INFÉRENCE DE CONNAISSANCES SÉMANTIQUES, APPLICATION AUX IMAGES SATELLITAIRES

Jean-Baptiste Bordes

Thèse

Présentée pour obtenir le grade de docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Signal et Images**

Jean-Baptiste Bordes

Inférence de connaissances sémantiques : application aux images satellitaires

Résumé

Une méthode probabiliste pour annoter des images satellites avec des concepts sémantiques est présentée. Cette méthode part de caractéristiques de bas-niveau quantifiées dans l'image et utilise une phase d'apprentissage à partir des concepts fournis par un utilisateur avec un lot d'images exemples. La contribution principale est la définition d'un formalisme pour la mise en relation d'un réseau sémantique hiérarchique avec un modèle stochastique. Les liens sémantiques de synonymie, méronymie, hyponymie sont mis en correspondance avec différents types de modélisations inspirées des méthodes utilisées en fouille de données textuelles. Les niveaux de structuration et de généralité des différents concepts utilisés sont pris en compte pour l'annotation et la modélisation de la base de données. Une méthode de sélection de modèle permet de déduire le réseau s émantique correspondant à la modélisation optimale de la base de données. Cette approche exploite ainsi la puissance de description des réseaux sémantique tout en conservant la flexibilité des approches statistiques par apprentissage. La méthode a été évaluée sur des bases de données SPOT5 et Quickbird.

Abstract

A novel method is presented for annotating satellite images. The labels used for annotation are given by a user with a set of example images. A learning step is then applied to learn the model. The originality of the method is to formulate the problem of semantic annotation to a further extent than a mere probabilistic classification task. The method takes into account the semantical relationships between the concepts by considering a *duality* between the structure of the model and the structure of the set of labels. The semantical structure of the labels is represented by a semantic network containing three semantical relationships : synonymy, meronymy, and hyponymy. The semantic network is constrained in a hierarchy induced by the links of hyponymy and meronymy. By a procedure of MDL model selection, it is possible to find the optimal semantical structure of the set of labels.

Table des matières

Chapitre 1

Introduction

1.1 Enjeux du problème d'indexation sémantique

Au cours de la dernière décennie, les quantités d'images détenues par les bases d'images satellitaires ont augmenté considérablement. Ces quantités deviennent encore plus énormes avec l'arrivée de nouveaux capteurs à haute résolution qui fournissent en permanence de nouvelles images de la Terre. Utiliser des opérateurs humains pour annoter toutes ces images étant d'un coût exorbitant, il devient important de développer des systèmes automatiques permettant d'accéder de façon fiable et simple à ces grandes bases de données afin qu'elles deviennent véritablement exploitables. Or, un utilisateur humain effectuant des requêtes à un niveau sémantique, il est crucial de parvenir à une description sémantique automatique de l'image avec le vocabulaire du langage naturel. Pourtant, les systèmes actuels d'indexation peinent à fournir une interprétation sémantique d'une image, car ils se basent sur des descripteurs extraits directement sur l'image comme la couleur, la texture, la forme ou toute autre description que l'on appellera ici de "bas-niveau" car ces caractéristiques sont extraites directement de la représentation numérique de l'image et n'ont pas de lien immédiat avec la sémantique présente dans l'image. Beaucoup de travaux sur la recherche d'image par le contenu ont utilisé directement ces caractéristiques de bas-niveau qui ont donné des résultats satisfaisants pour des requêtes du type "Requête par présentation d'images exemples" où l'utilisateur fournit au système une ou plusieurs images et lui demande de lui renvoyer un lot d'images similaires. Cependant, ces caractéristiques symboliques ne peuvent pas satisfaire pleinement les attentes des utilisateurs. La raison en est qu'un utilisateur pense sa requête en termes sémantiques (zone pavillonaire, zone portuaire etc.), et non en termes de valeur symbolique extraite (zone verte, texture rayée). De plus, il est difficile de trouver des descripteurs puissants pour l'image permettant de décrire des notions sémantiques. On appelle ce problème le "fossé sémantique" [15], il est défini comme : "le manque de concordance entre les informations qu'on peut extraire des données visuelles et l'interprétation qu'ont ces mêmes données pour un utilisateur dans une situation donnée" [124]. Ce "fossé" est une difficulté récurrente tant en indexation d'images fixes que dans le domaine de l'indexation audio ou vidéo et il correspond au problème de liaison entre une description de bas-niveau et une description de

haut-niveau d'une image.

Depuis quelques années, certains résultats intéressants ont été obtenus pour l'annotation sémantique automatique d'images, sur des images assez diverses allant des photos personnelles et des dessins aux images satellitaires et aériennes. Ces travaux émanent d'une prise de conscience de l'importance de ce sujet pour parvenir à franchir un cap dans le domaine de l'indexation. Dans une base d'images, un certain nombre d'images sont annotées manuellement, et le système doit, à partir de cet apprentissage, "propager" les annotations au restant de la base. Les annotations textuelles ainsi effectuées permettent ensuite de répondre plus facilement aux requêtes, elles aussi textuelles, de l'utilisateur. Dans le cas particulier des images satellitaires, les systèmes actuels d'annotations sémantiques sont souvent construits à partir de règles prédéfinies provenant des connaissances d'un photo-interprète [141]. Ces systèmes, bien qu'ils donnent des résultats satisfaisants, sont peu souples pour adapter ou rajouter des notions sémantiques nouvelles au système. Nous nous intéressons dans cette thèse au contraire à l'utilisation de méthodes statistiques permettant un apprentissage automatique à partir d'exemples, rendant cette méthode aisément adaptable et généralisable.

L'approche développée dans ce travail se place après une première étape de traitement d'image extrayant des caractéristiques de "bas-niveau" : coefficients de texture, extraction d'objets, extraction de réseaux routiers... Il s'agit ainsi de faire le lien entre ces caractéristiques de bas-niveau qui apportent en elles-mêmes peu d'information, et des notions sémantiques de haut-niveau décrites par le vocabulaire du langage naturel. Nous souhaitons donc développer des méthodes qui permettent d'apprendre des concepts sémantiques à partir d'images annotées par l'utilisateur, pour pouvoir ensuite propager ces annotations à des images non annotées.

1.2 Caractéristiques de l'approche proposée

D'une façon générale, on remarque que de plus en plus d'efforts de recherche se concentrent sur des interfaces entre plusieurs discplines : ainsi, de plus en plus de travaux de recherche en biologie font appel à l'informatique et à la physique. Le problème traité ici apparait également comme extrêmement pluri-disciplinaire. Comme on peut le voir, il fait en effet appel à des domaines aussi divers que ceux de la vision par ordinateur, de la fouille de données, de la sémantique, du traitement d'images (domaine en lui-même très pluri-disciplinaire), de l'intelligence artificielle, de l'apprentissage et même de la théorie de l'information. Parmi tous ces domaines, le champ d'investigation de la sémantique, ou plutôt des sémantiques comme nous verrons dans le deuxième chapitre, est sans doute le plus difficile à définir.

L'originalité principale du travail effectué ici consiste justement à poser le problème de l'extraction de sémantique d'une façon que nous souhaitons à la fois plus complète et plus sophistiquée que ce qui est développé dans la plupart des travaux de recherche en annotation d'images, à savoir comme une simple tâche de classification. En effet, considérons une base d'images d'apprentissage annotée par des concepts. La première phase, essentielle, est celle du processus de traitement d'im-

ages qui consiste à extraire des caractéristiques pertinentes de ces images. Ensuite, le processus d'apprentissage traditionnel consiste à définir pour chaque concept un sous-espace de l'espace des caractéristiques. Les concepts sont donc traités comme des simples classes. Or, une des bases fondamentales de la sémantique est que les concepts vivent dans un espace qui leur est propre et qui est structuré par des liens sémantiques. La sémantique retient 4 relations sémantiques principales : synonymie, antonymie, hyponymie et méronymie. Pour illustrer l'utilité d'une prise en compte de tels liens sémantiques, considérons un utilisateur souhaitant rechercher dans une base de données des images correspondant au concept de "végétation". Si, dans les modèles sémantiques du système, le concept de végétation est relié par une relation d'hyponymie aux concepts "prairie", "forêt" et "savane" et que ces 3 concepts sont associés à des modèles stochastiques, il n'est pas nécessaire d'estimer des nouveaux paramètres pour le concept végétation. Toutes les images annotées par les concepts "prairie", "forêt" et "savane" peuvent automatiquement être annotées par le concept "végétation". Ainsi, l'approche considérée, illustrée figure 1.1, consiste à mettre en relation la structure du lexique d'annotation, représentée par un réseau sémantique, avec un modèle statistique. Étant donné une base de données annotée, un algorithme de sélection de modèle peut déterminer la structure du modèle sémantique qui permet le mieux de décrire le signal de la base de données et donc déterminer des liens sémantiques entre les concepts.

