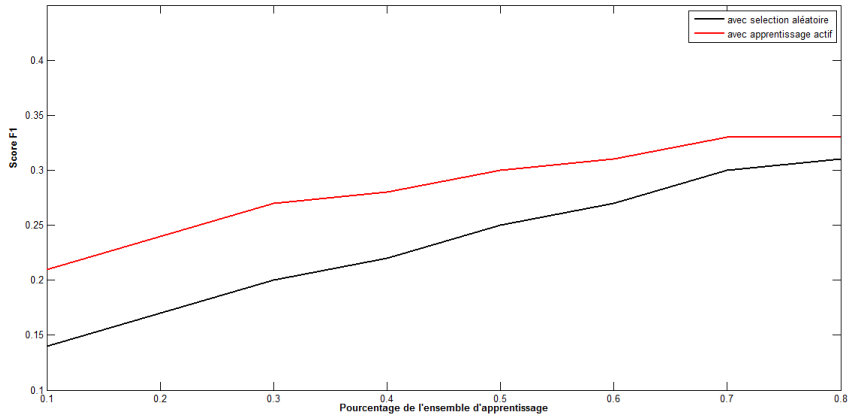


FIGURE 4.8 – Score F1 en fonction de l'ensemble d'apprentissage ESP-Game

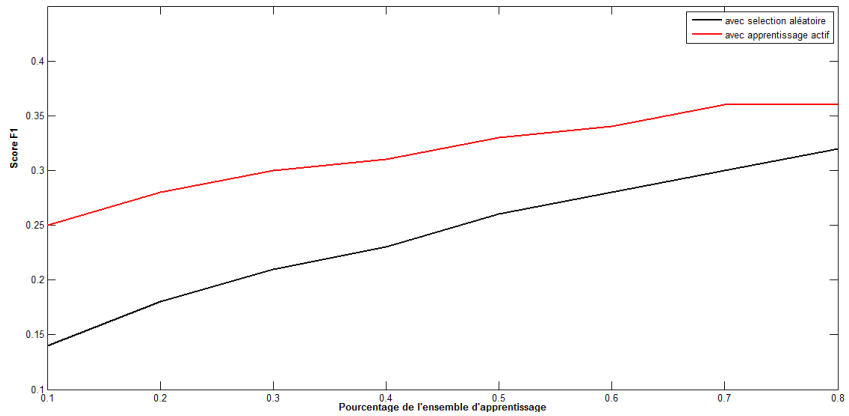


FIGURE 4.9 – Score F1 en fonction de l'ensemble d'apprentissage IAPRTC-12

Le tableau 4.4 montre les résultats d'annotation des images de test sur les trois ensembles de données à l'aide d'un apprentissage simple (sélection aléatoire de données) et d'un apprentissage actif. Tout d'abord, nous observons que la précision et le rappel des deux méthodes sont améliorés proportionnellement au pourcentage de l'ensemble d'apprentissage. C'est parce que pour la plupart des algorithmes d'apprentissage supervisés, plus d'exemples d'apprentissage conduisent généralement à une meilleure performance. Deuxièmement, par rapport au modèle de référence, la méthode d'apprentissage actif est sensiblement plus efficace pour améliorer à la fois la précision et le rappel de notre modèle d'annotation. Par exemple, Les résultats avec $10\%(P = 20; R = 29)$ des données d'apprentissage sélectionnées sont supérieures à $30\%(P = 19; R = 29)$ sélectionnées de manière aléatoire (ensemble de données Corel-5K). De même, pour les deux autres ensembles de données, la sélection des données d'apprentissage améliore grandement la performance du modèle. Pour les trois ensembles de données, nous pouvons remarquer l'efficacité de l'apprentissage actif, car, avec 50% d'exemples sélectionnés, nous obtenons des résultats similaires ou meilleurs que 80% d'exemples aléatoires. Nous obtenons ainsi un gain d'environ 30%.

TABLE 4.4 – Apprentissage actif (AA) contre sélection aléatoire (SA)

Ensemble	Corel-5K		ESP-Game		IAPRTC-12	
d'apprentissage %	SA	AA	SA	AA	SA	AA
	P R	P R	P R	P R	P R	P R
10	14 24	20 29	12 18	18 25	13 16	23 27
20	17 26	24 33	14 21	21 27	16 20	26 30
30	19 29	27 36	17 23	24 30	19 23	28 32
40	22 31	30 39	19 25	26 31	21 25	29 34
50	24 34	32 41	22 28	28 33	24 28	31 35
60	26 36	33 43	25 30	29 34	27 30	32 37
70	28 38	34 44	27 33	31 35	28 33	34 38
80	30 41	34 45	29 34	31 36	30 34	34 39

Nous avons clairement démontré, avec nos expériences, que l'algorithme d'apprentissage actif que nous proposons présente de multiples avantages en matière de gain de performance ainsi que de réduction des besoins en données d'apprentissage. Avec un apprentissage actif, nous réduisons le temps et les coûts nécessaires pour définir les données d'apprentissage. L'algorithme est en mesure de sélectionner des échantillons pertinents à partir des données de test afin d'améliorer considérablement la précision globale d'annotation.

4.5 Retour utilisateur à base des trois apprentissages

4.5.1 Méthode proposée

Pour profiter des avantages des trois retours utilisateurs précédents, nous les avons combiné en un seul retour utilisateur. Lorsque une image est sélectionnée par l'apprentissage actif, l'utilisateur l'annote par 2 mots-clés. L'extension automatique par le modèle est utilisée pour ajouter d'autres mots-clés. L'utilisateur intervient de nouveau pour corriger les fausses étiquettes. Cette nouvelle image avec ses mots-clés est ajoutée à l'ensemble d'apprentissage pour la mise à jour des paramètres du modèle. Le processus est répété avec d'autres images jusqu'à atteindre un nombre donné d'images d'apprentissage.

Les étapes du retour utilisateur résultat de la combinaison, comme illustré dans la figure 4.10, sont les suivantes :

1. L'apprentissage actif est utilisé pour sélectionner une image à annoter à partir de l'ensemble non annoté ;
2. L'utilisateur annote l'image sélectionnée par 2 mots-clés ;
3. L'apprentissage dans l'apprentissage est utilisé pour ajouter d'autres mots-clés ;
4. L'utilisateur intervient pour corriger les faux mots-clés de l'image sélectionnée ;
5. L'apprentissage incrémental est utilisé pour la mise à jour des paramètres du modèle ;

6. La nouvelle image est ajoutée à l'ensemble d'apprentissage annoté.

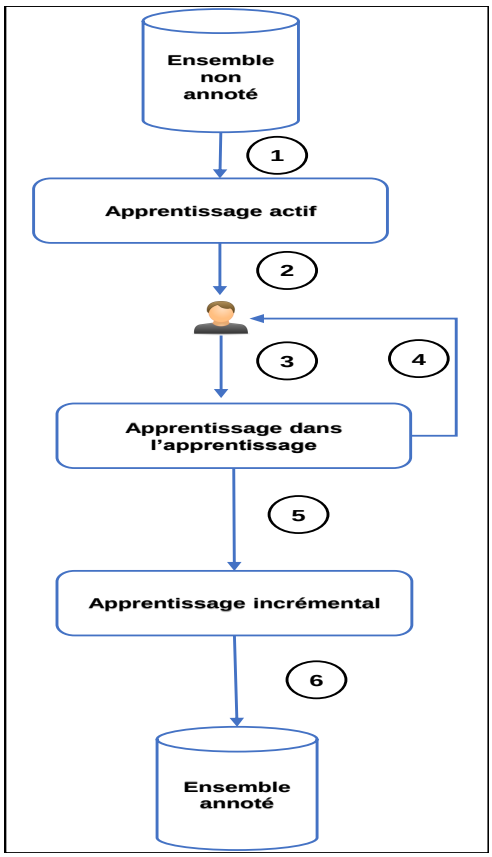


FIGURE 4.10 – Combinaison des trois retours utilisateurs

4.5.2 Évaluations

Pour évaluer la combinaison des trois utilisateurs, nous avons refait la même expérience présentée dans le tableau 4.2 avec 80% des données d'apprentissage. Nous avons obtenu les résultats présentés dans le nouveau tableaux 4.5. En utilisant 80% des données d'apprentissage avec le premier retour utilisateur (apprentissage dans l'apprentissage), nous avons obtenu 15.82%, 16.31% et 14.69% de gain en effort manuel respectivement sur les trois bases Corel-5K, ESP-Game et IAPRTC-12.

TABLE 4.5 – Gain en effort manuel avec 80% d'exemples d'apprentissage

	Corel-5K	ESP-Game	IAPRTC-12
3^{eme} mot-clé	354/2955	1712/12230	1579/13171
4^{eme} mot-clé	404/1835	2024/10664	1951/10852
Gain en effort manuel	15.82 %	16,31 %	14,69 %

En intégrant l'apprentissage actif et l'apprentissage incrémental, nous avons obtenu les résultats représentés dans le tableau 4.6. Par exemple pour la base Corel-5K, en ajoutant automatiquement deux mots-clés pour 10% d'images d'apprentissage sélectionnées par l'apprentissage actif, nous avons obtenu 47 mots-clés corrects sur les 235 mots-clés ajoutés. Ainsi, nous gagnons $741/2502 = 29.62\%$ d'effort manuel en total. La même expérience est effectuée pour les deux autres bases de données ESP-Game et IAPRTC-12 où nous atteignons un gain de $26,33\%$ et de $29,92\%$ respectivement. En moyenne sur les trois bases, nous avons gagné environ 13% d'effort manuel de plus en combinant les trois retours utilisateurs.

TABLE 4.6 – Gain en effort manuel après combinaison

Ensemble d'apprentissage %	Corel-5K	ESP-Game	IAPRTC-12
10	47/235	193/1073	263/1142
20	79/328	286/1360	416/1599
30	80/297	361/1503	405/1447
40	108/360	428/1646	375/1295
50	85/266	341/1216	543/1752
60	129/391	519/1789	609/1904
70	96/281	399/1288	466/1371
80	117/344	488/1574	570/1676
Total	741/2502	3015/11449	3647/12186
Gain en effort manuel	29,62 %	26,33 %	29,92 %

Dans les trois retours utilisateurs que nous avons présentés pour notre modèle, l'utilisateur intervient pour corriger l'annotation de certaines images ou annoter complètement les images. Les retours utilisateurs améliorent les performances d'annotation dans les trois types d'apprentissage et réduisent l'effort manuel dans l'apprentissage actif et l'apprentissage dans l'apprentissage. En effet, l'ensemble d'apprentissage dans l'apprentissage incrémental croit progressivement au fur et à mesure que l'utilisateur corrige les fausses annotations dans un processus de recherche d'images, ce qui permet d'améliorer les performances d'annotation. Dans l'apprentissage dans l'apprentissage, le modèle est utilisé pour générer un ensemble d'apprentissage, ce qui permet de minimiser l'effort manuel d'annotation. Dans l'apprentissage actif, les exemples les plus informatifs sont choisis pour améliorer la qualité de l'apprentissage de sorte à minimiser l'effort manuel d'annotation fourni par un utilisateur. La combinaison de ces algorithmes permet de mieux réduire l'effort manuel et de gagner environ 13% de plus.

4.6 Conclusion

Dans ce chapitre, nous avons intégré des retours utilisateurs dans notre modèle pour améliorer les performances et minimiser le laborieux effort manuel d'annotation. Les retours utilisateurs ont été effectués en utilisant l'apprentissage dans l'apprentissage, l'apprentissage incrémental et l'apprentissage actif. Dans le premier, l'utilisateur améliore la qualité de l'ensemble d'apprentissage en collaborant avec le modèle pour réduire l'effort fastidieux de l'annotation manuelle. Au deuxième niveau, lors d'une étape de recherche d'images, les données sont progressivement introduites dans le système pour améliorer l'ensemble d'apprentissage et ainsi améliorer les performances d'annotation. Au troisième niveau, l'apprentissage actif est utilisé pour réduire le nombre d'instances en choisissant l'exemple conduisant à la plus grande réduction du taux d'erreur. Nous avons aussi combiné ces trois retours utilisateurs, ce qui nous a permis de mieux réduire l'effort manuel. Le chapitre suivant sera consacré à intégrer une hiérarchie sémantique dans notre modèle pour améliorer et enrichir l'annotation.

Chapitre 5

Annotation d'images en utilisant une hiérarchie sémantique

Sommaire

5.1 Introduction

La plupart des approches d'annotation automatiques d'images sont basées sur la formulation d'une fonction de correspondance entre les caractéristiques de bas niveau, ou de niveau intermédiaire, et les concepts sémantiques en utilisant des techniques d'apprentissage automatique. Cependant, la seule utilisation des algorithmes d'apprentissage semble être insuffisante pour combler le problème bien connu du fossé sémantique [FZQ12, THA12], et donc pour produire des systèmes efficaces pour l'annotation automatique d'images. En effet, dans la plupart des approches d'annotation d'images, la sémantique est presque limitée à sa manifestation perceptuelle à travers l'apprentissage d'une fonction de correspondance associant les caractéristiques de bas niveau à des caractéristiques visuels de plus haut niveau sémantique. Les performances de ces approches dépend de la taille du vocabulaire et de la nature des données utilisées. L'utilisation de connaissances structurées, comme les hiérarchies sémantiques et les ontologies, semble être un bon moyen pour améliorer ces approches. Ces structures de connaissances permettent de modéliser de nombreuses relations sémantiques entre les mots-clés d'annotation. Ces relations se sont avérées être d'une importance primordiale pour la compréhension de la

sémantique d'images [THA12]. En outre, l'utilisation de ces modèles à base de connaissances structurées permet de réduire la complexité du problème de l'annotation d'images à grande échelle. Récemment, plusieurs travaux se sont intéressés à l'utilisation des hiérarchies sémantiques pour annoter des images. Ces structures peuvent être classées, comme mentionnée dans [BC13, THA12], en trois catégories principales :

1. Hiérarchies textuelles : basées uniquement sur les caractéristiques textuelles des images ;
2. Hiérarchies visuelles : basées uniquement sur les caractéristiques visuelles des images ;
3. Hiérarchies visuo-textuelles : basées à la fois sur les caractéristiques textuelles et visuelles.

Les hiérarchies textuelles sont des hiérarchies conceptuelles construites en utilisant une mesure de similarité entre les concepts. Plusieurs approches se sont basées sur la base lexicale WordNet [Mil95] pour la construction des hiérarchies conceptuelles [DDS+09, MS07, ATY+97, DGLW07, LS06, LSFF09, WML04]. Marszalek et al. [MS07] ont proposé un classificateur de hiérarchie basé sur WordNet. Leur hiérarchie est construite en extrayant le sous-graphe pertinent à partir de Wordnet en reliant tous les concepts du vocabulaire d'annotation. La hiérarchie obtenue est ensuite utilisée dans un ensemble de classificateurs hiérarchiques pour l'annotation des images. Kurtz et al. [KR14] ont proposé une approche de recherche d'images similaires dans une base de données qui combine une distance hiérarchique et une mesure de similarité ontologique. Dans cette approche, l'image requête est annotée par des termes sémantiques issues d'une ontologie. Un vecteur d'éléments modélisant cette image est construit. Ensuite, l'image est comparée aux autres images de la base en comparant leurs vecteurs représentatifs. Les vecteurs sont comparés en utilisant la distance HSBD (Hierarchical Semantic-Based Distance) [Kur12]. Bien que les approches de cette catégorie exploitent une représentation de connaissances pour fournir une annotation plus riche, ils ignorent l'information visuelle qui est une partie importante dans la tâche d'annotation d'images.

Les hiérarchies visuelles utilisent des caractéristiques visuelles de bas niveau plutôt que l'organisation conceptuelle du vocabulaire d'annotation. Dans ces hiérarchies, les images similaires sont généralement représentées dans les nœuds et les mots du vocabulaire dans les feuilles. Elles permettent d'accélérer et d'améliorer les performances de classification et de recherche. Bart et al. [BPPW08] ont proposé une méthode bayésienne pour trouver une taxonomie telle qu'une image soit générée à partir d'un chemin dans l'arbre. Les images similaires ont beaucoup de nœuds communs sur leurs chemins associés et donc une courte distance entre nœuds. Griffin et al. [GP08] ont construit une hiérarchie pour une classification plus rapide. Ils ont classifié d'abord les images pour estimer une matrice de confusion. Ils ont regroupé des catégories confuses de manière ascendante. Ensuite, ils ont construit également une hiérarchie descendante pour la comparaison, en divisant successivement les catégories. Les deux hiérarchies ont montré des résultats similaires pour la vitesse et la précision. Dans [SRZ+08], le modèle d'Allocation Dirichlet Latente hiérarchique (hLDA) a été utilisé pour regrouper les objets dans une hiérarchie visuelle suivant leurs similarités visuelles. Les hiérarchies de cette catégorie peuvent être utilisées pour la tâche de la classification hiérarchique afin d'accélérer et améliorer la classification. Cependant, elles présentent un problème majeur qui est la difficulté d'interprétation sémantique puisqu'elles sont basées uniquement sur des caractéristiques visuelles.

Les hiérarchies textuelles et visuelles ont résolu plusieurs problèmes en regroupant des objets dans des structures bien définies. Elles permettent d'augmenter la vitesse, la précision et ré-

duire la complexité des systèmes [THA12]. Elles ne sont pas adéquates en annotation d'images. En effet, la sémantique conceptuelle n'est pas toujours cohérente avec la visuelle des images, et est alors insuffisante pour construire de bonnes structures sémantiques pour bien annoter des images [WHY+12]. La sémantique visuelle toute seule ne peut pas conduire à une hiérarchie sémantique significative puisqu'elle est difficile à interpréter sémantiquement [BC13]. Il est nécessaire d'utiliser ces deux informations ensemble pour obtenir des hiérarchies sémantiques bien adaptées à la tâche d'annotation d'images. Bannour et al. [BC13] ont proposé une nouvelle approche pour la construction automatique des hiérarchies sémantiques qui peuvent être utilisées pour la classification et l'annotation d'images. Cette méthode est basée sur l'utilisation de trois mesure de similarité : visuelle, conceptuelle et contextuelle. Zhiming et al. [QZC16] se sont concentrés sur l'annotation d'images en deux niveaux en intégrant à la fois des caractéristiques visuelles globales et locales avec des hiérarchies sémantiques. Les deux niveaux d'annotation pour une image donnée sont la classification de sa scène et l'étiquetage des objets représentant ses régions. Pour la classification de la scène, plusieurs scènes spécifiques qui décrivent la plupart des cas des données ont été définies au préalable. Ensuite, les machines à vecteurs de support (SVM) sur la base des caractéristiques globales sont utilisées pour classifier la scène. Pour l'étiquetage des objets, un ensemble de noms abstraits en conformité avec Wordnet est défini. Puis, un nouveau modèle à base de champ aléatoire conditionnel est utilisé pour exploiter les corrélations scène-objet et objet-objet.

De manière générale, l'utilisation des hiérarchies sémantiques est de plus en plus intéressante pour la tâche d'annotation d'images. Elles permettent une meilleure performance lorsque le nombre de catégories augmente. Cependant, la construction d'une hiérarchie sémantique à partir d'un vocabulaire donné est complexe [LSLW12].

Dans ce chapitre, nous présentons notre méthode de construction d'une taxonomie sémantique à partir des mots-clés d'un vocabulaire donné. Cette taxonomie est basée sur l'utilisation d'informations visuelles, sémantiques et contextuelles. Nous montrons aussi que l'intégration de cette taxonomie dans notre modèle d'annotation permet d'augmenter les performances d'annotation et enrichir le vocabulaire utilisé.

5.2 Construction de la taxonomie

Une taxonomie est une collection de termes de vocabulaire organisés en une structure hiérarchique. Chaque terme d'une taxonomie se trouve dans une ou plusieurs relations parent-enfant avec d'autres termes de la taxonomie. Il peut y avoir différents types de relations parent-enfant dans une taxonomie (par exemple, partie entière, genre-espèce, type-instance), mais une bonne pratique limite toutes les relations parent-enfant à un seul parent pour être du même type. Récemment, de nombreux travaux ont été consacrés à la création automatique d'une ontologie ou d'une taxonomie propre à un domaine [SJN06, NVF11, FL12, LSLW12]. La construction d'une taxonomie manuelle est un processus laborieux, et la taxonomie résultante est souvent très subjective, comparée aux taxonomies construites par des approches basées sur les données. En outre, les approches automatiques ont le potentiel de permettre aux humains ou même aux machines de comprendre un domaine très ciblé et potentiellement évolutif. Cependant, le problème de l'induction d'une taxonomie à partir d'un ensemble de mots-clés présente un défi majeur [LSLW12]. Bien que l'utilisation d'un ensemble de mots-clés nous permet de caracté-

riser plus précisément un domaine spécifique, l'ensemble de mots-clés lui-même ne contient pas de relations explicites à partir desquelles une taxonomie peut être construite. Une façon de surmonter ce problème consiste à enrichir le vocabulaire d'annotation en ajoutant des nouveaux mots-clés. Liu et al. [LSLW12] ont présenté une nouvelle approche qui peut automatiquement dériver une taxonomie dépendante d'un domaine à partir d'un ensemble de mots-clés en exploitant à la fois une base de connaissances générale et une recherche par mot-clé. Pour enrichir le vocabulaire, ils ont utilisé la technique de conceptualisation en extrayant des informations contextuelles d'un moteur de recherche. Ils ont lancé des recherches en utilisant les mots-clés du vocabulaire comme requêtes et ont analysé les résultats de recherche pour en déduire les informations contextuelles partagées par les mots-clés. La taxonomie est ensuite construite par classification hiérarchique des mots-clés à l'aide d'un arbre bayésien planaire enraciné « Bayesian rose tree » [BTH12].

Dans la suite de cette section, nous présenterons les trois types d'informations utilisées ainsi que notre méthode de construction d'une taxonomie à partir d'un ensemble de mots-clés.

5.2.1 Information sémantique

L'information sémantique reflète la signification sémantique d'un mot-clé donné d'un point de vue linguistique. Il s'agit de chercher une représentation vectorielle qui caractérise la signification linguistique d'un mot-clé donné. De nombreux algorithmes d'apprentissage automatique sont incapables de traiter le texte dans son forme brute. Ils ont besoin de chiffres comme entrée pour effectuer n'importe quel type de travail, que ce soit la classification, la régression, etc. Avec l'énorme quantité de données présentes dans le format texte, il est impératif d'extraire des connaissances et de créer des applications. C'est pourquoi des méthodes de plongement de mots « Word Embedding » ont fait l'apparition. Ces méthodes tentent généralement de représenter un mot appartenant à un dictionnaire par un vecteur de nombres réels. Plusieurs stratégies ont été proposées pour le « Word Embedding » tels que les vecteurs TF-IDF et les vecteurs de co-occurrence. Cependant, ces méthodes se sont révélées limitées dans leurs représentations jusqu'à ce que Mitolov et al. [MCCD13] ont introduit word2vec dans la communauté de traitement du langage naturel. Word2vec est un groupe de modèles associés utilisés pour produire des plongements de mots. Ces modèles sont des réseaux neuronaux à deux couches, peu profonds, formés pour reconstruire les contextes linguistiques des mots. Il prend en entrée un grand corpus de texte et produit un espace vectoriel, typiquement de plusieurs centaines de dimensions, avec pour chaque mot unique du corpus un vecteur correspondant dans l'espace. Les vecteurs de mots sont positionnés dans l'espace vectoriel de sorte que les mots qui partagent des contextes communs dans le corpus sont situés à proximité les uns des autres dans l'espace.

Le modèle word2vec et ses applications ont récemment attiré beaucoup d'attention dans la communauté de l'apprentissage automatique. Il a été démontré que ces représentations vectorielles denses de mots appris par word2vec portaient des significations sémantiques et sont utiles dans un large éventail de cas d'utilisation allant du traitement de langage naturel à l'analyse de sentiments [GVD+17]. L'idée principale de ce modèle, appelée hypothèse distributionnelle, est que des mots similaires apparaissent dans des contextes similaires de mots autour d'eux. Il existe deux principaux modèles word2vec : Continuous Bag of Words (CBOW) et Skip-Gram.

5.2.1.1 Continuous bag-of-words

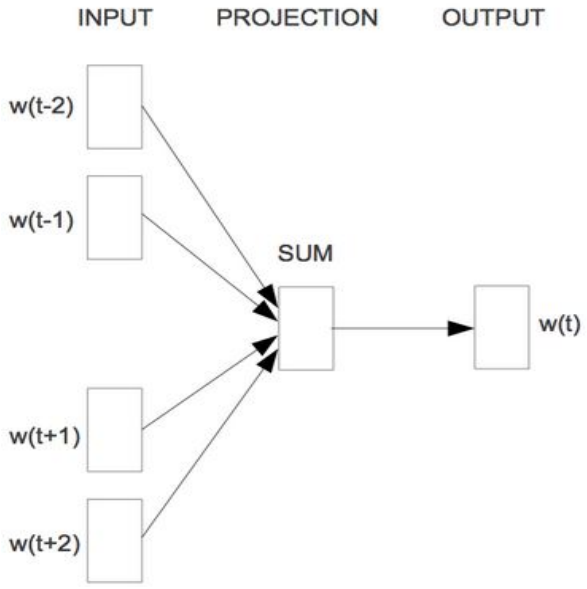


FIGURE 5.1 – Continuous bag-of-words [MCCD13]

Mikolov et al. [MCCD13] ont utilisé, comme représenté sur la figure 5.1, les n mots avant et après le mot cible w_t pour le prédire. Ils appellent ce modèle sac-de-mots continu (CBOW), car il utilise des représentations continues dont l'ordre est sans importance. Ce modèle présente l'avantage de ne pas avoir besoin d'une grande quantité mémoire. Cependant, il a l'inconvénient de prendre la moyenne du contexte d'un mot. Par exemple, le mot « Apple » peut être à la fois un fruit et une entreprise.

5.2.1.2 Skip-gram

Au lieu d'utiliser les mots environnants pour prédire le mot central comme CBOW, skip-gram [MCCD13] utilise le mot central pour prédire les mots environnants comme on peut le voir sur la figure 5.2. Le modèle skip-gram est une architecture neurale qui analyse un grand corpus et apprend à prédire les mots environnants avec un mot courant. Pendant la formation, ce modèle apprend des représentations vectorielles pour chaque mot, qui codent des informations sémantiques. Les mots de sens similaire auront des vecteurs proches l'un de l'autre dans l'espace vectoriel. Ces vecteurs de mots capturent d'autres relations sémantiques intéressantes qui sont cohérentes avec les opérations arithmétiques. Le modèle Skip-Gram peut capturer deux sémantiques pour un seul mot. Par exemple, il y aura deux représentations vectorielles pour le mot « Apple », une pour l'entreprise et l'autre pour le fruit.

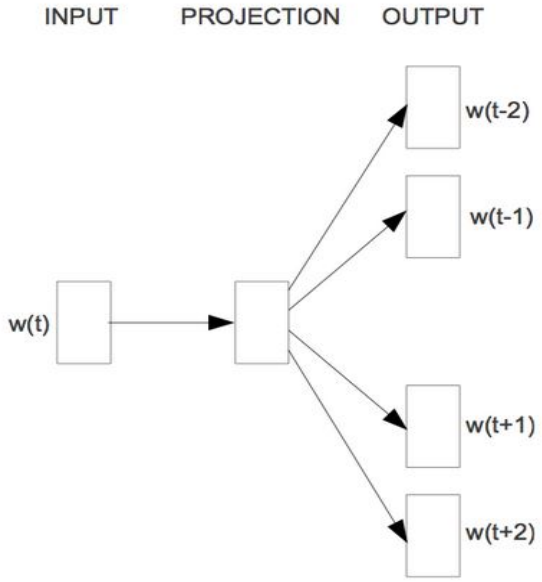


FIGURE 5.2 – Skip-gram [MCCD13]

5.2.2 Information visuelle

L'information visuelle reflète l'apparence visuelle d'un mot-clé donné dans les images annotées par ce mot-clé. Il s'agit donc de trouver une représentation vectorielle qui permet de caractériser cette apparence dans les images d'apprentissage. Nous avons utilisé la technique d'appariement (« image matching ») des images en utilisant les détecteurs et descripteurs SIFT [Low99]. En effet, La détection de caractéristiques et l'appariement d'images représentent deux tâches importantes dans le domaine de la vision par ordinateur, l'infographie, et certaines applications d'images. La première étape du processus d'appariement d'images implique la détection de points d'intérêt dans une image. Les détecteurs de caractéristiques sont utilisés pour trouver des points d'intérêt. Une fois les points d'intérêt extraits d'une image, un descripteur est généré pour chaque point décrivant les propriétés de son voisinage. Les points trouvés sont appariés ensuite en minimisant la distance entre leurs descripteurs. Le descripteur SIFT a été largement utilisé pour la tâche d'appariement d'images [MS05]. SIFT combine un détecteur de régions invariant à l'échelle et un descripteur basé sur la distribution du gradient dans les régions détectées. Le descripteur est représenté par un histogramme 3D des emplacements et des orientations du gradient.