La méthode proposée dans ce travail, appelée méthode d'"Annotation Sémantique Paradigmatique" (ASP), est conçue pour pouvoir s'appliquer à tout type d'image. Cependant, le choix de travailler avec des images satellitaires s'explique par des raisons de simplicité. En effet, contrairement à des images telles que des photographies personnelles, les images satellitaires comportent l'avantage de la connaissance précise du type d'images avec lequel on travaille : résolution, luminosité, angle d'observation, etc. Ainsi, une des difficultés de la tâche d'extraction de sémantiques est supprimée car on suppose que toutes les images ont le même type et correspondent au même contexte. On suppose de plus que l'utilisateur est intéressé par une application de type cartographique et que le vocabulaire avec lequel il souhaite travailler est un vocabulaire de type photo-interprète. Les mots employés servent à nommer des zones de taille variée pouvant correspondre à des régions de quelques milliers de pixels pouvant être annotée par des concepts tels que "hangar" ou "parc", mais peuvent aussi correspondre à des zones de plusieurs millions de pixels et annotées par des concepts très abstraits tels que "banlieue résidentielle" ou "complexe industriel". Pour évaluer les algorithmes proposés, nous avons ainsi travaillé avec deux types d'images différents : des images du satellite Spot5 à 2,5 mètres de résolution centrées sur des villes françaises, et des images Quickbird de Pékin à 0,6 mètres de résolution.

1.3 Structure du rapport

Nous commencerons dans le premier chapitre par définir la sémantique, la délimitation de son domaine et comment il est possible de la faire émerger par un pro-

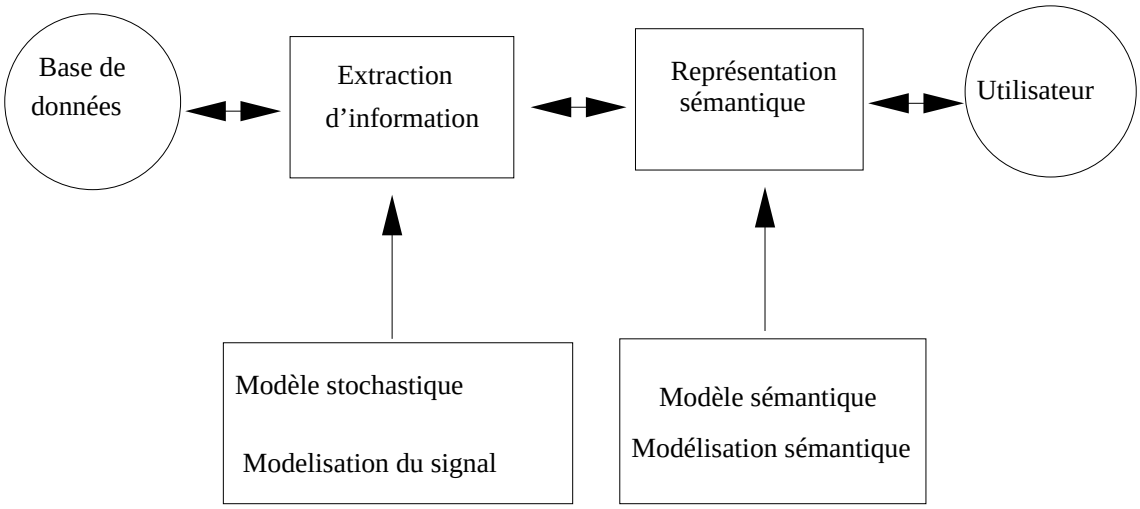


Fig. 1.1 – Dualité modèle stochastique/modèle sémantique, proposée par M. Datcu

cessus de fouille de données. Le deuxième chapitre exposera ensuite un état de l'art des techniques d'inférence traditionnellement utilisées en extraction de sémantique dans le texte et dans les images ainsi que les principaux modèles utilisés en indexation de données textuelles. Ensuite, les bases de la méthode hiérarchique proposée, qui constitue le cœur du travail, seront exposées dans le troisième chapitre : la problématique de mise en correspondance d'un réseau sémantique et d'un modèle stochastique sera développée ainsi qu'une évaluation des performances de construction d'un réseau sémantique. Le cinquième chapitre traitera ensuite de l'approche utilisée pour annoter des images test et des évaluations qui ont été effectuées.

En annexe A, un travail de classification de d'images satellitaires à haute résolution est présenté. Ce travail a été effectué au cours de la thèse et est utilisé en entrée de la modélisation qui est présentée pour les expériences effectuées sur la base d'images Quickbird. En annexe B, le cadre de la modélisation markovienne qui est utilisée en comparaison avec la méthode ASP est détaillé. Enfin, l'annexe C présente un autre modèle statistique d'annotation automatique qui a été formalisé et expérimenté au cours de la thèse. Ce modèle a été abandonné au profit du modèle ASP, ce dernier étant plus satisfaisant et efficace à tous points de vue. Cependant, ce modèle a été laissé dans l'annexe à titre indicatif, tel une étape intermédiaire vers la formalisation du modèle ASP.

Chapitre 2

Sémantique et fouille de données

Si le sens est une donnée immédiate et fondamentale de notre expérience des langues, il ne va pas de soi de passer du sens communément perçu d'un mot à un sens érigé en objet d'étude linguistique. C'est ce que confirme l'avènement tardif de la sémantique en tant que branche des sciences du langage ayant pour domaine d'étude les significations propres aux langues. On attribue en effet généralement la paternité du mot *sémantique* à Michel Bréal. Plutôt que 1883, année où il forge le terme de *sémantique* dans l'article "Les lois intellectuelles du langage : fragments de sémantiques", on retient généralement 1887 comme année de naissance officielle de la *science des significations* qu'il inaugure dans son *Essai de sémantique* paru en 1897 [17] : "L'étude où nous invitons le lecteur à nous suivre est d'espèce si nouvelle qu'elle n'a même pas encore reçu de nom. En effet, c'est sur le corps et sur la forme des mots que la plupart des linguistes ont exercé leur sagacité : les lois qui président à la transmission des sens, au choix d'expressions nouvelles, à la naissance et à la mort des locutions ont été laissées dans l'ombre ou n'ont été indiquées qu'en passant. Comme cette étude, aussi bien que la phonétique et la morphologie, mérite d'avoir son nom, nous l'appellerons la sémantique (du verbe *semainen*), c'est-à-dire la science des significations".

L'idée directrice de cette nouvelle science est que les mots, formes et sens mènent une existence qui leur est propre, et qu'il appartient à la linguistique comparée d'établir des lois de l'évolution des significations des mots tout comme elle établit les lois de leur évolution phonétique. Cependant, cette datation de la naissance de la sémantique n'est pas totalement satisfaisante car ce que les linguistes appellent aujourd'hui la *sémantique* n'a plus grand chose à voir avec la science des significations qu'avait fondée Bréal. La délimitation du domaine d'étude de la sémantique, tout comme les méthodologies qu'elle a employé ont évolué au cours du temps. C'est pourquoi il convient de définir plus précisément la sémantique et son domaine d'application avant de s'attarder sur les probématiques et les techniques d'extraction de sémantiques dans les bases de données couramment employées dans le texte et dans les images.

2.1 Qu'est ce que la sémantique ?

Les manuels contemporains définissent avec unanimité la sémantique pour son objet d'étude : le sens à travers les formes et les structures signifiantes des langues. Cependant, ils délimitent différemment le domaine d'investigation de la sémantique. On peut par exemple considérer la *sémantique lexicale*, qui limite le sens linguistique à celui des seules unités lexicales, mots simples ou expressions codées. Ou encore la *pragma-sémantique*, qui s'attaque à trois niveaux distincts d'organisation du sens : structuration lexicale au niveau des unités-mots, structuration grammaticale ou morpho-syntaxique au niveau des unités-phrases, organisation discursive au niveau des unités énoncés. Ainsi, J. Lyons [84] affirme que "la définition de la sémantique comme l'étude du sens reflète le seul point sur lequel les sémanticiens se soient mis d'accord". De même, la sémantique ne se caractérise pas non plus par une méthodologie qui lui est propre. Les méthodes descriptives varient selon les théories linguistiques et la sémantique a emprunté des outils d'analyse ou des principes explicatifs à divers domaines de la linguistique (phonétique, phonologie, syntaxe), aussi bien qu'à d'autres sciences et techniques : logique, mathématique, informatique, ou encore intelligence artificielle.

En résumé, la sémantique ne peut absolument pas être vue comme une théorie homogène, d'autant plus que l'histoire de la sémantique dépend fortement des grands courants théoriques qui ont jalonné la linguistique, dont elle est, rappelons le, un secteur particulier. On citera ici la sémantique évolutionniste, la sémantique structurale, la sémantique des grammaires formelles, et la sémantique cognitive. Leur influence, qui n'est du reste pas exclusive, permet de distinguer quatre grandes périodes :

- La période évolutionniste (à partir de 1897) : Une sémantique historique y domine et, inspirée par la théorie de l'évolution de Darwin, cherche des lois générales à l'évolution du sens des mots au cours de l'Histoire.
- La période structurale (à partir de 1931) : Il s'agit de dégager l'organisation intrinsèque du lexique en faisant appel à un ensemble fini d'éléments sémiques (*plus petites unités de sens*) et à un compartimentage en champs lexicologiques.
- La période des grammaires formelles (à partir de 1963) : La sémantique est transférée du lexique aux phrases, avec pour principal centre d'intérêt les rapports entre les structures syntaxiques et sémantiques des phrases, incluant certains systèmes indexicaux.
- La période cognitive (à partir de 1978) : La sémantique cognitive vise à naturaliser le sens linguistique en le rattachant au fonctionnement général du cerveau.