Pour un mot-clé donné Kw_i, un ensemble d'images R_{Kw_i} est sélectionné à partie de l'ensemble d'apprentissage T de taille n. Toutes les images de l'ensemble R_{Kw_i} doivent être annotées par Kw_i. Ainsi,

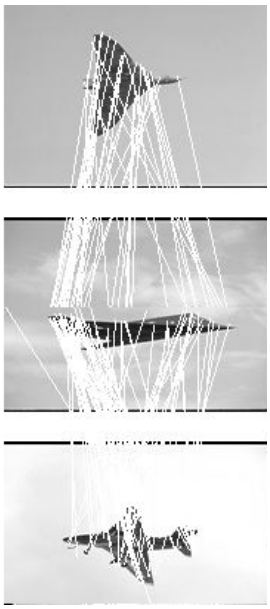
$$R_{Kw_i} = \bigcup_{1 \leq j \leq n} \{I_j\} \ / Kw_i \in W_{I_j} \tag{5.1}$$

W_{I_j} représente l'ensemble des mots-clés annotant l'image I_j dans T. Pour chaque image de l'ensemble R_{Kw_i}, des points d'intérêt sont détectés en utilisant les détecteurs SIFT. Pour chaque point d'intérêt trouvé, un descripteur SIFT est calculé. Les images sont appariés ensuite deux

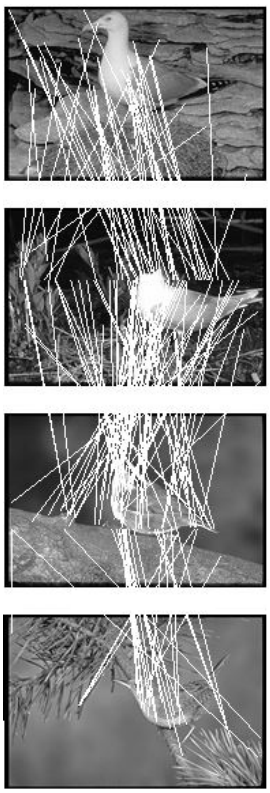
à deux et le résultat de cet appariement est pris comme information visuelle représentant le mot-clé Kw_i. Ainsi, l'information visuelle d'un mot-clé Kw_i, notée par $Vis(Kw_i)$, est définie par :

$$Vis(Kw_i) = \bigcap matching(SIFT(I_i), SIFT(I_j)) \ \forall \ I_i, I_j \in R_{Kw_i} \tag{5.2}$$

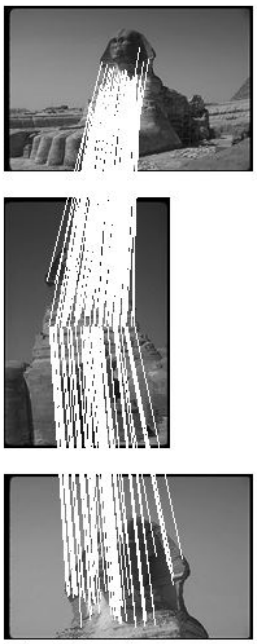
La figure 5.3 représente trois exemples de calcul de l'information visuelle pour 3 mots-clés de la base Corel-5K. L'image (a) représente l'appariement de 3 images annotées par le mot-clé « jet ». L'image (b) représente l'appariement de 4 images annotées par le mot-clé « bird ». L'image (c) représente l'appariement de 3 images annotées par le mot-clé « statue ». Par exemple pour l'image (a), nous observons clairement que l'intersection des appariements entre les 3 images est la région représentant l'objet « jet ». Ainsi, en cherchant l'intersection des descripteurs SIFT des trois images, nous obtenons une information visuelle qui caractérise bien le mot-clé « jet » dans la base d'apprentissage.



(a) matching des images annotées par « jet »



(b) matching des images annotées par « bird »



(c) matching des images annotées par « statue »

FIGURE 5.3 – Appariement d'images pour calculer l'information visuelle

5.2.3 Information contextuelle

Puisque les objets du monde réel ont tendance à exister dans un contexte, l'incorporation d'informations contextuelles est importante pour aider à comprendre la sémantique de l'image. Les informations contextuelles permettent de déterminer le contexte d'apparition des mots-clés en reliant celles qui apparaissent souvent ensemble dans l'annotation d'images même s'ils sont distants visuellement ou sémantiquement. Par exemple, les deux mots-clés « cheval » et « herbe » peuvent annoter ensemble une image pour représenter une scène naturelle, alors qu'il n'ont ni similarité visuelle ni similarité sémantique puisque « cheval » appartient à la famille des animaux et « herbe » appartient à la famille des plantes. Une méthode simple pour représenter l'information contextuelle est de chercher la fréquence de co-occurrence d'un paire de mots-clés. Cette information dépend uniquement du vocabulaire d'annotation utilisé. Par exemple, le paire de mots-clés « sea, sky » apparaît plus fréquemment que la paire de mots-clés « sea, sand » dans la base de données Corel-5k.

Nous utilisons l'information mutuelle pour caractériser l'information contextuelle. Cette mesure a été utilisée dans [BC13]. Soit Kw_i et Kw_j deux mots-clés du vocabulaire d'annotation. L'information contextuelle de Kw_i et Kw_j, notée par $cont(Kw_i, Kw_j)$, est définie par :

$$cont(Kw_i, Kw_j) = log\frac{P(Kw_i, Kw_j)}{P(Kw_i)P(Kw_j)} \tag{5.3}$$

$P(Kw_i)$ représente la probabilité du mot-clé Kw_i dans les images de la base de données. $P(Kw_i, Kw_j)$ représente la probabilité jointe d'apparition des deux mots-clés Kw_i et Kw_j ensemble.

Soit :

- V : la taille du vocabulaire d'annotation
- T : la taille de la base d'images
- n_i : le nombre d'images annotées par le mot-clé Kw_i
- n_{ij} : le nombre d'images annotées simultanément par les deux mots-clés Kw_i et Kw_j

Nous obtenons alors :

$$P(Kw_i) = \frac{n_i}{T}$$

$$P(Kw_i, Kw_j) = \frac{n_{ij}}{T}$$

L'information contextuelle de Kw_i et Kw_j devient donc :

$$cont(Kw_i, Kw_j) = log\frac{T * n_{ij}}{n_i * n_j} \tag{5.4}$$

Si les deux mots-clés Kw_i et Kw_j sont indépendants, nous avons :

$$P(Kw_i, Kw_j) = P(Kw_i) * P(Kw_j)$$

et donc

$$cont(Kw_i, Kw_j) = 0$$

En calculant l'information contextuelle de chaque mot-clé Kw_i avec tous les autres mot-clés du vocabulaire, nous obtenons une matrice symétrique dont chaque ligne (colonne) représente un mot-clé de l'ensemble de données. Chaque ligne (colonne) est de taille V caractérisant la quantité d'informations partagée avec les autres mots-clés du vocabulaire.

5.2.4 Méthode proposée

Une fois que nous avons estimé les informations visuelles, contextuelles et sémantiques pour chaque mot-clé du vocabulaire, il est important de les regrouper dans une taxonomie sémantique en se basant sur ces informations. Les trois informations sont utilisées ensemble sous forme d'un seul vecteur de caractéristiques pour la construction de la taxonomie.

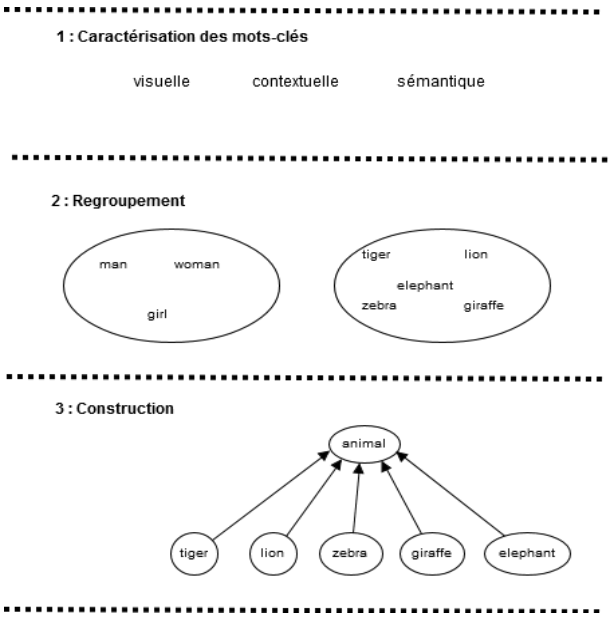


FIGURE 5.4 – Les étapes de construction de la taxonomie

La construction de la taxonomie, comme illustrée dans la figure 5.5, est divisée en 3 étapes principales :

1. Caractérisation : il s'agit de calculer les informations sémantiques, visuelles et contextuelles définies dans les sections 5.2.1, 5.2.2 et 5.2.3 pour chaque mot-clé du vocabulaire. Un vecteur qui caractérise chaque mot-clé est définit en concaténant les trois types d'informations.
2. Regroupement : il s'agit de regrouper les mots-clés les plus proches suivant les informations définies dans un groupe sémantique. Nous avons utilisé l'algorithme traditionnel de regroupement kmeans avec les vecteurs caractéristiques des mots-clés pour les regrouper dans K groupes.
3. Construction : il s'agit de construire une hiérarchie pour chaque groupe sémantique trouvé dans l'étape précédente. Tout d'abord, un nouveau mot-clé est ajouté pour chacun des K groupes. Ce nouveau mot-clé représente le concept ou la famille partagés par tous les

mots-clés du groupe. Ensuite, des arcs sont ajoutés entre tous les mots-clés du groupe et le nouveau mot-clé ajouté. Ces arcs représentent la relation parent-enfant entre les mots-clés du groupe (enfants) et le nouveau mot-clé ajouté (parent).

Les 3 étapes précédentes sont répétées en utilisant l'algorithme 3 dans un processus ascendant jusqu'à atteindre le nœud racine de la hiérarchie. Cet algorithme représente les 3 étapes de la figure 5.5 de façon détaillée. En effet, la première étape de la figure 5.5 est illustrée par les itérations 2 à 5, la deuxième étape est illustrée par l'itération 6. Les itérations 7 à 10 représentent algorithmiquement la troisième étape de la figure 5.5.

Algorithm 3 : Algorithme de construction de la taxonomie

Entrées : Ensemble d'images d'apprentissage, Vocabulaire d'annotation V

Sorties : Taxonomie, Nouveau vocabulaire d'annotation New_V

1 : $T = V$; $New\_V = V$;

Répéter

pour Chaque mot-clé Kw_i dans T **faire**

2 : Calculer l'information visuelle $Vis(Kw_i)$;

3 : Calculer l'information contextuelle $Cont(Kw_i)$;

4 : Calculer l'information sémantique $Sem(Kw_i)$;

5 : $Vec\_caract(Kw_i) = concat(Vis(Kw_i), Cont(Kw_i), Sem(Kw_i))$;

fin pour

6 : Utiliser k-means pour regrouper les mots-clés de T en K groupes ;

pour i=1 :K **faire**

7 : Ajouter un nouveau mot-clé P_i comme parent du groupe i ;

8 : Ajouter un arc entre les mots-clés du groupe i et le nouveau mot-clé P_i ;

9 : $New\_V = New\_V + P_i$;

10 : $T = T - Kw_j + P_i$, $\forall\ Kw_j \in$ groupe i

fin pour

Jusqu'à $|T| = 1$

Cet algorithme prend en entrée le vocabulaire d'annotation V et les images d'apprentissage qui seront utilisées pour calculer les informations visuelles et contextuelles. La sortie de cet algorithme est un nouveau vocabulaire d'annotation New_V et une taxonomie construite avec les mots-clés de New_V. Les nœuds de la taxonomie construite représentent les mots-clés du nouveau vocabulaire et les arcs représentent les relations sémantiques entre elles.

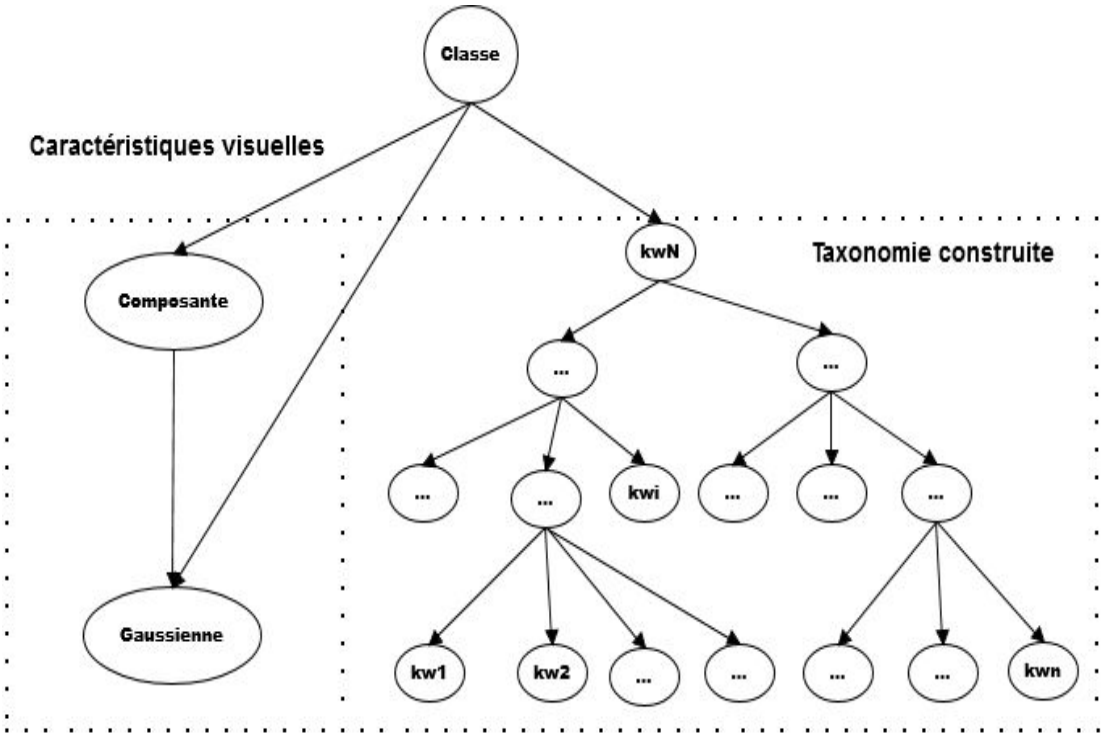


FIGURE 5.5 – Intégration de la taxonomie dans le modèle d'annotation

5.3 Modèle d'annotation en utilisant la taxonomie

Une fois la taxonomie construite, elle est intégrée dans le modèle comme illustré dans la figure 5.5. La structure obtenue est presque la même que celle présentée dans la section 3.4 du chapitre 3 sauf que les nœuds modélisant les caractéristiques textuelles sont remplacés par la structure de la taxonomie obtenue. Les caractéristiques visuelles d'une image donnée sont considérées comme des variables continues. Elles suivent une loi dont la fonction de densité est une densité de mélange de Gaussiennes. Elles sont modélisées par les deux nœuds : *Gaussienne* et *Composante*. Les caractéristiques textuelles d'une image donnée sont modélisées par les nœuds de la taxonomie construite. Chaque nœud est représentée par une variable aléatoire discrète qui suit une distribution de Bernoulli. Cette variable prend deux valeurs possibles : 0 et 1. La valeur 1 prise par la variable représentant le nœud kw_i indique que l'image est annotée par le mot-clé i du vocabulaire New_V et la valeur 0 indique l'absence de ce mot-clé dans l'annotation de l'image. Le modèle obtenu est donc un mélange de distributions de Bernoulli et de mélanges de Gaussiennes. L'avantage de ce modèle par rapport à l'ancien modèle est de pouvoir représenter les relations sémantiques entre les mots-clés.

Étant donnée une nouvelle image Im_i représentée par ses caractéristiques visuelles $VC_1, ..., VC_M$ et ses mots-clés existants $Kw_1, ..., Kw_n$, nous pouvons utiliser l'algorithme d'arbre de jonction [LS88], comme pour l'ancien modèle, pour étendre l'annotation de cette image avec de nouveaux mots-clés. Nous pouvons donc calculer la probabilité *a posteriori*

$$P(Kw_i|I_i) = P(Kw_i|VC_1, ..., VC_M, Kw_1, ..., Kw_n) \, \forall Kw_i \ \in New\_V \tag{5.5}$$

Nous pouvons également calculer la probabilité *a posteriori*

$$P(C_i|I_i) = P(C_i|VC_1, ..., VC_M, Kw_1, ..., Kw_n) \tag{5.6}$$

pour classifier l'image. Cette dernière est affectée à la classe C_i maximisant la probabilité définie dans l'équation 5.6.

La plupart des méthodes d'annotation automatique des images supposent une longueur d'annotation fixe k (5 généralement) pour chaque image. Ainsi, pour annoter une image avec k mots-clés avec notre modèle, les k mots avec la plus grande probabilité $P(Kw_i|I_i)$ sont sélectionnés comme annotation. Toutefois, l'annotation de longueur fixe peut entraîner des annotations insuffisantes ou trop longues. Lorsque la longueur est petite, il est probable qu'un certain contenu de l'image n'est pas capturé par l'annotation. En revanche, lorsque la longueur est longue, il est probable que les annotations générées contiennent des mots qui ne sont pas pertinents pour le contenu. Ces deux cas ne sont pas souhaitable pour supporter efficacement la recherche dans les bases de données d'images. Par exemple, les mots non pertinents résultant d'une annotation excessive peuvent dégrader gravement la précision de la recherche d'images. Ainsi, pour résoudre ce problème, nous pouvons définir un seuil λ sur la probabilité d'un mot-clé. Dans ce cas, nous aurons des annotations de différentes tailles. Une image ne sera annotée par un mot-clé Kw_i si et seulement si :

$$P(Kw_i|I_i) > \lambda \tag{5.7}$$

5.4 Évaluations

Dans cette section, nous présentons l'évaluation de notre modèle après l'intégration de la hiérarchie sémantique sur la base Corel-5K. Concernant les informations sémantiques, nous avons utilisé le modèle Word2vec pré-entrainé sur le corpus Google News [14]. La longueur de chaque vecteur obtenu par le modèle est de 300 caractéristiques. Ainsi, pour chaque mot-clé, nous obtenons un vecteur de taille 300 qui permet de coder suffisamment d'informations sémantiques pour décrire le mot-clé correspondant. Une visualisation en deux dimensions de ces vecteurs en utilisant t-SNE est représentée dans la figure 5.6. La méthode t-SNE [MH08] est une technique de réduction de la dimensionnalité qui peut être utilisée pour visualiser des vecteurs de grandes dimensions, tels que les « word embeddings ». Cette visualisation nous permet d'identifier les distances sémantiques entre les mots-clés. Par exemple, en effectuant un zoom (figure 5.7) sur une partie de la figure 5.6, nous remarquons que les mots-clés « buddha », « buddhist », « bengal » et « indian » sont très proches sémantiquement. Concernant le choix des images de l'ensemble R_{Kw_i}, nous avons préféré les images annotées par le minimum possible de mots-clés (y compris Kw_i) tout en définissant un seuil (6 images) sur la taille de cet ensemble. Ce choix nous permet de garantir que les images sélectionnées ne représentent pas plusieurs objets et que l'appariement de ces images représente bien l'objet décrit par Kw_i.

14. https ://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz

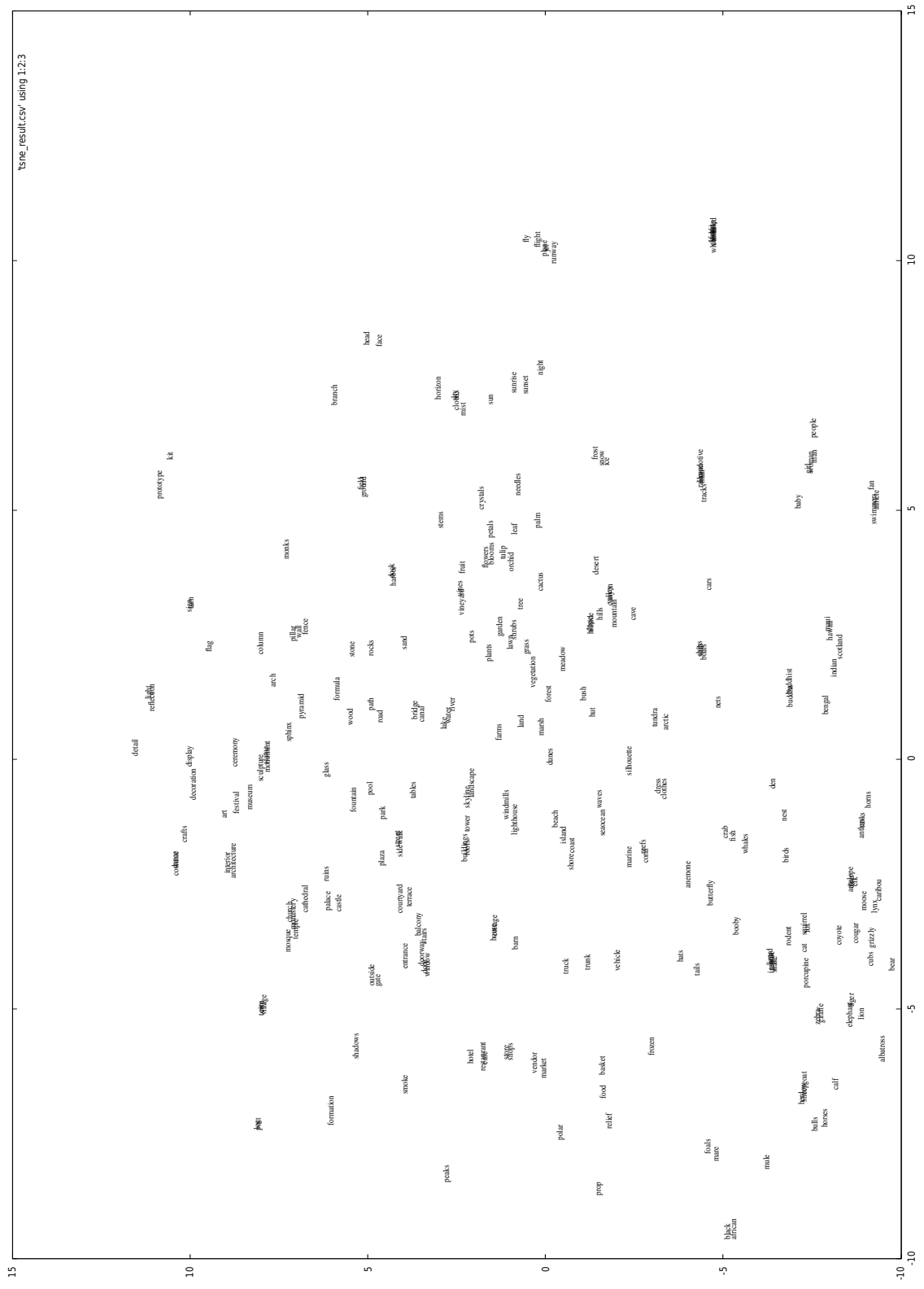


FIGURE 5.6 – Visualisation en deux dimensions des « word embeddings » du vocabulaire d'annotation de Corel-5K

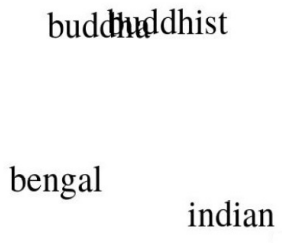


FIGURE 5.7 – Zoom sur les mots-clés « buddha », « buddhist », « bengal » et « indian »

En utilisant les informations visuelles, contextuelles et sémantiques, nous avons regroupé les 260 mots-clés du vocabulaire d'annotation de la base Corel-5k en 30 groupes suivant l'algorithme 3. Nous avons fait ce choix pour ne pas avoir une hiérarchie trop profonde ce qui agit sur la complexité du modèle. Pour chaque groupe, un nouveau mot-clé est ajouté comme parent des membres du groupe. Le parent doit décrire le concept sémantique partagé par l'ensemble du groupe. Pour ajouter automatiquement les parents des groupes, nous avons essayé d'utiliser le modèle word2vec. Nous avons utilisé le centre du groupe généré par Kmeans comme entrée du modèle pour chercher le mot le plus proche correspondant à ce vecteur. Les résultats obtenus n'étaient pas acceptables. Par exemple, pour le groupe contenant les mots-clés (« people », « baby », « man », « woman », et « girl »), les mots-clés les plus proches obtenus sont : « boy », « teenage_girl », « teenager », « mother », « toddler » et « newborn_baby ». Ces mots-clés obtenus peuvent être ajoutés comme membres du même groupe mais pas comme parent du groupe parce qu'ils ne représentent pas un concept partagé par les membres. Nous avons aussi essayé d'utiliser Wordnet en cherchant l'ancêtre commun le plus proche entre les membre du groupe mais dans la majorité des cas nous avons obtenu la racine de Wordnet. Par exemple, pour le groupe contenant les mots-clés (« bear », « cat », « tiger », « fox », « lion », « elephant », « iguana », « lizard », « giraffe », « zebra », « rodent », « squirrel » et « porcupine »), le résultat obtenu est « entity » alors que le concept partagé le plus logique est « animal ». Nous avons choisi donc d'ajouter les nouveaux mots-clés manuellement même si c'est une tâche difficile puisqu'il faut analyser la signification sémantique de tous les membres du même groupe pour en déduire leur parent. Dans les figures 5.8 à 5.11, nous avons représenté quelques résultats du regroupement par kmeans. Par exemple, la figure 5.8 représente le groupe contenant les mots-clés (« people », « fan », « athlete », « swimmers », « baby », « man », « woman », et « girl »). Pour ce groupe, nous avons choisi le mot-clé « human » comme parent. Les 30 nouveaux mots-clés obtenus du regroupement ont été à leur tour regroupés en 7 nouveaux groupes. Le résultat de ce deuxième regroupement est représenté dans la figure 5.12. Par exemple, les 6 mots-clés « herbivorous animal », « various animal », « smal animal », « big animal », « wild animal », et « predator » ont été regroupés dans le même groupe auquel nous avons ajouté le nouveau mot-clé « animal » comme parent. Ainsi, partant d'un vocabulaire de 260 mots-clés, nous avons obtenu un nouveau vocabulaire de 298 mots-clés (37 nouveaux mots-clés + racine) organisé sous forme d'une taxonomie. Cette dernière permet de représenter les relations sémantiques entre les mots-clés.

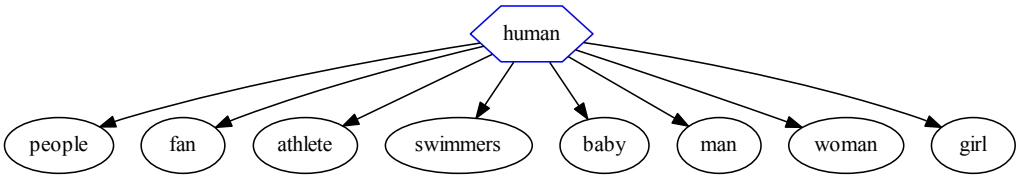


FIGURE 5.8 – Représentation graphique du groupe « human »

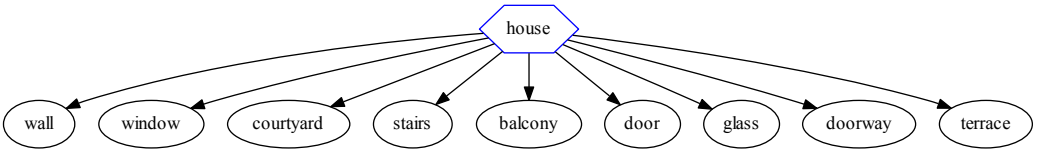


FIGURE 5.9 – Représentation graphique du groupe « house »

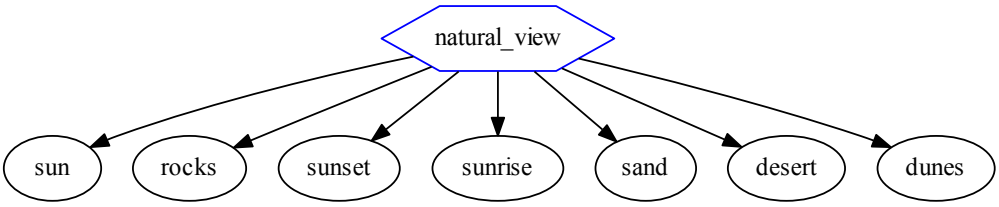


FIGURE 5.10 – Représentation graphique du groupe « natural_view »

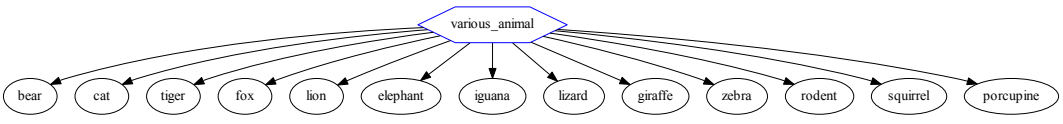


FIGURE 5.11 – Représentation graphique du groupe « various_animal »