Détaillons maintenant le propos de chacune de ces sémantiques :

2.1.1 La période évolutionniste

Dans son *Essai de sémantique*, Michel Bréal conçoit ainsi sa nouvelle science d'étude des significations linguistiques dans une perspective diachronique. Il cherche en effet une réponse à la question de savoir pourquoi et comment les mots changent

de sens : "Laissant de côté les changements de phonétique, qui sont du ressort de la grammaire physiologique, j'étudie les causes intellectuelles qui ont présidé à la transformation de nos langues. Pour mettre de l'ordre dans cette recherche, j'ai rangé les faits sous un certain nombre de *lois*". La sémantique évolutionniste emprunte ainsi à la fois l'objectif du darwinisme ambiant et sa méthode scientifique d'observation empirique de phénomènes que sont les "faits de sens".

Les sémanticiens de cette époquent analyseront des faits d'évolution du langage dans un grand nombres de langues et dialectes depuis l'antiquité jusqu'à nos jours. Meillet mettra alors en évidence la nécéssité de faire intervenir la biologie ou la sociologie pour mettre en lumière des principes explicatifs des changements du sens dans les langues : "Par le fait même qu'ils dépendent immédiatement des causes extérieures à la langue, les changements sémantiques ne se laissent pas restituer par des hypothèses proprement linguistiques" [87].

Les premiers tenants de la linguistique structurale, préférant étudier des données en apparence plus formelles (phonologiques, morphologiques ou syntaxiques), ont alors manifesté à l'égard de la sémantique évolutionniste une méfiance taxée d'idéalisme qui l'a conduit à rester dans l'ombre jusqu'aux années soixante. Irène Tamba analyse ainsi le paradoxe de l'apparition de la sémantique [130] : "D'un côté, cette optique a libéré la sémantique de la tutelle de la philosophie, de la logique, et de la psychologie, mais d'un autre côté, la recherche des lois générales de l'évolution sémantique conduit à demander à la biologie ou à la sociologie des principes explicatifs des changements de sens dans les langues. En prenant pour objet d'étude l'évolution du sens des mots, la sémantique a donc choisi un domaine qu'elle ne contrôle pas complètement".

2.1.2 La sémantique structurale

Le structuralisme est un mouvement linguistique qui trouve son origine dans le "Cours de linguistique générale" de Saussure en 1916. Le structuralisme, ensuite devenu une branche des sciences humaines, a vu peu à peu son sens devenir quelque peu galvaudé. Cependant, en linguistique, le structuralisme a gardé une définition relativement rigoureuse. Georges Mounin, dans [93], précise ainsi que "étudier les structures linguistiques c'est, rigoureusement, étudier la construction de certains ensembles lignuistiques ; c'est-à-dire essayer de déceler, d'après les fonctions linguistiques, les unités réelles qui construisent ces ensembles, et les règles d'emploi pour construire ces ensembles".

Le structuralisme a ainsi engendré l'émergence d'une sémantique synchronique qui tente de structurer la sémantique d'une langue : ses *signifiés*, et son lexique : ses *signifiants*. Cette volonté de structuration n'est pas totalement désintéressée puisqu'elle vise également à constituer des dictionnaires avec des organisations plus rationnelles qu'à l'aide d'un tri alphabétique.

Dans [93], Georges Mounin distingue deux types de méthode de structurations possibles de la langue :

- La structuration formelle, qui se base sur des marques présentes dans les signifiants eux même.

– La structuration conceptuelle qui se base au contraire sur les signifiés.

Quelque soit la méthode utilisée, la sémantique structurale vise à introduire un ordre dans une partie du lexique, voire dans sa totalité.

2.1.3 La sémantique des grammaires formelles

La sémantique des grammaires formelles étudie principalement les rapports entre les structures syntaxiques et sémantiques des phrases. Les grammaires formelles reposent sur la définition d'un certain nombre de catégories syntaxiques et des relations existant entre elles qui modélisent les structures de phrase du langage. La sémantique des grammaires formelles analyse une langue à l'aide d'un ensemble d'états finis, permettant de caractériser la phrase sur la base de ses constituants, de leurs interprétations, et la façon dont ils sont regroupés.

Une grammaire formelle est constituée d'un 4-upplet : (N, T, R, a) où :

– N est un ensemble fini non vide de symboles dit *alphabet non terminal*
– T est un ensemble fini non vide de symboles dit *alphabet terminal*, dont les éléments sont appelés *symboles terminaux*. Les ensembles N et T sont disjoints et leur union définit l'alphabet global V.
– R est l'ensemble fini et non vide des règles grammaticales, ou productions : chaque production est de la forme $\alpha \rightarrow \beta$ où $\alpha \in V$ et $\beta \in V$. α, appelé "tête", contient au moins un symbole non terminal.
– a est appelé l'axiome, ou symbole de départ, et est un élément particulier de N.

Les symboles non terminaux correspondent aux catégories syntaxiques, et les symboles non terminaux correspondent aux mots constitutifs de la phrase lorsque le processus de génération se termine. Le processus de génération consiste à appliquer à chaque pas un règle de production jusqu'à ce qu'aucune règle ne puisse être appliquée ou que l'on ait éliminé tous les symboles non terminaux. En introduisant un certain nombre de limitations sur la forme des règles de production, Chomsky a introduit en 1956 une classification hiérarchique des grammaires et des langages qui est très largement acceptée.

Cependant, d'autres grammaires basées sur des contraintes, comme la Grammaire Lexicale Fonctionnelle ([18]), la Grammaire d'Arbres Adjoints ([67]) analysent la phrase en termes de constituants syntaxiques. Les Grammaires Sémantiques ([21]) telles que la Grammaire de Cas ([41]) analysent la structure de la phrase à un niveau sémantique plutôt que syntaxique.

Si la grammaire est implémentée sous forme de règles qui sont établies manuellement, les structures qui en résultent se révèlent assez complexes dès lors que l'on envisage des questions de maintenance et de compatibilité. En effet, le jeu de règles est en général adapté afin d'obtenir des performances optimales pour une tâche spécifique. Il a été envisagé d'étendre les méthodes statistiques telles que celles fondées sur les Modèles de Markov à l'analyse sémantique. Ces méthodes donnent de très bons résultats aux niveaux acoustique et linguistique. L'information sémantique étant encodée dans un corpus au lieu d'utiliser des règles explicites, ces grammaires sont plus flexibles et réutilisables. Après une analyse automatique des

données, l'information sémantique est mémorisée sous forme des paramètres d'un modèle stochastique. Un exemple de ce principe d'analyse est observable dans les Modèles de Compréhension Cachés ([114]).

2.1.4 La sémantique cognitive

Selon [130], la sémantique cognitive est baptisée en 1978 par la publication d'un article de L. Talmy intitulé "The relation of Grammar to Cognition : A synopsis". La sémantique cognitive a profité de l'émergence des neurosciences cognitives qui, couplées au développement de l'imagerie cérébrale, ont mis en lumière certaines relations entre activité cérébrale et fonctions cognitives. Dans [105], François Rastier décrit la sémantique cognitive de la manière suivante : "Contestant la sémantique formelle au profit d'un mentalisme généralisé, la sémantique cognitive pose que le sens linguistique consiste en représentations ou processus mentaux, ce qui la conduit à s'appuyer sur une psychologie ou une phénoménologie spontanée. Dans tous les cas, la mise en relation du linguistique et du mental soulève des difficultés considérables.". Ainsi, la sémantique cognitive s'est construite par le rejet de la sémantique des grammaires formelles et préfère les modèles connexionnistes des réseaux de neurones interconnectés plutôt que les modèles logico-mathématiques opérant sur des symboles. Dans [130], Irène Tamba distingue trois grands chantiers d'études de la sémantique cognitive :

- Les *expressions spatiales* dont le projet général est expliqué par François Rastier dans [104] comme étant d'"analyser la langue de façon à faire apparaître en elle les représentations et processus cognitifs nécessaires à la production et à la compréhension de ces expressions spatiales."
- La *catégorisation*, introduite par Rosch [111], part d'expériences psychologiques afin d'amener à la notion de *prototype*. Le but est de proposer une méthode universaliste pour définir des classes lexicales.
- L'*activité de conceptualisation métaphorique*, qui physicalise la pensée abstraite [78]

2.2 Relations paradigmatiques et syntagmatiques

La sémantique structurale distingue deux axes de structure : la structuration paradigmatique, qui concerne les relations sémantiques entre les éléments du lexique, et la structuration syntagmatique qui concerne leurs possibilités de combinaison. Dans [84], Lyons appelle *information paradigmatique* l'information provenant de la possibilité de choisir une unité lexicale plutôt qu'une autre, et *information paradigmatique* l'information provenant de la possibilité de combiner entre elles des unités lexicales.

2.2.1 Les relations paradigmatiques

La sémantique lexicale retient 4 types principaux de relations paradigmatiques : la synonymie, l'antonymie, l'hyponymie/hyperonymie, et la méronymie.

Nous détaillons à présent le sens précis de ces différentes relations entre concepts.

La synonymie On dit que deux unités lexicales sont synonymes dans un contexte donné si il est possible de les échanger sans modifier le sens communicatif de l'énoncé. Deux types de synonymies peuvent être considérés : la synonymie dite *complète*, qui correspond à deux unités interchangeables quelque soit le contexte, et la synonymie dite *partielle* pour laquelle les concepts ne sont interchangeables que dans un nombre limité de contexte. Notons que la synonymie ne s'applique que pour des unités lexicales de même nature.