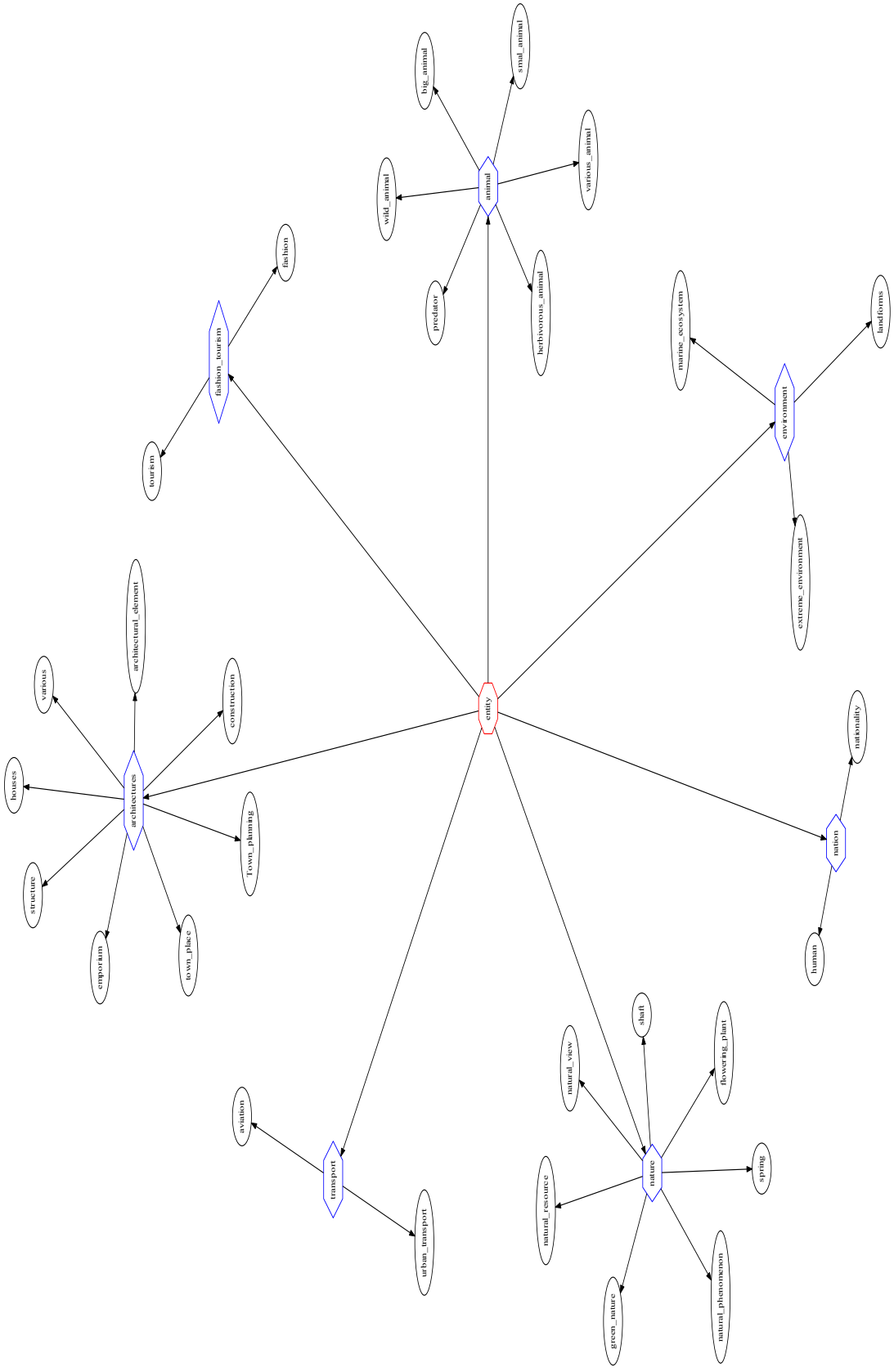


FIGURE 5.12 – Représentation graphique des 37 nouveaux mots-clés ajoutés dans la hiérarchie

TABLE 5.1 – Performance de notre modèle avec hiérarchie sémantique par rapport aux autres modèles d'annotation de l'état de l'art sur l'ensembles de données Corel-5k

Méthode	Corel-5K			
	P	R	$F1$	$N+$
CRM [LMJ04]	16	19	17	107
MBRM [FML04]	24	25	25	122
SKL-CRM [ML14]	39	46	42	184
SVM-DMBRM [MCM14]	36	48	41	197
JEC [MPK10]	27	32	29	139
TagProp [GMVS09]	33	42	37	160
NMF-KNN [KIS14]	38	**56**	45	150
2PKNN [VJ12]	44	46	45	191
CNN-R [MMM15]	32	41	37	166
HHD [MSCM16]	31	49	38	194
GS [ZHH$^+$10]	30	33	31	146
MSC [WYZZ09]	25	32	28	136
MLDL [JWL$^+$16]	**45**	49	**47**	**198**
SLED [CZG$^+$15]	35	51	42	196
SML [CCMV07]	23	29	26	137
RF [FZQ12]	29	40	34	157
RFC-PSO[EBKHS14]	26	22	24	109
Fuzzy[MY16]	27	32	29	–
Corr-LDA [CBL09]	21	36	27	131
GMM-Mult [BT08]	27	38	32	154
Notre méthode sans HS	34	45	39	175
Notre méthode avec HS	42	47	44	182

Dans le tableau 5.1, nous avons repris les deux premières colonne du tableau 3.1 (chapitre 3) auquel nous avons ajouté les performances de notre modèle après intégration de la hiérarchie construite (dernière ligne). Dans ce tableau, nous avons annoté automatiquement chaque image dans la base de test (Corel-5K) par 5 mots-clés et nous avons calculé le rappel, la précision, la mesure F1 et la mesure N+. Nous remarquons que l'intégration de la hiérarchie sémantique construite dans le modèle augmente considérablement les performances d'annotations et surtout au niveau de la précision. En effet, nous avons obtenu une précision de 34% avec l'ancien modèle (« Notre méthode sans HS » dans le tableau) et après l'intégration de la hiérarchie sémantique, nous arrivons à une précision de 42% (« Notre méthode avec HS » dans le tableau).

Cette augmentation de performances nous a permis de dépasser quelques méthodes de l'état de l'art au niveau de la mesure F1 tels que [ML14] et [CZG+15]. En plus des avantages de notre modèle cités dans le chapitre 3, un autre avantage s'ajoute qui est le fait de pouvoir enrichir l'annotation en utilisant des nouveaux mots-clés qui n'appartenaient pas au vocabulaire d'annotation initial.

La figure 5.13 illustre l'annotation de quelques images de la base Corel-5k où les étiquettes de la vérité terrain sont données. Nous remarquons clairement que les images ne sont pas annotées par le même nombre de mots-clés à cause de l'utilisation d'un seuil λ défini expérimentalement à 0.75. Nous remarquons aussi qu'il y a apparition de nouveaux mots-clés n'appartenant pas au vocabulaire initial. Par exemple, la quatrième image est annotée manuellement par trois mots-clés (« water », « boats » et « bridge »), sept nouveaux mots-clés (« arch », « pyramid », « natural_resource », « town », « structure », « architectures » et « nature ») sont automatiquement ajoutés après l'extension automatique. Les deux mots-clés (« arch » et « pyramid ») appartiennent au vocabulaire d'annotation initial et les autres cinq mots-clés appartiennent au nouveau vocabulaire ajouté. Cette image est annotée par le mot-clé « natural_resource » parce qu'elle est annotée par « water » qui appartient au groupe de parent « natural_resource ». Le mot-clé « town » est apparu parce qu'il est parent de « bridge ». Le mot-clé « structure » est apparu parce qu'il est parent de « arch » et « pyramid ». De même pour les deux mots-clés « architectures » et « nature » qui sont parents respectivement de « structure » et « natural_resource ».

5.5 Conclusion

Dans ce chapitre, nous avons présenté une méthode pour construire une hiérarchie sémantique à partir d'un ensemble de mots-clés. Cette hiérarchie se base sur l'utilisation de trois types d'informations pour chaque mot-clé : visuelles, contextuelles et sémantiques. L'information contextuelle permet de déterminer le contexte d'apparition d'un mot-clé dans l'ensemble d'annotation d'images d'apprentissage. L'information visuelle permet de décrire l'apparition visuelle d'un mot-clé donné dans les images d'apprentissage. L'information sémantique reflète la signification sémantique d'un mot-clé donné de point de vue linguistique. Après la construction de la hiérarchie, nous l'avons intégré dans notre modèle d'annotation d'images. Le modèle obtenu avec cette hiérarchie est un mélange de distributions de Bernoulli et de mélanges de Gaussiennes. L'intégration de la hiérarchie sémantique construite dans le modèle augmente considérablement les performances d'annotations comme montré dans la partie expérimentale. En outre, elle permet d'enrichir l'annotation d'images en utilisant de nouveaux mots-clés qui n'appartenaient pas au vocabulaire d'annotation initial. Dans le chapitre suivant, nous proposons une conclusion générale de ce mémoire de thèse et des futures directions de recherche.

 sky, sun, clouds, tree	 sky, jet, plane	 bear, polar, snow, tundra
sky, sun, clouds, tree, palm,natural_view, shaft, natural_phenomenon, nature	**sky, jet, plane, f-16, aviation, natural_view, transport, nature**	**bear, polar, snow, ice, various_animal, extreme_environment, animal**
 water, boats, bridge	 tree, horses, mare, foals	 sky, buildings, flag
water, boats, bridge, arch, pyramid, natural_resource, town, structure, architectures, nature	**tree, horses, mare, foals, field, herbivorous_animal, shaft, animal, nature**	**sky, buildings, skyline, street, architectural_element, natural_view, architectures, nature**
 flowers, house, garden	 plants	 mountain, sky, water, tree
flowers, house, garden, lawn, spring, flowering_plant, town_planning, nature	**plants, leaf, stems, green_nature, houses, shaft, nature**	**mountain, sky, water, landforms, natural_view, emporium, architectures, nature, environment**

FIGURE 5.13 – Exemples d'annotation d'images de la base Corel-5k après intégration de la hiérarchie sémantique

Chapitre 6

Conclusion et travaux futurs

Sommaire

6.1 Conclusion

Dans cette thèse, nous avons abordé le problème de l'extension d'annotation d'images. Après une étude approfondie de l'état de l'art (chapitre 2), nous avons opté pour les modèles graphiques probabilistes et plus précisément les réseaux bayésiens. En effet, ces derniers réunissent la théorie des probabilités et la théorie des graphes. Ils bénéficient donc des avantages de ces deux théories. Ils utilisent une structure graphique pour fournir une représentation explicite des informations et un calcul probabiliste pour représenter l'incertitude. En outre, un avantage des modèles graphiques probabilistes est de traiter le problème des données manquantes et d'offrir la possibilité de combiner plusieurs sources d'informations.

Nous avons présenté un modèle graphique probabiliste pour l'extension d'annotation des images (chapitre 3). Ce modèle est un mélange de distributions multinomiales et de mélanges de Gaussiennes qui combinent des caractéristiques visuelles et textuelles. Les résultats expérimentaux sur trois bases d'images ont montré que notre modèle est compétitif par rapport à des modèles de l'état de l'art. Notre modèle présente l'avantage de maintenir à la fois une précision élevée et un rappel élevé de manière équilibrée. Il a été utilisé aussi pour l'annotation et la classification des documents.

Le modèle proposé nécessite une étape préliminaire d'apprentissage où les images sont annotées manuellement. L'annotation manuelle est efficace puisqu'elle permet d'obtenir une description plus appropriée du contenu d'une image mais elle est très longue et coûteuse. Pour réduire ce coût et améliorer la qualité de l'annotation obtenue, nous avons intégré des retours utilisateurs dans notre modèle (chapitre 4). Les retours utilisateurs ont été effectués en utilisant l'apprentissage dans l'apprentissage, l'apprentissage incrémental et l'apprentissage actif. Dans le premier, l'utilisateur améliore la qualité de l'ensemble d'apprentissage en collaborant avec le modèle pour réduire l'annotation manuelle. Ce retour utilisateur nous a permis de gagner en moyenne 17% d'effort manuel sur trois bases d'images tests. Dans le deuxième, lors

d'une étape de recherche d'images, les données sont progressivement introduites dans le système pour améliorer l'ensemble d'apprentissage et ainsi améliorer les performances d'annotation. Dans le troisième, l'apprentissage actif est utilisé pour réduire le nombre d'instances en choisissant l'exemple conduisant à la plus grande réduction du taux d'erreur. Dans les trois retours utilisateurs présentés pour notre approche, l'utilisateur intervient pour corriger l'annotation de certaines images ou annoter complètement les images. En utilisant l'apprentissage actif, avec 50% des données apprentissage nous avons obtenu de meilleurs résultats qu'avec 80% des données apprentissage sans apprentissage actif. Les retours utilisateurs améliorent les performances d'annotation dans les trois types d'apprentissage et réduisent l'effort manuel dans l'apprentissage actif et l'apprentissage dans l'apprentissage. La combinaison de ces trois retours utilisateurs permet de mieux réduire l'effort manuel et de gagner environ 13% de plus.

L'apprentissage d'une fonction de correspondance associant les caractéristiques de bas niveau à des concepts visuels de plus haut niveau sémantique est insuffisante pour combler le problème du fossé sémantique, et donc pour produire des systèmes efficaces pour l'annotation automatique d'images. Pour résoudre ce problème, nous avons utilisé une hiérarchie sémantique (chapitre 5) pour améliorer notre modèle en modélisant de nombreuses relations sémantiques entre les mots-clés d'annotation. Nous avons donc proposé une méthode semi-automatique pour construire une hiérarchie sémantique à partie d'un ensemble de mots-clés. Cette hiérarchie se base sur l'utilisation de trois types d'informations pour chaque mot-clé : visuelle, contextuelle et sémantique. L'information contextuelle permet de déterminer le contexte d'apparition d'un mot-clé dans l'ensemble d'annotation d'images d'apprentissage. L'information visuelle permet de décrire l'apparition visuelle d'un mot-clé donné dans les images d'apprentissage. L'information sémantique reflète la signification sémantique d'un mot-clé donné d'un point de vue linguistique. Après la construction de la hiérarchie, nous l'avons intégré dans notre modèle d'annotation d'images. Le modèle obtenu avec la hiérarchie est un mélange de distributions de Bernoulli et de mélanges de Gaussiennes. L'intégration de la hiérarchie sémantique construite dans le modèle a augmenté considérablement les performances d'annotations. En outre, elle permet d'enrichir l'annotation d'images en utilisant de nouveaux mots-clés qui n'appartenaient pas au vocabulaire d'annotation initial.