L'antonymie Le terme d'antonymie, introduit au milieu du XIXème siècle, est utilisé aujourd'hui en sémantique pour désigner une relation de *contraire*, opposé à celle de synonymie. Comme la synonymie, elle relie des unités lexicales appartenant à une même catégorie grammaticale. Cependant, vue sous un angle sémantique, l'antonymie diffère de la synonymie par une forte binarité et par une plus grande difficulté à être caractérisée de manière précise. Dans [130], Irène Tamba précise que, plutôt que de parler d'un seul type d'antonymie, il vaut mieux distinguer quatre types d'opposition dichotomique :

- l'antonymie contradictoire : intérieur/extérieur
- l'antonymie polaire : court/long
- l'antonymie inverse : monter/ descendre
- l'antonymie réciproque : acheter/vendre

Hyperonymie/hyponymie :

Les termes d'hyponymie et d'hyperonymie n'apparaissent en sémantique qu'à la fin des années 1960. C'est J.Lyons qui forge le néologisme d'hyponymie [84] en le définissant comme une implication unilatérale : "je nourris un chat" implique "je nourris un animal", mais réciproquement, 'je nourris un animal" n'implique pas "je nourris un chat". Aussi dira-t-on que *chat* est un hyponyme de l'hyperonyme *animal*. Suivant cette définition, on a affaire à une relation paradigmatique caractéristique de la structuration verticale du lexique. La classe d'objets à laquelle s'applique, par définition le nom hyperonymique d'*animal*, peut être délimité par des phrases génériques telles que : *le chat est un animal* ou *les chats sont des animaux*. Ainsi, la classe des chats est reclassée dans celle plus générale des animaux. Cette classification par *superordination* permet un emboîtement successif de classes de plus en plus générales.

Au niveau lexical, les relations d'hyponymie/hyperonymie servent à construire des structures hiérarchiques à l'aide d'une échelle de généralité descendante ou ascendante, intrinsèque à une catégorie. Disposant d'une expression catégorielle unifiée avec les phrases définitoires génériques du type *un X est une sorte de Y*, cette dimension verticale du lexique repose donc sur la formulation verbale qui la stipule ou l'entérine dans les dictionnaires. Elle permet aussi bien d'exprimer les classifications hiérarchiques du sens commun que celles scientifiques ou techniques de spécialistes.

La méronymie Le terme de méronymie a été introduit par A. Cruse [28] pour différencier clairement ce type de structuration lexicale hiérarchique de celle induite par la relation d'hyponymie. En effet, si ces deux relations partagent les caractéristiques d'inclusion et d'asymétrie, leur sens est très différent et les deux hiérarchies qui en découlent naturellement sont incompatibles. Si l'hyponymie est basée sur la relation "sorte de", la méronymie est basée sur une relation "partie de". Ces deux sortes de structuration sont totalement incompatibles, comme l'atteste l'impossibilité d'inclure une partie dans un tout à l'aide de la relation "est une sorte de", et d'inclure une classe dans une autre à l'aide de la relation "est composé de". Dans [130], Irène Tamba précise que, si il est clair que les structurations induites par ces deux relations sont de nature totalement différente, "il est de plus délicat d'établir l'existence d'une structuration lexicale de méronymie. En effet, les travaux sur les relations partie-tout sont unanimes à signaler leur émiettement, leurs variations de langue à langue et l'absence d'un terme relationnel générique. On peut donc se demander si la méronymie constitue une catégorie relationnelle d'ordre lexical ou si l'on a affaire à des rapports entre parties et totalité qui restent dépendantes d'une linguistique référentielle. Auquel cas les relations méronymiques reposeraient davantage sur des données perceptuelles et pragmatiques que sur une catégorie relationnelle lexicale".

2.2.2 Les relations syntagmatiques

La relation syntagmatique correspond à l'enchainement d'unités lexicales selon un certain ordre. Une séquence de mots qui peuvent se succéder sans enfreindre les règles de syntaxe constituent alors un *syntagme*. Le mot, le syntagme, et la phrase constituent différents paliers de détermination du sens. Le mot isolé, hors-contexte, correspond à une première signification assez floue et peut être polysémique. Le syntagme permet déja une compréhension plus claire. La phrase correspond en principe à une unité de sens parfaitement définie.

Notons que les distinctions de sens qui sont propres à la structure, ou au système d'une langue ne se retrouvent pas nécessairement dans la structure d'une autre langue. La traduction d'un texte montre bien la difficulté, parfois même l'impossibilité lexicale, et pas seulement syntaxique, de trouver des lexèmes qui se correspondent termes à termes. En effet, il arrive fréquemment qu'une langue concentre dans une seule unité lexicale (et rend donc paradigmatique), une information qui, dans une autre langue, exige un groupe de mots (c'est-à-dire une réalisation syntagmatique). Ainsi, à titre d'exemple, la notion de "grand frère" est exprimée par un seul mot en chinois (*gege*) tandis qu'elle nécessite un syntagme de deux mots en français.

2.3 Sémantique et Sémiotique

Comme nous l'avons vu dans la section précédente, la sémantique est l'étude des significations *linguistiques*, et il faut prendre garde à ne pas appeler sémantique l'étude de toutes les significations. En effet, n'importe quel objet ou n'importe quel

événement du monde réel peut se voir attribuer une signification et il convient de ne pas confondre la signification des ces faits avec la signification linguistique à proprement parler, à savoir les "faits linguistiques qui ont pour fonction de transmettre des significations linguistiques" [93].

2.3.1 Rapports entre signifiants et significations

Si l'on s'accorde à considérer la sémantique comme l'étude des significations linguistiques, il reste à définir précisément sur ce que l'on entend par *signification linguistique*. Sur ce point, il faut savoir que ce problème est encore loin d'être clarifié et, si les les linguistes se refusent généralement à éluder la question du sens, ils le font intervenir à partir de l'axiome structuraliste de Bloomfeld, de Hjelmselv et de Harris selon lequel : "une distinction est pertinente sur un plan si elle suffit à établir une distinction sur un autre plan". Cela revient à dire qu'une différence sur le plan des signifiants correspond à une différence sur le plan des signifiés. Saussure dit ainsi dans [31] :"On ne peut pas définir une forme à l'aide de la figure vocale qu'elle représente, pas davantage à l'aide du sens que contient cette figure vocale. On est obligé de poser comme fait primordial le fait général, complexe et composé de deux faits négatifs : de la *différence* générale des figures vocales jointe à la *différence* générale des sens qui s'y peuvent attacher".

Cela revient à dire, comme l'a développé Saussure, que le signe linguistique présente deux faces : d'une part sa face signifiante, constituée d'une séquence de phonèmes, et de l'autre sa face signifiée qui correspond à ce qui est compris par l'énonciataire. De plus, le concept saussurien du signe introduit un troisième terme d'une relation triangulaire : le *concept*. Le concept (que la plupart des linguistes actuels appellent à présent le *référent*) correspond à l'entité ou l'ensemble d'entités du monde réel ou sensible désignées par le signifiant auquel il est lié. Ainsi, il va de soi qu'en tant qu'objet extralinguistique, son étude est du domaine des sciences physiques, naturelles, ou humaines et il n'est donc pas du ressort de la linguistique.

2.3.2 Sémiotique et sémiologie

Le but de la sémiotique est de rendre compte de la signification d'un *objet sémiotique*. Cet objet s'exprime à travers un certain nombre de *canaux* qui correspondent aux cinq sens. Plusieurs canaux peuvent être utilisés à la fois : dans le cas d'un clip publicitaire, l'information est émise en utilisant à la fois des images et du son (musique plus parole). Précisons également que ce que l'on entend par sémiotique correspond à la tradition française de la sémiotique initiée par Saussure et analyse le rapport entre les signes et leur signification. La sémiologie (ce terme vient du mot anglais "semiotics") insiste davantage sur la classification et la typologie des signes et leurs formes de communication.

Saussure donne initialement de la sémiotique la définition suivante : "Science qui étudie la vie des signes au sein de la vie sociale". Ainsi, Saussure considérait la sémiotique comme l'étude des signes dans un environnement socioculturel donné. En effet, si le recours à des signes est universel, il n'existe pas de signes universels

et certains signes peuvent avoir des sens très différents selon les cultures. De plus, dans une culture donnée, le sens de certains signes et la relation d'un signifié à un signifiant a parfois beaucoup évolué au cours du temps ("fenestre" en ancien français est devenu "fenêtre" par la suite).

L'objet sémiotique étudié doit être clairement délimité afin d'éviter de dévier dans une subjectivité incontrôlée. Il est en effet totalement différent d'étudier simplement un extrait d'*Eugénie Grandet*, le roman en entier, ou encore tout la *Comédie humaine* de Balzac. Cet objet sémiotique correspond ainsi à un *ensemble signifiant* que L. Hjelmselv appelle "plan de l'expression" [59].

La sémiotique part du postulat de base selon lequel le sens n'est accessible qu'à travers des différences (d'expression et de contenu) et s'emploie généralement à décomposer le signifant, sensé correspondre à un continuum, en un ensemble d'unités distinctes qui correspondent au signifié. La première étape de la sémiotique est donc de délimiter un ensemble fini d'unités descriptives de l'objet sémiotique considéré, puis ensuite d'élaborer un certain nombre de relations entre elles.