6.2 Travaux futurs

Dans le futur, nous envisageons d'apporter les améliorations suivantes :

– Nous souhaitons automatiser la construction de la hiérarchie sémantique. En effet, la méthode que nous avons proposé est semi-automatique puisque le concept sémantique partagé par les membres d'un groupe est ajouté manuellement. Le nouveau concept peut être ajouté en utilisant la méthode proposée dans [LSLW12]. Cette méthode se base sur l'utilisation d'un moteur de recherche. Les mots-clés du groupe sont lancés comme requête dans un moteur de recherche et les pages Web trouvées sont analysées pour trouver le concept partagé entre les requêtes.
– Nous souhaitons également utiliser des caractéristiques visuelles plus pertinentes en employant les fonctionnalités CNN comme entrée du réseau bayésien. La combinaison des CNN avec les réseaux bayésiens a été utilisée dans les systèmes proposés dans [SPS+17, TW16]. En effet, il y a eu une explosion médiatique pour les architectures de réseaux à

neurones profonds pour la vision par ordinateur ces dernières années. Ces architectures permettent d'apprendre de plus en plus de fonctionnalités abstraites qui permettent de mieux représenter l'image originale.

– Nous souhaitons expérimenter notre modèle sur des corpus plus larges en nombre d'images et en taille du vocabulaire tel que ImageNet [DDS+09] qui contient plus de 14 millions images et plus de 21 milles concepts.

Annexe A

Exemple d'un réseau bayésien

Dans cet exemple, nous présentons le réseau de diagnostic médical "ASIA" introduit par Lauritzen et Spiegelhalter [LS88]. Dans ce réseau, le fait d'être fumeur (F) ou non a une influence sur la possibilité d'avoir une bronchite (B) ou un cancer (C) . La présence ou l'absence des deux maladies Cancer (C) et tuberculoses (T) ont une influence sur la possibilité d'avoir un examen radio (R). Ce modèle permet de décrire un domaine de connaissances médicales (diagnostic médical).

La structure de ce réseau est décrite dans la figure A.1. Elle est constitué de 8 nœuds et 8 arcs. Les variables qui représentent les nœuds sont les suivants :

- V : Voyage en Asie
- F : Fumeur
- T : Tuberculose
- C : Cancer du poumon
- B : Bronchite
- R : Résultat de la radio
- D : Dyspnée
- X : Rayons X

Les valeurs possibles pour les variables sont représentées dans le tableau A.1.

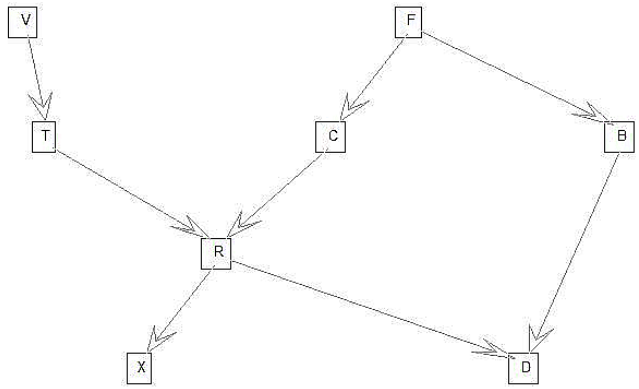


FIGURE A.1 – Le réseau bayésien ASIA

TABLE A.1 – Valeurs possibles des variables du réseau ASIA

Variable	valeurs
V	Oui/Non
F	Oui/Non
T	Oui/Non
C	Oui/Non
B	Oui/Non
R	Normale/Anormale
D	Oui/Non
X	Positif/Négatif

A.1 Apprentissage de structure

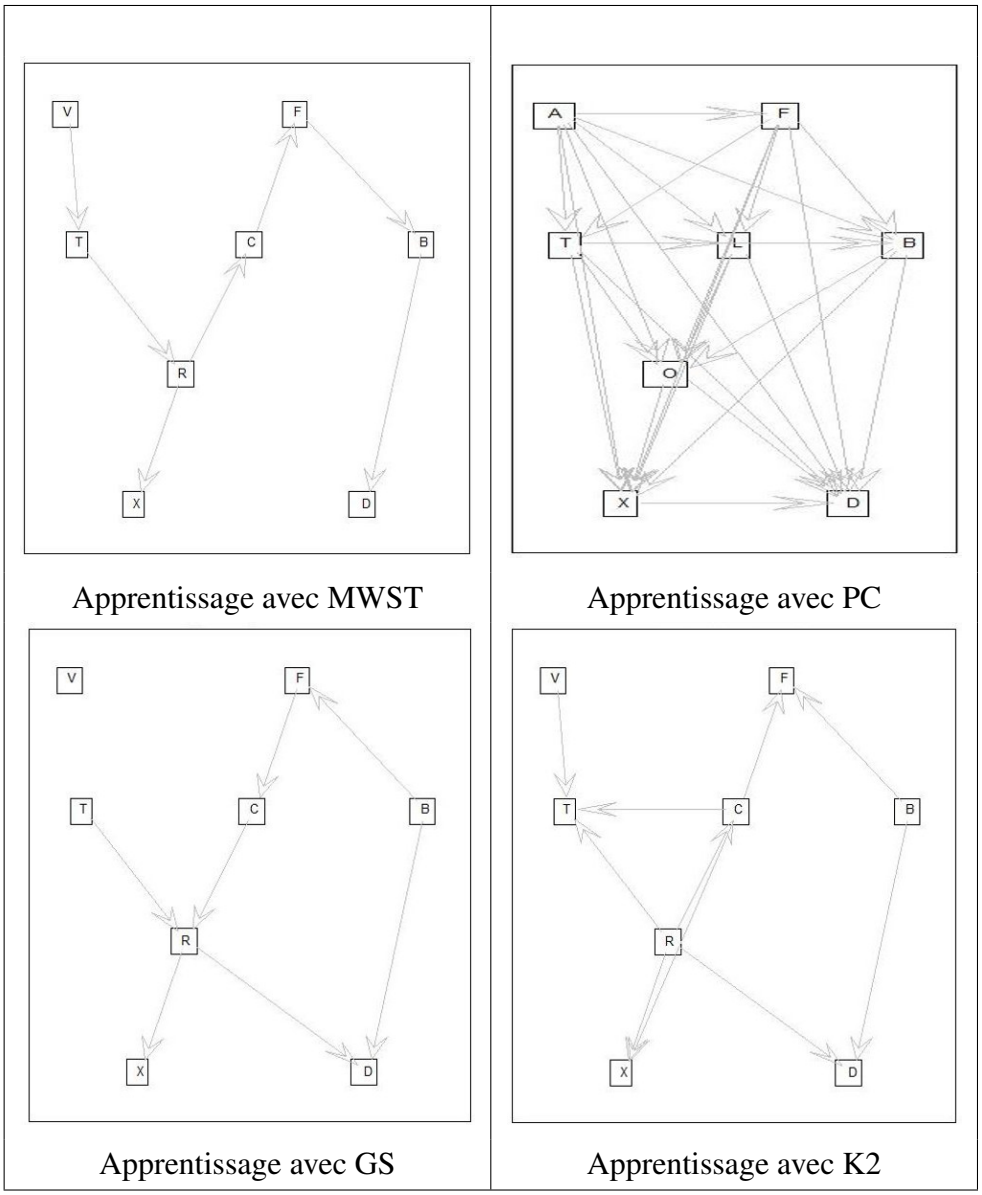


FIGURE A.2 – Apprentissage de struture du réseau ASIA avec différentes méthodes

Nous avons utilisé la boîte à outil Bayes Net Toolbox [Mur04] pour l'apprentissage des paramètres et la la boîte à outil Structure Learning Package [LF04] pour l'apprentissage de la structure. Nous avons utilisé aussi la méthode de [FL07] pour générer une base de 1000 exemples. La structure du réseau ASIA représentée dans la figure A.1 est donnée par un expert. Toutefois, nous pouvons déterminer cette structure par un apprentissage automatique en utilisant la base d'exemple générée. Les structures obtenues avec les algorithmes d'apprentissage de structures PC [SGS00], GS [CGH95], K2 [CH92] et MWST [Gei92] sont représentées dans la figure A.2.

A.2 Apprentissage des paramètres

Après la définition de la structure du réseau bayésien, il faut trouver les tables de probabilités conditionnelles associées aux variables. C'est ce qu'on appelle l'apprentissage de paramètres. Nous allons commencer par une initialisation aléatoire des paramètres et nous les déterminons ensuite par un apprentissage automatique.

Les tables de probabilités conditionnelles (TPC) associées aux variables (V, F, T, C, B, R, X et D) sont données respectivement dans les tableaux (A.2, A.3, A.4, A.5, A.6, A.7, A.8 et A.9).

TABLE A.2 – TPC(V) sans apprentissage

Oui	Non
0.01	0.99

TABLE A.3 – TPC(F) sans apprentissage

Oui	Non
0.5	0.5

TABLE A.4 – TPC(T) sans apprentissage

P(T|V)	Oui	Non
Oui	0.05	0.01
Non	0.95	0.99

TABLE A.5 – TPC(C) sans apprentissage

P(C|F)	Oui	Non
Oui	0.10	0.01
Non	0.90	0.99

TABLE A.6 – TPC(B) sans apprentissage

P(B|F)	Oui	Non
Oui	0.6	0.3
Non	0.4	0.7

TABLE A.7 – TPC(R) sans apprentissage

P(R|T,C)	Non, Non	Non, Oui	Oui, Non	Oui, Oui
Anormale	1.0	0.0	0.0	0.0
Normale	0.0	1.0	1.0	1.0

TABLE A.8 – TPC(X) sans apprentissage

P(X|R)	Normale	Anormale
Positif	0.98	0.05
Négatif	0.02	0.95

TABLE A.9 – TPC(D) sans apprentissage

P(D|B,R)	Non, Non	Non, Oui	Oui, Non	Oui, Oui
Non	0.9	0.2	0.3	0.1
Oui	0.1	0.8	0.7	0.9

Dans le tableau A.2, la variable V prend seulement 2 valeurs possibles (oui ou non) puisqu'elle ne possède pas de parents. La variable T prend 2 valeurs (oui ou non) et elle possède la variable V comment parent qui, à son tour, prend 2 valeurs possibles. C'est pourquoi nous avons 4 états possibles dans le tableau A.4 qui représentent $p(T = oui|V = oui)$, $p(T = oui|V = Non)$, $p(T = Non|V = oui)$ et $p(T = Non|V = Non)$.

Pour l'apprentissage des paramètres, nous avons utilisé l'algorithme EM [DLR77] et la base générée précédemment. Les tables de probabilités conditionnelles (TPC) associées aux variables (V, F, T, C, B, R, X et D) après apprentissage automatique sont données respectivement dans les tableaux (A.10, A.11, A.12, A.13, A.14, A.15, A.16 et A.17).

TABLE A.10 – TPC(V) avec apprentissage

Oui	Non
0.0125	0.9875

TABLE A.11 – TPC(F) avec apprentissage

Oui	Non
0.5139	0.4861

TABLE A.12 – TPC(T) avec apprentissage

P(T|V)	Oui	Non
Oui	0.9933	0.0009
Non	0.0067	0.9991

TABLE A.13 – TPC(C) avec apprentissage

P(T|V)	Oui	Non
Oui	0.0231	0.0111
Non	0.9769	0.9889

TABLE A.14 – TPC(B) avec apprentissage

P(C|F)	Oui	Non
Oui	0.6854	0.2601
Non	0.3146	0.7399

TABLE A.15 – TPC(R) avec apprentissage

P(R|T,C)	Non, Non	Non, Oui	Oui, Non	Oui, Oui
Anormale	0.9998	0.0000	0.0002	0.000
Normale	0.0002	1.0000	0.9998	1.0000

TABLE A.16 – TPC(X) avec apprentissage

P(X|R)	Normale	Anormale
Positif	0.9474	0.0685
Négatif	0.0526	0.9315

TABLE A.17 – TPC(D) avec apprentissage

P(D|B,R)	Non, Non	Non, Oui	Oui, Non	Oui, Oui
Non	0.8805	0.1714	0.2038	0.1056
Oui	0.1195	0.8286	0.7962	0.8944

A.3 Inférence

Après la définition du réseau bayésien (structure et paramètres), nous pouvons prévoir l'état probable de certains attributs après avoir observé l'état d'un ensemble de descripteurs. C'est ce qu'on appelle inférence. Par exemple, après avoir observé qu'un individu est fumeur $(F = oui)$ et a visité l'Asie $(V = oui)$, on peut calculer la probabilité d'avoir un cancer des poumons. $(F = oui, V = oui)$ s'appelle une évidence. En injectant l'évidence $(F = oui, V = oui)$ dans le réseau, nous obtenons $P(C = oui) = 0.5069$ et $P(C = non) = 0.4931$.

Annexe B

Publications dans le cadre de la thèse

- Bouzaieni, A. (2014, March). Réseaux Bayésiens et quelques applications en traitement d'images. In CORIA-CIFED (pp. 377-380).

- Bouzaieni, A., Barrat, S. and Tabbone, S. (2015, August). Automatic annotation extension and classification of documents using a probabilistic graphical model. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on (pp. 316-320). IEEE.

- Bouzaieni, A., Tabbone, S. and Barrat, S. (2015, September). Automatic Images Annotation Extension Using a Probabilistic Graphical Model. In International Conference on Computer Analysis of Images and Patterns (pp. 579-590). Springer International Publishing.

- Bouzaieni, A., Barrat, S. and Tabbone, S. (2016, March). Extension automatique d'annotation et classification de documents en utilisant un modèle graphique probabiliste. In CORIA-CIFED (pp. 564-574).

- Bouzaieni, A., Barrat, S. and Tabbone, S. (2016, June). Extension automatique d'annotation d'images en utilisant un modèle graphique probabiliste. Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes.

- Bouzaieni, A. and Tabbone, S. (2017, September). Images Annotation Extension Based on User Feedback. In International Conference on Advanced Concepts for Intelligent Vision Systems (pp. 418-430). Springer, Cham.

Bibliographie

[ALLQ13] Nasullah Khalid Alham, Maozhen Li, Yang Liu, and Man Qi. A mapreduce-based distributed svm ensemble for scalable image classification and annotation. *Computers & Mathematics with Applications*, 66(10) :1920–1934, 2013.