2.4 Réseaux sémantiques

2.4.1 Définition

La notion de réseau sémantique est à présent relativement ancienne dans la littérature des sciences cognitives et de l'intelligence artificielle et a été développée pour beaucoup d'applications et à travers différentes méthodes ces vingt dernières années. Actuellement, le terme "réseau sémantique" tel qu'il est désigné actuellement correspond davantage à une famille de schémas de représentation plutôt qu'à un formalisme précis. Nous dressons ici un rapide historique des réseaux sémantiques :

Le terme de réseau sémantique a été introduit pour la première fois dans la thèse de Ross Quillian [103] en 1968 comme une façon de décrire l'organisation des "mots-concepts" dans la mémoire humaine. Mais l'idée d'un réseau sémantique comme un réseau d'association de concepts liés entre eux est plus tardive (Anderson et Bower [16]). Plus spécifiquement, les réseaux sémantiques sont vus comme "un format de représentation permettant de stocker le sens des mots de manière à permettre une utilisation de ces sens telle que pourrait le faire un humain" [103], et, comme dans la plupart des travaux de recherche sur les réseaux sémantiques depuis la proposition originale de Quillian, ils ont été conçus pour représenter la part non émotive, autrement dit "objective" du sens : les propriétés des choses, plutôt que notre perception de celles-ci.

La représentation mentale que constitue les réseaux sémantiques dépasse celle de la définition d'un simple dictionnaire : les réseaux sémantiques reflètent la façon complexe dont est structurée la connaissance humaine. Chaque concept trouve sa place dans un réseau de relations entre concepts. Nous pouvons ainsi représenter la connaissance d'une personne par un graphe dont les noeuds sont des concepts individuels et des arcs étiquettés reliant ces noeuds entre eux. Ainsi, quelques années après la thèse de Quillian, un psychologiste, Alain Collins, mena une série d'-expériences avec Quillian pour tester la plausibilité des réseaux sémantiques en tant

que modèles de l'organisation de la mémoire et des mécanismes d'inférences humains. Ils développèrent un réseau sémantique disposant d'une structure hiérarchique dans laquelle les concepts sont stockés avec un certain nombre de propriétés. Selon cette étude, le modèle explique, avec des principes très simples, la quantité de temps relative pour vérifier une phrase, en fonction du nombre de niveaux de la hiérarchie qui doivent être parcourus pour trouver une relation entre un concept et une propriété, ou entre deux concepts.

Un réseau de représentation du sens d'une phrase a été explicité par Robert Simmons ([118]). Simmons abandonne la représentation hiérarchique de Quillian au profit d'une structure constituée d'un noeud, représentant le verbe principal d'une phrase, attaché à des liens représentant des champs qui sont associés au verbe. Ainsi, la phrase : "Philippe joue au poker avec Yves" sera représentée par un réseau sémantique où le noeud central représente l'action, les autres noeuds sont étiquettés avec les participants de l'action, et les liens entre les noeuds représentent les relations entre les participants de l'action : c'est Philippe qui joue, le jeu est le poker, et Yves est le partenaire de jeu. Lorsque l'action est le jeu, les champs de relations resteront toujours les mêmes : quelqu'un joue, il y a un partenaire, et un jeu. Le réseau peut donc être vu comme le remplissage d'un certain nombre de champs dans un schéma abstrait.

2.4.2 Les différents types de réseaux sémantiques

Le point commun à tous les réseaux sémantiques est qu'il s'agit d'une représentation graphique qui peut être utilisée aussi bien pour représenter de la connaissance que comme base pour faire des raisonnements à partir de connaissance pour des systèmes automatiques. Voici les quatre types de réseaux sémantiques que l'on peut considérer comme les plus couramment utilisés [126] :

- Les réseaux sémantiques de définition (Definitional network) utilisent de façon systématique la relation "sous-type" ou "est un" entre un concept et un sous-type de ce concept
- Les réseaux d'affirmation (Assertionnal networks) sont construits pour affirmer des propositions. Certains réseaux d'affirmation ont été proposés comme modèles pour la structure de la sémantique du langage.
- Les réseaux d'implication (Implicational networks) utilisent l'implication comme relation de base entre noeuds connectés. Ils peuvent être utilisés pour représenter des causes ou des inférences.
- Les réseaux exécutables (Executable networks) qui contiennent des mécanismes dínférence, ou rechercher un certain nombre de motifs ou d'associations.
- Les réseaux d'apprentissage (Learning networks) construisent ou étendent leur représentations en acquérant de la connaissance. La nouvelle connaissance peut changer l'ancien réseau en ajoutant ou supprimant des noeuds, ou en modifiant des valeurs numériques, appelés poids, associés avec les noeuds et les arcs.
- Les réseaux hybrides (Hybrid Networks) qui combinent plusieurs des techniques précdemment évoqué.

Certains de ces réseaux ont été élaborés explicitement pour implémenter des hypothèses concernant les mécanismes cognitifs humains, tandis que d'autres ont été élaborés avec pour objectif une efficacité informatique.

2.4.3 Ontologies

2.4.3.1 Définition de l'ontologie

En philosophie, l'ontologie est définie comme l'étude des propriétés générales de tout ce qui est, et est rattachée à la métaphysique. Cependant, depuis trente ans, on parle des ontologies comme d'un domaine de plus en plus autonome qui a connu un développement très important. Dans [106], François Rastier fait remarquer que "les ontologies sont des réseaux sémantiques comme on en connaissait depuis vingt ou trente ans. La nouveauté réside dans leur échelle sans précédent (par dizaine de milliers de "concepts") et dans leur utilisation pour servir de base de connaissance interlingue."

La définition la plus générale que l'on puisse faire d'une ontologie est qu'il s'agit d'une représentation graphique définissant formellement un domaine de connaissance. Il s'agit en général simplement d'expliciter un vocabulaire en définissant les termes nécessaires pour partager la connaissance liée à un domaine, ainsi que les relations entre ces termes et les contraintes du domaine dans un but de clarification.

2.4.3.2 Principaux types d'ontologies

Ontologie d'un domaine L'ontologie d'un domaine est utilisée pour représenter un domaine (la génétique, l'immobilier, les composants informatiques) sous forme de bases de connaissances. Par exemple, l'ontologie d'un site web peut être intéressante pour comprendre sa structure et peut être réalisée avant sa création. Elle présentera alors les mots-clé, les attributs, les instances relatives au domaine. Pour réaliser ce type d'ontologie, il existe des éditeurs de structure de base de connaissances tels que "Protégé" [52], qui est open-source et gratuit , ainsi que des formats vers lesquels ces ontologies peuvent être exportées, tels que RDF [97] et OWL [125]. Un document RDF (ou OWL) correspond à un multi-graphe orienté et étiqueté. Chaque arc est étiqueté par un prédicat et relie un noeud qui est le sujet à un noeud cible qui est l'objet. Ainsi, des mécanismes d'inférence peuvent être mis en oeuvre sur ces formats qui déduisent intégralement les conséquences des prédicats.

Ontologie informatique Les ontologies informatiques sont des outils qui permettent de représenter précisement un corpus de connaissances sous une forme utilisable par une machine. Elles représentent un ensemble structuré de concepts. Ceux-ci sont organisés dans un graphe dont les relations peuvent être des relations sémantiques et/ou des relations de composition et d'héritage (au sens objet). Un langage tel que le langage UML (Unified Modeling Language) permet de coder des ontologies informatiques.

2.4.3.3 Ontologies et sémantique

Remarquons que la construction d'ontologies ne relève pas du domaine de la linguistique. Le domaine des ontologies s'intéresse en effet aux référents et non aux signifiants et s'appuie par conséquent sur des connaissances extralinguistiques. Le but des ontologies est en effet de normer des relations entre des référents qui sont supposés être indépendants des langues. Les grandes ontologies telles que WordNet [39] ou EuroWordNet [138] comportent des réseaux différents pour les verbe, les noms, les adjectifs et les adverbes. Ainsi, la notion de morphème, qui est cruciale est linguistique, n'est pas du tout exploitée et des mots tels que "nager" et "nage" seront compartimentés dans des réseaux différents. Cependant, les relations sémantiques sont particulièrement importantes dans les ontologies et sont utilisées pour relier les termes entre eux. A titre d'exemple, EuroWordNet, développé en 1996, retient six sortes de relations conceptuelles : hyponymie, hyperonymie, holonymie, méronymie, synonymie, antonymie.

2.5 Fouille d'images et problématique d'indexation

Le domaine de la fouille d'images concerne l'ensemble des méthodes ayant pour but de permettre à un utilisateur d'accéder rapidement et efficacement à des images dans une base de données. Ce domaine a pris une importance particulière avec les progrès technologiques accomplis ces dernières décennies en matière d'acquisition et de stockage. En effet, l'émergence de grandes bases d'images pose le problème de leur organisation, de leur visualisation, et de la recherche d'une image correspondant à certains critères.

2.5.1 Indexation et besoin des usager

L'indexation n'est pas particulière aux bases d'image et peut s'appliquer à tout type de document. De plus, elle ne prend son sens que lorsqu'un ensemble d'usagers vont vouloir accéder aux documents de cette base en poursuivant un objectif donné. L'indexation d'un *document source* correspond alors à un *document de description* qui lui est attaché et qui permet d'évaluer la pertinence du document source au regard de l'objectif poursuivi par l'usager. Dans [5], Bachimont fait remarquer qu'aucun document n'est en soi un document source ou un document de description et que tout dépend du contexte d'usage. Il prend l'exemple d'une chronique écrite au sujet d'un film : cette chronique peut être lue pour elle-même, elle est vue alors comme un document source, mais elle peut être utilisée également pour indexer le contenu audiovisuel concerné. Ainsi, dans le cadre du domaine de l'indexation, les documents de description peuvent être d'une grande diversité quant à leur forme.