[ATY+97] Y Alp Aslandogan, Chuck Thier, Clement T Yu, Jon Zou, and Naphtali Rishe. Using semantic contents and wordnet in image retrieval. In *ACM SIGIR Forum*, volume 31, pages 286–295. ACM, 1997.

[BBB16] Mohamed-Rafik Bouguelia, Yolande Belaïd, and Abdel Belaïd. An adaptive streaming active learning strategy based on instance weighting. *Pattern Recognition Letters*, 70 :38–44, 2016.

[BBUDB15] Lamberto Ballan, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Data-driven approaches for social image and video tagging. *Multimedia Tools and Applications*, 74(4) :1443–1468, 2015.

[BC13] Hichem Bannour and Hudelot Céline. Construction de hiérarchies sémantiques pour l'annotation d'images. *Revue des Sciences et Technologies de l'Information-Série RIA : Revue d'Intelligence Artificielle*, 27(1) :11–37, 2013.

[BDF+03] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb) :1107–1135, 2003.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3) :346–359, June 2008.

[BJ15] Priyam Bakliwal and CV Jawahar. Active learning based image annotation. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2015 Fifth National Conference on*, pages 1–4. IEEE, 2015.

[BL09] Byungki Byun and Chin-Hui Lee. An incremental learning framework combining sample confidence and discrimination with an application to automatic image annotation. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1441–1444. IEEE, 2009.

[Bor14] Eugene Borovikov. A survey of modern optical character recognition techniques. *arXiv preprint arXiv :1412.4183*, 2014.

[BPPW08] Evgeniy Bart, Ian Porteous, Pietro Perona, and Max Welling. Unsupervised learning of visual taxonomies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[BT08] Sabine Barrat and Salvatore Tabbone. Classification and automatic annotation extension of images using bayesian network. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 937–946, 2008.

[BT10] Sabine Barrat and Salvatore Tabbone. Modeling, classifying and annotating weakly annotated images using bayesian network. *Journal of Visual Communication and Image Representation*, 21(4) :355–363, 2010.

[BTH12] Charles Blundell, Yee Whye Teh, and Katherine A Heller. Bayesian rose trees. *arXiv preprint arXiv :1203.3468*, 2012.

[CB07] Nawei Chen and Dorothea Blostein. A survey of document image classification : Problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition*, 10(1) :1–16, 2007.

[CBL97] Jie Cheng, David A Bell, and Weiru Liu. Learning belief networks from data : An information theory based approach. In *Proceedings of the sixth international conference on Information and knowledge management*, pages 325–331. ACM, 1997.

[CBL09] Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

[CCL06] Ajay Chakravarthy, Fabio Ciravegna, and Vitaveska Lanfranchi. Cross-media document annotation and enrichment. In *1st Semantic Web Authoring and Annotation Workshop (SAAW2006)*, 2006.

[CCL07] Bertrand Coüasnon, Jean Camillerapp, and Ivan Leplumey. Access by content to handwritten archive documents : generic document recognition method and platform for annotations. *International Journal of Document Analysis and Recognition*, 9 :223–242, 2007.

[CCMV07] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 29(3) :394–410, 2007.

[CCS04] C Cusano, G Ciocca, and R Schettini. Image annotation using svm, in proceedings of internet imaging iv, vol. SPIE, 2004.

[CD85] Gilles Celeux and Jean Diebolt. The sem algorithm : a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2(1) :73–82, 1985.

[CFCF10] Zenghai Chen, Hong Fu, Zheru Chi, and Dagan Feng. A neural network model with adaptive structure for image annotation. In *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*, pages 1865–1870. IEEE, 2010.

[CFCF12] Zenghai Chen, Hong Fu, Zheru Chi, and David Dagan Feng. An adaptive recognition model for image annotation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6) :1120–1127, 2012.

[CG92] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3) :315–332, 1992.

[CGH95] Do Chickering, Dan Geiger, and David Heckerman. Learning bayesian networks : Search methods and experimental results. In *proceedings of fifth conference on artificial intelligence and statistics*, pages 112–128, 1995.

[CGJ96] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[CH92] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4) :309–347, 1992.

[Chi13] Cheng-Chieh Chiang. Interactive tool for image annotation using a semi-supervised and hierarchical approach. *Computer Standards & Interfaces*, 35(1) :50–58, 2013.

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

[CK12] Dongjin Choi and Pankoo Kim. Automatic image annotation using semantic text analysis. *Multidisciplinary Research and Practice for Information Systems*, pages 479–487, 2012.

[CLBS11] Yue Cao, Xiabi Liu, Jie Bing, and Li Song. Using neural network to combine measures of word semantic similarity for image annotation. In *Information and Automation (ICIA), 2011 IEEE International Conference on*, pages 833–837. IEEE, 2011.

[CSB06] Nawei Chen, Hagit Shatkay, and Dorothea Blostein. Exploring a new space of features for document classification : figure clustering. In *Proceedings of the conference of the Center for Advanced Studies on Collaborative research*, pages 369–372, 2006.

[Csu17] Gabriela Csurka. Domain adaptation for visual applications : A comprehensive survey. *arXiv preprint arXiv :1702.05374*, 2017.

[CYC+11] X. Chen, X. T. Yuan, Q. Chen, S. Yan, and T. S. Chua. Multi-label visual classification with label exclusive context. In *2011 International Conference on Computer Vision*, pages 834–841, 2011.

[CZG+15] Xiaochun Cao, Hua Zhang, Xiaojie Guo, Si Liu, and Dan Meng. Sled : Semantic label embedding dictionary representation for multilabel image annotation. *IEEE Transactions on Image Processing*, 24(9) :2746–2759, 2015.

[DBdFF02] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation : Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pages 97–112. Springer, 2002.

[DCCO14] Benjamin Duthil, Mickaël Coustaty, Vincent Courboulay, and Jean Marc Ogier. Annotation sémantique de documents administratifs. In *EGC*, pages 47–52, 2014.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet : A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[DE95] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 150–157. The Morgan Kaufmann series in machine learning,(San Francisco, CA, USA), 1995.

[DEG+03] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A Tomlin, et al. Semtag and seeker : Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, pages 178–186. ACM, 2003.

[DFPSS07] Fabio Del Frate, Fabio Pacifici, Giovanni Schiavon, and Chiara Solimini. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4) :800–809, 2007.

[DGL+11] Stamatia Dasiopoulou, Eirini Giannakidou, Georgios Litos, Polyxeni Malasioti, and Yiannis Kompatsiaris. A survey of semantic image and video annotation tools. *Knowledge-driven multimedia information extraction and ontology evolution*, pages 196–239, 2011.

[DGLW07] Ritendra Datta, Weina Ge, Jia Li, and James Z Wang. Toward bridging the annotation-retrieval gap in image search. *IEEE MultiMedia*, 14(3) :24–35, 2007.

[Die93] Francisco Javier Diez. Parameter adjustment in bayes networks. the generalized noisy or-gate. In *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*, pages 99–105. Morgan Kaufmann Publishers Inc., 1993.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[DTXC09] Lixin Duan, Ivor W Tsang, Dong Xu, and Tat-Seng Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 289–296. ACM, 2009.

[DTXM09] Lixin Duan, Ivor W Tsang, Dong Xu, and Stephen J Maybank. Domain transfer svm for video concept detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1375–1381. IEEE, 2009.

[EBKHS14] Nashwa El-Bendary, Tai-hoon Kim, Aboul Ella Hassanien, and Mohamed Sami. Automatic image annotation approach based on optimization of classes scores. *Computing*, 96(5) :381–402, 2014.

[EHG+10] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding*, 114(4) :419–428, 2010.

[FKQ11] Motofumi Fukui, Noriji Kato, and Wenyuan Qi. Multi-class labeling improved by random forest for automatic image annotation. In *IAPR Conference on Machine Vision Applications*, pages 202–205, 2011.

[FL07] OCH François and P Leray. Incomplete datasets generation using the bayesian networks formalism. In *Proceedings of the International Joint Conferences on Neural Networks (IJCNN 2007), Orlando, Florida*, 2007.

[FL12] Trevor Fountain and Mirella Lapata. Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 466–476. Association for Computational Linguistics, 2012.

[FM13] Ali Fakhari and Amir Masoud Eftekhari Moghadam. Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Applied Soft Computing*, 13(2) :1292–1302, 2013.

[FML04] SL Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition*, volume 2, pages 1002–1009. IEEE, 2004.

[Fra06] Olivier François. *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes.* PhD thesis, INSA de Rouen, 2006.

[Fri98] Nir Friedman. The bayesian structural em algorithm. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 129–138. Morgan Kaufmann Publishers Inc., 1998.

[FZQ12] Hao Fu, Qian Zhang, and Guoping Qiu. Random forest for image annotation. In *European Conference on Computer Vision*, pages 86–99. Springer, 2012.

[Gar04] Lars Marius Garshol. Metadata ? thesauri ? taxonomies ? topic maps ! making sense of it all. *Journal of information science*, 30(4) :378–391, 2004.

[GCL05] K-S Goh, Edward Y Chang, and Beitao Li. Using one-class and two-class svms for multiclass image annotation. *IEEE Transactions on knowledge and data engineering*, 17(10) :1333–1346, 2005.

[GCT11] S. Gao, L. T. Chia, and I. W. H. Tsang. Multi-layer group sparse coding for concurrent image classification and annotation. In *CVPR 2011*, pages 2809–2816, 2011.

[GCTR14] S. Gao, L. T. Chia, I. W. H. Tsang, and Z. Ren. Concurrent single-label image classification and annotation via efficient multi-layer group sparse coding. *IEEE Transactions on Multimedia*, 16(3) :762–771, 2014.

[Gei92] Dan Geiger. An entropy-based learning algorithm of bayesian conditional trees. In *Proceedings of the Eighth international conference on Uncertainty in artificial intelligence*, pages 92–97. Morgan Kaufmann Publishers Inc., 1992.

[Gil01] D Gillies. Judea pearl causality : Models, reasoning, and inference. *British Journal for the Philosophy of Science*, 52(3) :613–622, 2001.

[GMVS09] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop : Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[GP08] Gregory Griffin and Pietro Perona. Learning and using taxonomies for fast visual categorization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[GPV13] Albert Gordo, Florent Perronnin, and Ernest Valveny. Large-scale document image retrieval and classification with runlength histograms and binary embeddings. *Pattern Recognition*, 46(7) :1898–1905, 2013.

[GTCZ10] S. Gao, I. W. H. Tsang, L. T. Chia, and P. Zhao. Local features are not lonely - laplacian sparse coding for image classification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3555–3561, 2010.

[GVD+17] Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69 :214–224, 2017.

[Han08] Allan Hanbury. A survey of methods for image annotation. *Journal of Visual Languages & Computing*, 19(5) :617–627, 2008.

[HTO99] Yasuhide Mori Hironobu, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *In MISRM'99 first international workshop on multimedia intelligent storage and retrieval management*, pages 405–409, 1999.

[HWWT14] Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification : A comprehensive study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3) :493–506, 2014.

[Jen96] Finn V Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[JHCZ09] Lixing Jiang, Jin Hou, Zeng Chen, and Dengsheng Zhang. Automatic image annotation based on decision tree machine learning. In *Cyber-Enabled Distributed Computing and Knowledge Discovery, 2009. CyberC'09. International Conference on*, pages 170–175. IEEE, 2009.

[JKWA05] Yohan Jin, Latifur Khan, Lei Wang, and Mamoun Awad. Image annotations by combining multiple evidence & wordnet. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 706–715. ACM, 2005.

[JLM03] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[JWL+16] Xiao-Yuan Jing, Fei Wu, Zhiqiang Li, Ruimin Hu, and David Zhang. Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing*, 25(6) :2712–2725, 2016.

[JZHJ14] Ping Ji, Na Zhao, Shijie Hao, and Jianguo Jiang. Automatic image annotation by semi-supervised manifold kernel density estimation. *Information Sciences*, 281 :648–660, 2014.

[KH02] Kazuhiro Kuroda and Masafumi Hagiwara. An image retrieval system by impression words and specific object names–iris. *Neurocomputing*, 43(1) :259–276, 2002.

[KHPG06] Aditya Kalyanpur, James Hendler, Bijan Parsia, and Jennifer Golbeck. Smore-semantic markup, ontology, and rdf editor. *Defense Technical Information Center*, 2006.

[KIS14] Mahdi M Kalayeh, Haroon Idrees, and Mubarak Shah. Nmf-knn : Image annotation using weighted multi-view non-negative matrix factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 184–191, 2014.

[KKPS02] José Kahan, M-R Koivunen, Eric Prud'Hommeaux, and Ralph R Swick. Annotea : an open rdf infrastructure for shared web annotations. *Computer Networks*, 39(5) :589–608, 2002.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1) :79–86, 1951.

[KL11] N. Kulkarni and B. Li. Discriminative affine sparse codes for image classification. In *CVPR 2011*, pages 1609–1616, 2011.

[KP06] Halina Kwasnicka and Mariusz Paradowski. On evaluation of image auto-annotation methods. In *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*, volume 2, pages 353–358. IEEE, 2006.

[KPK04] Sungyoung Kim, Sojung Park, and Minhwan Kim. Image classification into object/non-object classes. In *CIVR*, pages 393–400. Springer, 2004.

[KR14] Camille Kurtz and Daniel L Rubin. Utilisation de relations ontologiques pour la comparaison d'images décrites par des annotations sémantiques. In *EGC*, pages 609–614, 2014.

[Kur12] Camille Kurtz. Une distance hiérarchique basée sur la sémantique pour la comparaison d'histogrammes nominaux. In *EGC*, pages 77–88, 2012.

[KYD13] Jayant Kumar, Peng Ye, and David Doermann. Structural similarity for document image classification and retrieval. *Pattern Recognition Letters,* , pages 119–126, November 2013.

[LCZ+06] Xirong Li, Le Chen, Lei Zhang, Fuzong Lin, and Wei-Ying Ma. Image annotation by large-scale content-based image retrieval. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 607–610. ACM, 2006.

[LF04] Philippe Leray and Olivier Francois. Bnt structure learning package : Documentation and experiments. *Laboratoire PSI, Universitè et INSA de Rouen, Tech. Rep*, 2004.

[LG94] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[LHWW15] Zhiwu Lu, Peng Han, Liwei Wang, and Ji-Rong Wen. Semantic sparse recoding of visual content for image applications. *IEEE Transactions on Image Processing*, 24(1) :176–188, 2015.

[LJ97] Steffen L Lauritzen and Finn Verner Jensen. Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence*, 21(1) :51–69, 1997.

[LLL+09] Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. *Pattern recognition*, 42(2) :218–228, 2009.