Une solution à ce problème est d'annoter les images, à savoir ajouter des métadonnées sur celles-ci. On peut distinguer deux types d'annotation. Tout d'abord, les annotations sur le contexte de l'image : pour une photographie, on

pourra ainsi mettre la date à laquelle elle a été prise, l'endroit où elle a été prise etc. Et les annotations sur le contenu, qui décrivent des éléments d'information présents dans l'image.

2.5.2 Problématiques de l'annotation d'images

2.5.2.1 Particularités de l'image par rapport au signe linguistique

Dans [5], Bachimont insiste sur une différence fondamentale existant entre l'image, en tant que forme sémiotique, et le signe linguistique en terme de relation entre signifiant et signifié (voir section 2.3.1). Il fait remarquer que le signe linguistique se caractérise par une relation arbitraire entre la forme signifiante et le contenu signifié, tandis que l'image est fortement liée au contenu signifié étant donné que son aspect comporte souvent une forte analogie avec celui du monde visible. Bachimont résume ainsi cette opposition : "Le signe qu'est l'image est un signe qui montre mais non un signe qui dit". Ainsi, l'image ne comporte pas l'équivalent des *mots*, à savoir une décomposition en unités signifiantes qui constituent une réalité objective de la linguistique et qui peuvent être utilisés directement pour accéder à des documents textuels. Bachimont en conclut que les images d'une base de données doivent être associée à "une sémiotisation explicite, dès lors qu'on veut les exploiter en fonction d'une certain valeur ou signification. Ainsi la pratique actuelle de la documentation à l'INA repose-t-elle sur la description linguistique puis en mots clés des programmes audiovisuels".

2.5.2.2 Choix des termes d'annotation

L'annotation des images d'une base par des mots du langage naturel pose un certain nombre de problèmes sémantiques que nous listons ici.

Polysémie Un mot annotant une image peut comporter plusieurs significations. Par exemple, "côte" peut désigner à la fois un os du corps humain (ou d'un animal) ou un terrain en montée. Les problèmes de polysémie sur les labels d'annotation tendent à détériorer les résultats de requêtes en diluant les résultats voulus dans un ensemble d'images non désirées. Des labels complémentaires sont alors nécessaires pour préciser la requête.

Hyponymie Le contenu d'une image peut être décrit par un ensemble de mots plus ou moins précis : "animal", "reptile", ou "serpent" sont différentes manières tout à fait raisonnables de décrire une même entité. Les personnes qui annotent ou recherchent des images vont considérer un niveau de spécification qui correspond à leur objectif ou à leur degré de connaissance. Un manque de concordance entre le niveau de spécification de la requête et celui de l'annotation va entraîner le fait que des images qui auraient intéressé l'utilisateur ne lui seront pas retournées par le système.

Synonymie La synonymie, ou d'une manière plus large le fait que des labels puissent avoir un sens proche, est un problème capital pour l'annotation d'images. En effet, un manque de cohérence dans l'annotation entraîne le fait qu'une requête ne suffira pas à obtenir l'ensemble des images que souhaite un utilisateur. Ainsi, certaines images de télévisions pourront être annotées "télé", "téléviseur" ou encore "télévisions", et plusieurs requêtes seront alors nécessaires à un utilisateur pour obtenir l'ensemble des images souhaitées. Dans les systèmes d'annotation collaborative (qui seront présentés en section 2.5.3.2), ce problème est particulièrement critique étant donné que les utilisateurs ne sont pas contraints d'utiliser un vocabulaire particulier.

2.5.2.3 Description d'une image

Des études mettent en évidence le fait qu'un groupe d'indexeurs utilise souvent un grand ensemble de mots différents pour annoter un même document [48] [53]. Ce problème de cohérence entre indexeurs provient de la difficulté même que constitue la tâche de description d'une image. Logiquement, la description devrait précéder l'interprétation et relever de l'objectif et de l'explicite. Or, il s'avère que, face à une image, on glisse rapidement vers le subjectif, et que dans la lecture du message visuel, la composante propre au récepteur est particulièrement importante. Dans [137], Kumiko Vézina évoque ainsi une expérience particulièrement intéressante de Sunderland où un enfant de 12 ans, un profane et un historien d'art décrivent un même tableau de J-E Millet "Le christ dans la maison de ses parents". Il est alors apparu que chacune des trois descriptions qui ont été données ont été significativement différentes :

- L'historien d'art a identifié précisément l'oeuvre quant à son exécution (auteur, date etc.), son style et sa symbolique.
- Le profane a dit qu'il s'agissait d'une image religieuse représentant la famille Sainte et qu'elle dégageait une impression de bonheur familiale.
- L'enfant a décrit les éléments présents dans la salle : il y a une femme à genou qui tient un enfant, il y a des copeaux de bois sur la plancher etc.

Chacune des trois personnes a ainsi décrit l'image en fonction de son vécu et da connaissance. Au final, les termes employés pour la description ont été totalement différents.

2.5.3 Méthodes d'annotation d'images

2.5.3.1 Annotation de documents à partir d'ontologies

Une ontologie est par définition un vocabulaire partagé par une communauté pour un domaine de connaissance donné (voir section 2.4.3). Ainsi, annoter des documents en utilisant une ontologie présente l'avantage d'utiliser des termes sur lesquels une communauté s'est mise d'accord et qui sont donc dépourvus d'ambigüités pour celle-ci. De plus, le fait d'avoir à disposition un lexique prédéfini attire l'attention de l'annotateur sur des structures sémantiques à annoter. Dans le cadre du projet de Web sémantique [10], de nombreux outils d'annotation de ressources web à partir

d'ontologies ont été créés, tels que OntoAnnotate [127], ou MnM [134]. Les possibilités propres aux ontologies peuvent être utilisées à partir des annotations ainsi produites pour mener à bien des mécanismes d'inférence ou effectuer une navigation conceptuelle. Cependant, certaines ontologies peuvent être amenées à changer : des concepts peuvent disparaître et d'autres apparaître. Cela signifie que des mises à jour doivent être effectuées automatiquement entre les documents annotés et les ontologies qui ont évolué. De plus, une ontologie concerne un domaine bien délimité et pour annoter des corpus très vastes, de nombreuses ontologies sont nécessaires. Ainsi, des recoupements peuvent apparaître et une traçabilité peut être nécessaire pour apparier des éléments appartenant à différentes ontologies [34]. Dans le cadre des travaux effectués sur le Web Sémantique, un certain nombre de travaux visent à annoter automatiquement des ressources web à partir d'ontologies [72], [143].

2.5.3.2 Annotation collaborative d'images

Récemment, les systèmes d'annotation (ou *tagging systems*), tels que Flickr connaissent un engouement grandissant. Ces systèmes permettent à des utilisateurs d'attacher des mots-clés (ou "tags") à des ressources internet (images, vidéos, pages internet) sans avoir de contraintes sur vocabulaire utilisé. Le succès d'un système comme Flickr, qui a dépassés le cap des trois milliards de photos en 2008 et rassemble 8,5 millions d'utilisateurs inscrits, prouve le réel intérêt d'un grand nombre d'internautes pour associer des informations sur le contenu de leurs images et la rendre ainsi accessible à un public très large. Il est intéressant d'étudier la manière dont les utilisateurs annotent les images dans ce type de systèmes et quels types de labels ils ont tendance à utiliser. Dans [119], les auteurs ont ainsi étudié divers aspects de ces questions, tels que la fréquence des labels, la quantité de labels attachés aux photos ainsi que la répartition de ces labels dans les catégories WordNet les plus courantes. En termes quantitatifs, il ressort de cette étude que 64% des images sont annotées avec moins de 3 labels. Les auteurs considèrent qu'un tel nombre de labels est insuffisant pour une annotation complète. Environ 23% des images sont annotées avec un nombre compris entre 4 et 6 labels et 14% des images sont annotées avec 7 labels ou plus. En termes de contenu, les auteurs affirment que 52% des labels peuvent être classifiés dans des catégories WordNet courantes. Parmis ceux-là, la catégorie la plus courante est le lieu qui correspond à 28% des labels classifiables. Viennent ensuite des objets (16%), des gens ou des groupes (13%), des actions ou des événements (9%), et des annotation de temps ou de date (7%). Ainsi, il ressort de cette étude d'une part que les utilisateurs n'annotent pas seulement les photos quant à son contenu, mais également quant au contexte dans lequel elle a été prise, et d'autre part que les photos sont très souvent insuffisamment annotées pour permettre une recherche efficace. Les auteurs proposent donc un certain nombre de stratégies permettant d'enrichir automatiquement les annotations produites par les utilisateurs.

2.6 Conclusion

À la lumière de ce chapitre, nous voyons que dans le titre même des travaux de thèse exposés ici, "Inférence de connaissances sémantiques" est un abus de langage. Nous nous l'autorisons néanmoins car il est couramment employé dans la littérature scientifique portant sur des travaux pour lesquels il s'agit d'annoter des images avec du vocabulaire provenant du langage naturel. Afin de donner davantage d'explications sur le syntagme "Inférence de connaissances sémantiques", il convient d'étudier comment les linguistes traitent la relation entre les mots et leur signification. Ce titre signifie que l'on souhaite mettre en relation des *faits de sens* présents dans les images que nous étudions avec un certain nombre de signifiants qui seront des mots du langage naturel et pourront être vus comme des représentants de *classes sémantiques visuelles* ayant une signification bien délimité pour un photo-interprète.