[LLLC12] Xi Liu, Rujie Liu, Fei Li, and Qiong Cao. Graph-based dimensionality reduction for knn-based image annotation. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1253–1256. IEEE, 2012.

[LMJ04] Victor Lavrenko, R Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *Advances in neural information processing systems*, pages 553–560, 2004.

[LMVG14] Jinpeng Li, Harold Mouchère, and Christian Viard-Gaudin. An annotation assistance system using an unsupervised codebook composed of handwritten graphical multi-stroke symbols. *Pattern Recognition Letters*, 35 :46–57, 2014.

[Low99] David G Low. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[LS88] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[LS06] Tony Lam and Rahul Singh. Semantically relevant image retrieval by combining image and linguistic analysis. *Advances in Visual Computing*, pages 770–779, 2006.

[LSFF09] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding : Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE, 2009.

[LSLW12] Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1433–1441. ACM, 2012.

[LSW09] Xirong Li, Cees GM Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7) :1310–1322, 2009.

[LTCT14] Weifeng Liu, Dacheng Tao, Jun Cheng, and Yuanyan Tang. Multiview hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding*, 118 :50–60, 2014.

[LTS03] Ching-Yung Lin, Belle L Tseng, and John R Smith. Videoannex : Ibm mpeg-7 annotation tool for multimedia indexing and concept learning. In *IEEE International Conference on Multimedia and Expo*, pages 1–2, 2003.

[LW03] Jia Li and James Ze Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9) :1075–1088, 2003.

[LWL+07] Jing Liu, Bin Wang, Mingjing Li, Zhiwei Li, Weiying Ma, Hanqing Lu, and Songde Ma. Dual cross-media relevance model for image annotation. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 605–614. ACM, 2007.

[LYRZ10] Dong Liu, Shuicheng Yan, Yong Rui, and Hong-Jiang Zhang. Unified tag analysis with multi-edge graph. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 25–34. ACM, 2010.

[LZL08] Ying Liu, Dengsheng Zhang, and Guojun Lu. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8) :2554–2570, 2008.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013.

[MCM14] Venkatesh N Murthy, Ethem F Can, and R Manmatha. A hybrid model for automatic image annotation. In *Proceedings of International Conference on Multimedia Retrieval*, pages 369–376. ACM, 2014.

[MDGW07] Raphaël Marée, Marie Dumont, Pierre Geurts, and Louis Wehenkel. Random subwindows and randomized trees for image retrieval, classification, and annotation. *Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 6th European Conference on Computational Biology (ECCB)*, 2007.

[MGP04] Florent Monay and Daniel Gatica-Perez. Plsa-based image auto-annotation : constraining the latent space. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 348–351. ACM, 2004.

[MGPW05] Raphael Maree, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In *CVPR*, volume 1, pages 34–40. IEEE, 2005.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov) :2579–2605, 2008.

[Mil95] George A Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41, 1995.

[ML12] Julian McAuley and Jure Leskovec. Image labeling on a network : using social-network metadata for image classification. *Computer Vision–ECCV 2012*, pages 828–841, 2012.

[ML14] Sean Moran and Victor Lavrenko. Sparse kernel learning for image annotation. In *Proceedings of International Conference on Multimedia Retrieval*, page 113. ACM, 2014.

[MMM15] Venkatesh N Murthy, Subhransu Maji, and R Manmatha. Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 603–606. ACM, 2015.

[MN98] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pages 359–367. Citeseer, 1998.

[MPK10] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. *International Journal of Computer Vision*, 90(1) :88–105, 2010.

[MS05] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10) :1615–1630, 2005.

[MS07] Marcin Marszalek and Cordelia Schmid. Semantic hierarchies for visual object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[MSCM16] Venkatesh N Murthy, Avinash Sharma, Visesh Chari, and R Manmatha. Image annotation using multi-scale hypergraph heat diffusion framework. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 299–303. ACM, 2016.

[Mur04] K Murphy. Bayes net toolbox v5 for matlab, 2004.

[MVC11] Thomas Mensink, Jakob Verbeek, and Gabriela Csurka. Learning structured prediction models for interactive image labeling. In *CVPR*, pages 833–840. IEEE, 2011.

[MY16] Vafa Maihami and Farzin Yaghmaee. Fuzzy neighbor voting for automatic image annotation. *Journal of Electrical and Computer Engineering Innovations*, 4(1) :1–8, 2016.

[MYSTM05] Prem Melville, Stewart M Yang, Maytal Saar-Tsechansky, and Raymond Mooney. Active learning for probability estimation using jensen-shannon divergence. In *ECML*, pages 268–279. Springer, 2005.

[NHGT14] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4240, 2014.

[NLS+02] Milind R Naphade, Ching-Yung Lin, John R Smith, Belle L Tseng, and Sankar Basu. Learning to annotate video databases. In *Storage and Retrieval for Media Databases*, pages 264–275, 2002.

[NST+06] Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE multimedia*, 13(3) :86–91, 2006.

[NVF11] Roberto Navigli, Paola Velardi, and Stefano Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, volume 2, page 2, 2011.

[NWL+11] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. *Réseaux bayésiens*. Editions Eyrolles, 2011.

[OMF12] Mustapha Oujaoura, Brahim Minaoui, and Mohammed Fakir. Image annotation using moments and multilayer neural networks. *IJCA Special Issue on Software Engineering, Databases and Expert Systems SEDEX*, 1 :46–55, 2012.

[OPH96] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1) :51–59, 1996.

[OT01] Aude Oliva and Antonio Torralba. Modeling the shape of the scene : A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3) :145–175, 2001.

[PAN10] Duangmanee Putthividhy, Hagai T Attias, and Srikantan S Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415. IEEE, 2010.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, 2014.

[PLK04] Soo Beom Park, Jae Won Lee, and Sang Kyoon Kim. Content-based image classification using a neural network. *Pattern Recognition Letters*, 25(3) :287–300, 2004.

[PYFD04] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Gcap : Graph-based automatic image captioning. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 146–146. IEEE, 2004.

[QH07] Xiaojun Qi and Yutao Han. Incorporating multiple svms for automatic image annotation. *Pattern Recognition*, 40(2) :728–741, 2007.

[QZC16] Zhiming Qian, Ping Zhong, and Jia Chen. Integrating global and local visual features with semantic hierarchies for two-level image annotation. *Neurocomputing*, 171 :1167–1174, 2016.

[RLF12] Michael Rubinstein, Ce Liu, and William T Freeman. Annotation propagation in large image databases via dense image correspondence. In *European Conference on Computer Vision*, pages 85–99. Springer, 2012.

[RLL+07] Xiaoguang Rui, Mingjing Li, Zhiwei Li, Wei-Ying Ma, and Nenghai Yu. Bipartite graph reinforcement model for web image annotation. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 585–594. ACM, 2007.

[RM01] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *International Conference on Machine Learning*, pages 441–448, 2001.

[Rob77] Robert W Robinson. Counting unlabeled acyclic digraphs. *Combinatorial mathematics V*, 622(1977) :28–43, 1977.

[RoKBL12] Marçal Rusinõl, Dimosthenis Karatzas, Andrew D Bagdanov, and Josep Lladós. Multipage document retrieval by textual and visual representations. In *International Conference on Pattern Recognition*, pages 521–524, 2012.

[SB91] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1) :11–32, 1991.

[SCG02] Gerard Sychay, Edward Chang, and Kingshy Goh. Effective image annotation via active learning. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 209–212. IEEE, 2002.

[SCR08] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[SDW01] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[SFCL04] Rui Shi, Huamin Feng, Tat-Seng Chua, and Chin-Hui Lee. An adaptive image content representation and segmentation approach to automatic image annotation. *Image and Video Retrieval*, pages 1951–1951, 2004.

[SG12] Saurabh Sharma and Vishal Gupta. Recent developments in text clustering techniques. *International Journal of Computer Applications*, 37(6) :14–19, January 2012.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[Sha48] Claude E Shannon. A mathematical theory of communication, part i, part ii. *Bell Syst. Tech. J.*, 27 :623–656, 1948.

[SJN06] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, 2006.

[SM12] Neepa Shah and Sunita Mahajan. Document clustering : A detailed review. *International Journal of Applied Information Systems*, 4(5) :30–38, October 2012.

[SOS92] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[SPS13] Pankaj Savita, Deepshikha Patel, and Amit Sinhal. A novel technique to image annottion using neural network. *International Journal of computer Applications & Information Technology*, 2, 2013.

[SPS+17] Luca Surace, Massimiliano Patacchiola, Elena Battini Sönmez, William Spataro, and Angelo Cangelosi. Emotion recognition in the wild using deep neural networks and bayesian classifiers. *arXiv preprint arXiv :1709.03820*, 2017.

[SRZ+08] Josef Sivic, Bryan C Russell, Andrew Zisserman, William T Freeman, and Alexei A Efros. Unsupervised discovery of visual object class hierarchies. In

Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[STL+14] Fuming Sun, Jinhui Tang, Haojie Li, Guo-Jun Qi, and Thomas S Huang. Multi-label image categorization with sparse factor representation. *IEEE Transactions on Image Processing*, 23(3) :1028–1037, 2014.

[Sut82] John Sutherland. Image collections : librarians, users and their needs. *Art Libraries Journal*, 7(2) :41–49, 1982.

[SWS+00] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12) :1349–1380, 2000.

[TF17] Amara Tariq and Hassan Foroosh. Image annotation using multi-layer sparse coding. *arXiv preprint arXiv :1705.02460*, 2017.

[THA12] Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. Semantic hierarchies for image annotation : A survey. *Pattern Recognition*, 45(1) :333–345, 2012.

[THY+11] Jinhui Tang, Richang Hong, Shuicheng Yan, Tat-Seng Chua, Guo-Jun Qi, and Ramesh Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2) :14, 2011.

[TK01] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Advances in neural information processing systems*, pages 647–653, 2001.

[TW16] Youbao Tang and Xiangqian Wu. Text-independent writer identification via cnn features and joint bayesian. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 566–571. IEEE, 2016.

[TYH+09] Jinhui Tang, Shuicheng Yan, Richang Hong, Guo-Jun Qi, and Tat-Seng Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 223–232. ACM, 2009.

[UBBDB13] Tiberio Uricchio, Lamberto Ballan, Marco Bertini, and Alberto Del Bimbo. An evaluation of nearest-neighbor methods for tag refinement. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[VAD04] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[VJ12] Yashaswi Verma and CV Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *European Conference on Computer Vision*, pages 836–849. Springer, 2012.

[VR11] Carl Vondrick and Deva Ramanan. Video annotation and tracking with active learning. In *Advances in Neural Information Processing Systems*, pages 28–36, 2011.

[Wan11] Feichao Wang. A survey on automatic image annotation and trends of the new age. *Procedia Engineering*, 23 :434–438, 2011.

[WDS+01] Liu Wenyin, Susan T Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, and Brent A Field. Semi-automatic image annotation. In *Interact*, volume 1, pages 326–333, 2001.

[WH11] Meng Wang and Xian-Sheng Hua. Active learning in multimedia annotation and retrieval : A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2) :10, 2011.

[WHS+06] Meng Wang, Xian-Sheng Hua, Yan Song, Xun Yuan, Shipeng Li, and Hong-Jiang Zhang. Automatic video annotation by semi-supervised learning with kernel density estimation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 967–976. ACM, 2006.

[WHY+12] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance : a relationship measure for visual concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5) :863–875, 2012.

[WJZZ07] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Content-based image annotation refinement. In *CVPR*, pages 1–8. IEEE, 2007.

[WLW12] Mei Wang, Feng Li, and Meng Wang. Collaborative visual modeling for automatic image annotation via sparse model coding. *Neurocomputing*, 95 :22–28, 2012.

[WML04] Xin-Jing Wang, Wei-Ying Ma, and Xing Li. Data-driven approach for bridging the cognitive gap in image retrieval. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 3, pages 2231–2234. IEEE, 2004.

[WP94] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[WYM+16] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn : A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016.

[WYZZ09] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. Multi-label sparse coding for automatic image annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1643–1650. IEEE, 2009.

[WZJM06] Xin-Jing Wang, Lei Zhang, Feng Jing, and Wei-Ying Ma. Annosearch : Image auto-annotation by search. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1483–1490. IEEE, 2006.

[WZL+10] Xin-Jing Wang, Lei Zhang, Ming Liu, Yi Li, and Wei-Ying Ma. Arista-image search to annotation on billions of web photos. In *CVPR*, pages 2987–2994. IEEE, 2010.

[YH08] Oksana Yakhnenko and Vasant Honavar. Annotating images and image objects using a hierarchical dirichlet process model. In *Proceedings of the 9th International Workshop on Multimedia Data Mining : held in conjunction with the ACM SIGKDD 2008*, pages 1–7. ACM, 2008.

[YHC12] Ning Yu, Kien A Hua, and Hao Cheng. A multi-directional search technique for image annotation propagation. *Journal of Visual Communication and Image Representation*, 23(1) :237–244, 2012.

[YMS14] Mussarat Yasmin, Sajjad Mohsin, and Muhammad Sharif. Intelligent image retrieval techniques : a survey. *Journal of applied research and technology*, 12(1) :87–103, 2014.

[YWSZ13] Ying Yuan, Fei Wu, Jian Shao, and Yueting Zhuang. Image annotation by semi-supervised cross-domain learning with group sparsity. *Journal of Visual Communication and Image Representation*, 24(2) :95–102, 2013.

[ZC03] Cha Zhang and Tsuhan Chen. Annotating retrieval database with active learning. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–595. IEEE, 2003.

[ZHH+10] Shaoting Zhang, Junzhou Huang, Yuchi Huang, Yang Yu, Hongsheng Li, and Dimitris N Metaxas. Automatic image annotation using group sparsity. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3312–3319. IEEE, 2010.

[ZIL12] Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1) :346–362, 2012.

[ZLT17] Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval : A literature survey. *arXiv preprint arXiv :1706.06064*, 2017.

[ZLX10] Shile Zhang, Bin Li, and Xiangyang Xue. Semi-automatic dynamic auxiliary-tag-aided image annotation. *Pattern Recognition*, 43(2) :470–477, 2010.

[ZTH+14] Shiliang Zhang, Qi Tian, Gang Hua, Qingming Huang, and Wen Gao. Object-patchnet : Towards scalable and semantic image annotation and retrieval. *Computer Vision and Image Understanding*, 118 :16–29, 2014.

[ZWH08] Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1129–1136. Association for Computational Linguistics, 2008.

[ZZZP08] Yufeng Zhao, Yao Zhao, Zhenfeng Zhu, and Jeng-Shyang Pan. A novel image annotation scheme based on neural network. In *Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on*, volume 3, pages 644–647. IEEE, 2008.