Habituellement, dans la littérature de vision par ordinateur, l'*annotation sémantique d'images* correspond généralement à un ensemble d'algorithmes et de procédures mettant automatiquement en rapport des images avec des termes lexicaux provenant d'une langue naturelle (généralement l'anglais). Joseph Courtès, dans [27], définit les langages naturels (le français, l'anglais, le chinois etc.) comme étant ceux pour lesquels où "le sujet humain qui les reçoit n'est qu'un usager et ne peut les modifier à sa guise". Les usages des termes d'un langage naturel sont le fruit d'une histoire et d'une culture et si le sujet humain le modifie, il ne sera plus compris par son environnement socioculturel (Joseph Courtès donne l'exemple des réactions suscitées par l'emploi que le mouvement surréaliste faisait de la langue française, tant au niveau syntaxique que sémantique). Au contraire, un langage artificiel est manipulable au gré des besoins, et la notion d'"usage" n'est pas prise en compte.

Annoter des images avec le langage naturel permet ainsi de permettre de les décrire par des termes intelligibles par une communauté socio-culturelle importante et de les rendre exploitables dès lors qu'on souhaite les utiliser en vue d'objectifs bien définis. Mais nous pensons ici que la richesse inhérente à la quasi-totalité des langues naturelles doit être exploitée au moins dans la composante structurelle de son lexique pour que le mot d'*annotation sémantique* ne soit pas usurpé. En effet, annoter une image par les mots "forêt", "végétation", et "toundra" sans prendre en compte en aucune manière les relations qui peuvent exister entre ces termes revient à faire du langage naturel une utilisation totalement minimaliste pour la fouille de données. En effet, dans [93], Georges Mounin dit que même si la sémantique structurale ne l'a pas en évidence de façon formelle, "tous les linguistes d'aujourd'hui demeurent d'accord qu'il doit y avoir une organisation quelconque du lexique, et de son contenu : les significations". Tous demeurent d'accord qu'il est impossible de penser que les mots sont présents d'une manière ou d'une autre dans notre tête sous la forme d'éléments totalement isolés les uns des autres. C'est cette conviction qui s'exprime quand on répète qu'une langue n'est pas une nomenclature (Saussure), un sac à mots (Harris), un empilement de noms (Whorf)."

Chapitre 3

État de l'art de l'extraction de sémantique

3.1 Extraction de sémantique dans les bases de données textuelles

La fouille textuelle est un domaine antérieur à celui de la fouille d'images et dispose à ce titre d'une littérature plus abondante. C'est pourquoi il est intéressant de l'étudier avant de s'intéresser à l'état de l'art dans le domaine de l'image. Notons cependant une différence de taille entre le domaine de l'extraction de sémantique dans l'image et celui de l'extraction de sémantique dans le texte : le constituant élémentaire du texte, le mot, contient intrinsèquement une "quantité" non négligeable de *sémantique*, tandis que le constituant élémentaire de l'image, le pixel, n'en contient quasiment pas. Et même si des caractéristiques de bas-niveau seront extraites dans l'image pour aider à son interprétation (texture, contours, etc.), ces caractéristiques de bas-niveau extraites dans l'image comporteront peu de signification. Ainsi, il semble que le *fossé semantique* existant entre l'image et le sens est bien plus large que celui existant entre le texte et le sens.

La fouille textuelle est devenu un domaine de recherche véritablement actif à partir du milieu des années 80. Abordé initialement à partir de méthodes logiques issues du domaine de l'intelligence artificielle, les méthodes statistiques à base d'apprentissage se sont progressivement imposées. Elles ont été portées notamment par l'émergence de grands corpus de données textuelles et par l'apparition des premières campagnes d'évaluation telles que TREC (Text Retrieval Evaluation Conference), et plus récemment DEFT (Défi francophone de Fouille de Textes). Les applications des méthodes de fouille textuelle sont multiples. Dans [102], les auteurs les classent selon les axes suivants :

- La recherche d'information classique, qui vise à retrouver des textes ou des extraits de texte pertinents à partir d'une requête exprimée avec des mots-clés.
- L'extraction de connaissances nouvelles ou la vérification d'un réseau ou arbre de connaissances existant.

- La classification de documents, qui peut soit correspondre à une tâche supervisée s'il existe préalablement un découpage des documents du corpus en différentes catégories, soit à une tâche non-supervisée s'il s'agit de faire émerger des regroupements entre documents.
- La segmentation de textes, qui a pour but de délimiter différentes parties ayant une certaine cohérence au sein d'un texte, et éventuellement de les nommer.
- Le profilage : il s'agit d'associer des aspects sémantiques ou lexicaux à des textes ou des fragments de texte dans le but de :
  - Reconnaître un auteur dans un ensemble de documents textuels.
  - Rechercher un des documents correspondant à des préférences fournies par des utilisateurs.
  - Détecter des tendances ou des opinions dans des discours.
  - etc.

La première étape de fouille textuelle est généralement de représenter un document textuel sous une forme permettant un traitement informatique, puis d'utiliser des techniques de fouille de données sur ces objets numérisés. Nous commençons donc ici par étudier les différentes représentations possibles d'un document textuel, avant de nous intéresser à différents modèles d'inférence de sémantique dans le texte.

3.1.1 Représentations vectorielles de documents

Nous ne discuterons ici que des représentations d'un document en sac-de-mots ("bag-of-words" en anglais), à savoir les représentations qui ne prennent pas en compte l'ordre des mots dans un document, et qui sont les plus fréquemment utilisées en fouille textuelle.

3.1.1.1 Vecteur binaire :

La représentation par vecteur binaire est la représentation la plus simple pour un document. Elle consiste à considérer un ensemble de M mots-clés $\{m_1, ..., m_M\}$ et à représenter un document par un vecteur à M composantes où l'indice i vaut 1 si le mot-clé i est présent dans le document et 0 s'il est absent de ce document.

Autrement dit, si d est le vecteur binaire représentant le document D, et V la taille du vocabulaire,

$$\forall i \epsilon [1...V], d(i) = \begin{cases} 1 & \text{si } m_i \epsilon \text{D} \\ 0 & \text{sinon} \end{cases}$$

3.1.1.2 Vecteur fréquentiel :

La représentation binaire a un pouvoir de description d'un document limité car la fréquence d'apparition d'un mot dans un document peut être une information importante. Ainsi, la représentation fréquentielle est une extension de la représentation binaire qui prend en compte l'occurrence d'un mot-clé dans un document. Le vecteur correspondant à un document aura donc sa composante i égale au nombre d'apparitions du mot-clé numéro i dans le document.

Plus formellement, si d est le vecteur fréquentiel représentant le document D, et n_D^i la fonction donnant le nombre d'occurrences du mot-clé i dans le document D, nous définissons d de la façon suivante :

$$\forall i \epsilon [1...V], \quad d(i) = n_D^i$$

.

Une des limites de la représentation fréquentielle est qu'un document long aura un vecteur de norme plus élevée qu'un document plus court, ce qui peut engendrer des problèmes dans certaines méthodes de clustering ou de classification où un document plus long sera pénalisé par rapport à un document plus court. Il est donc plus logique d'utiliser un vecteur qui a été normalisé par la longueur du document dont il a été extrait, et qui code donc la probabilité d'apparition d'un mot dans un document.

3.1.1.3 Vecteur TF-IDF :

Le vecteur TF-IDF tente de donner une représentation plus informative que la représentation fréquentielle en utilisant une normalisation par l'importance relative de chaque mot dans le corpus. En effet, certains mots dans un document apportent beaucoup d'information sur la nature ou le contenu d'un document même s'ils y sont peu souvent présents. Certains mots peuvent apparaitre au contraire avec des occurrences très élevées mais apportent une quantité d'information très faible. Une manière d'améliorer la représentation fréquentielle est donc de pondérer les différentes composantes du vecteur fréquentiel par un terme correspondant à la quantité d'information apportée par le mot associé à cette composante. Un exemple d'une telle mesure de quantité d'information est la loi de Zipf qui donne une observation empirique des fréquences d'apparition des mots dans un texte. Elle prévoit que dans un texte donné, la fréquence d'occurrence $f(n)$ d'un mot est liée à son rang n dans l'ordre des fréquences par la loi : $f(n) * n = K$, où K est une constante [147]. Des études ont montré que cette loi décrit de façon très efficace des corpus de textes en anglais [73] ainsi que d'en d'autres langues [82].
Mandelbrot généralise cette observation empirique en se basant sur l'hypothèse que le coût d'utilisation d'un mot est directement proportionnel au coût de stockage. On obtient ainsi la loi de Mandelbrot, dont la loi de Zipf n'est qu'un cas particulier : $f(n) * (a + bn)^c = K$, où K est une constante [85].
Plusieurs formules ont été proposées pour pondérer les composantes du vecteur fréquentiel de manière pertinente en utilisant la loi de Zipf. Le modèle le plus classique est celui pour lequel la première valeur est égale à la fréquence du mot dans le document (noté tf_i^d pour *term frequency*) et la seconde valeur est égale à $\log(\frac{|D|}{df_i})$ où $|D|$ est le nombre de documents du corpus et df_i est le nombre de documents qui contiennent le mot clé i. (df signifie *document frequency*). On peut l'écrire formellement de la façon suivante :

$$\forall i \epsilon [1...|V|], d_{tf-idf}^i = tf_i^d \log(\frac{|D|}{df_i})$$

.

3.1.1.4 Analyse sémantique latente (LSI/LSA)

La technique d'indexation la plus simple consiste à répondre à une requête d'utilisateur en construisant une représentation vectorielle de cette requête (vue comme un document textuel) et en la comparant aux représentations vectorielles extraites dans le corpus. Cette méthode se heurte à des limites liées à des problèmes de polysémie et de synonymie :

– Il est possible que dans certains documents du corpus, les termes de la requête soient présents mais employés dans un autre sens que celui recherché par l'utilisateur.
– A l'inverse, il arrive que le mot demandé ne se trouve pas dans un document pourtant pertinent pour le thème car c'est un synonyme qui est employé.

En effet, l'approche TF-IDF a des caractéristiques intéressantes en termes de discrimination mais elle apporte peu de réduction et donne relativement peu d'information en ce qui concerne la structure statistique du document. Pour résoudre ce problème, des méthodes de réduction de dimension ont été proposées, notamment la *Latent Semantic Analysis* (LSA) [32]. L'idée de Deerwester et al est de trouver un moyen de considérer le contexte d'un mot et les liens sous-jacents entre des termes dans le corpus pour régler en partie ces problèmes. En effet, si l'utilisateur cherche le mot *villa* et qu'un document contient uniquement le mot *pavillon*, le fait que d'autres termes du champ lexical tels que *bâtiment* et *jardin* soient, d'une part, présents en nombre dans ce texte et, d'autre part, en cooccurrence fréquente avec le terme de la requête *villa* ailleurs dans le corpus, permet d'affirmer que le texte est probablement pertinent.
Pour découvrir une telle structure *latente* dans le corpus, étant donné un corpus de N documents contenant M valeurs discrètes possibles, la LSA est fondée sur la matrice d'occurrences termes/documents A associée, normalisée ou non. Cette matrice est définie de la façon suivante : chaque colonne représente un document et chaque ligne i représente un terme. La valeur en (i, j) de cette matrice est donc le nombre d'occurrences du terme i dans le document j. Ainsi, si n_j^i est le nombre d'occurrences du i-ème terme dans le j-ème document, la matrice A s'écrit de la façon suivante :

$$\mathbf{A} = \begin{pmatrix} n_1^1 & n_1^2 & \dots & n_1^N \\ n_2^1 & n_2^2 & \dots & n_2^N \\ \vdots & \vdots & \ddots & \vdots \\ n_M^1 & n_M^2 & \dots & n_M^N \end{pmatrix}$$

La LSA consiste à employer la méthode de décomposition en valeurs singulières sur le matrice A. La décomposition en valeurs singulières d'une matrice M fournit la meilleure approximation aux sens des moindres carrés de la matrice M par une matrice de rang k. Ainsi, A est approximée par une matrice A_k de rang k, écrite comme le produit des trois matrices U, S_k et V :

$$A_k = US_kV^t$$

La matrice U, de dimension $M * k$ donne les coordonnées de chacun des M termes dans un espace linéaire de dimension k dans lequel les documents vont être projetés. La matrice S_k est une matrice diagonale de taille $k * k$, ses éléments diagonaux sont appelées valeurs singulières de A : ce sont les racines carrées non nulles des M valeurs propres de AA^t. V est une matrice de taille $k * N$ qui contient les coordonnées de chaque document dans le sous-espace linéaire. Les vecteurs de sac de mots sont ainsi projetés dans un espace réel de plus petite dimension dans lequel les documents peuvent être comparés à partir d'une mesure basée sur le calcul du cosinus de l'angle formé par deux vecteurs, pondéré par la matrice S_k. Notons que k est déterminé de façon empirique en fonction du corpus utilisé et du degré de performance voulu. On imagine aisément que, plus k est faible, plus on accélère le processus, mais plus on perd d'information. Cette approche permet des réductions significatives pour de grands ensembles de textes et améliore les performances d'indexation en prenant en compte le contexte d'utilisation d'un terme par rapport aux autres. On observe alors que les caractéristiques qui sont extraites par la LSA permettent de mettre en relief des notions linguistiques de base comme la synonymie ou la polysémie. En effet, dans l'espace de dimension réduite dans lequel sont projetés les vecteurs TF-IDF, on voit apparaitre des groupements correspondant à termes dont le sens est équivalent [9] [11].

3.1.2 Modélisations probabilistes du texte

Au cours des quinze dernières années, les méthodes de fouilles textuelles à base d'apprentissage sont montées en puissance, fortement soutenues par l'émergence de campagnes d'évaluation comportant de grands corpus linguistiques. Nous présentons ici ces modèles suivant deux grands axes : les modèles de mélange, sous-tendus par une hypothèse d'interchangeabilité entre les mots, et les modèles séquentiels, à travers les modèles dits *n-grammes*.

3.1.2.1 Modèles de mélange

Décrire des documents textuels selon un modèle de mélange revient à faire l'hypothèse que les mots d'un même texte sont interchangeables, à savoir que la loi jointe d'un ensemble de mots reste la même quel que soit leur ordre. En effet, un théorème classique formulé par de Finetti affirme que la loi de toute collection de variables aléatoires interchangeables peut s'écrire comme un mélange [30]. Ainsi, l'interchangeabilité entre variables aléatoires ne signifie pas qu'elles sont indépendantes et identiquement distribuées, mais qu'elles sont indépendantes et identiquement distribuées conditionnellement à une variable latente. Les modèles présentés dans cette section ont en commun d'être des modèles de mélange, quoiqu'ils soient de complexité variable.

Modèle "bayésien naïf" Le modèle bayésien naïf est un modèle génératif classique qui est utilisé notamment pour la classification de documents textuels plats dans les années 90. Sa simplicité et sa robustesse en font un modèle de référence qui est encore utilisé par des applications récentes tels que le contrôle parental ou le filtrage de spam [108] [35].
Soit un document D composé d'une séquence de mots $(w_1, w_2, ..., w_N)$, où N représente la longueur de la séquence et où les w_k sont à valeurs dans un vocabulaire de taille finie m. On note Θ l'ensemble des paramètres du modèle. Le modèle bayésien naïf repose sur l'indépendance conditionnelle des éléments de la séquence entre eux. On tire ainsi de l'équation précédente :

$$P(D|\Theta) = \prod_{k=1}^{N} P(w_k|\Theta)$$

L'expression du modèle bayésien naïf est donc simple, le modèle possède une inférence de complexité linéaire en fonction de la longueur de la séquence. Θ est composé de m paramètres θ_i correspondant chacun à la probabilité d'apparition du mot w_i.
Si le document D est représenté par un vecteur fréquentiel $(n_D^1, n_D^2, ..., n_D^m)$, où n_D^i représente donc le nombre d'occurrences du terme i dans le document, la probabilité de génération de ce document sachant le modèle Θ s'écrit :

$$P(D|\Theta) = \prod_{i=1}^{m} \theta_i^{n_D^i}$$

Modèle de mélange d'unigrammes Le modèle de mélange d'unigrammes consiste à introduire une variable latente z prenant ses valeurs dans un ensemble fini de thèmes possibles, conditionnellement à laquelle les mots sont indépendants [96]. Ce modèle est une généralisation du modèle bayésien naïf au sens où celui-ci correspond au cas particulier du modèle de mélange d'unigrammes pour lequel l'ensemble des valeurs possibles de z est un singleton. Plus précisement, les N mots w_i d'un document D sont générés selon le processus génératif suivant :
- z est choisie avec probabilité $p(z)$
- Pour chacun des N mots w_k :
  - w_k est choisi avec probabilité $p(w_k|z)$

Ainsi, la probabilité de D est :

$$p(D) = \sum_{z} p(z) \prod_{k=1}^{N} p(w_k|z)$$

Ce modèle suppose une homogénéité thématique du document car la valeur de la variable latente est choisie une seule fois par document et tous les mots sont ensuite choisis avec les paramètres correspondant à ce thème.

Modèle pLSA Le modèle pLSA (Probabilistic Latent Semantic Analysis) [9] est un autre modèle génératif de textes faisant l'hypothèse qu'un document D et un mot w_k sont conditionnellement indépendants étant donné une variable latente z prenant ses valeurs dans un ensemble fini. Elle suppose le processus génératif suivant :
Pour chaque mot w_k du document :
– Choisir un thème z_k selon la loi $p(z|D)$
– Choisir w_k suivant la loi $p(w_k|z_k)$, loi de probabilité conditionnée par le thème z_k.
Ainsi, la probabilité jointe d'un document et d'un mot est donnée par :

$$P(D, w_k) = p(D) \sum_z p(w_k|z)p(z|D)$$

pLSA essaie d'assouplir l'hypothèse d'homogénéité thématique faite par le modèle de mélange d'unigrammes. Elle "capture" ainsi la possibilité qu'un document contienne plusieurs thèmes, et $p(z|D)$ représente ainsi les poids de ces mélanges pour le document. Étant donné un nouveau document, ces probabilités doivent être recalculées. Ainsi, la taille du modèle croit linéairement avec le nombre de textes présents dans l'apprentissage. Le modèle LDA garde un principe similaire de génération du document mais propose une correction du dernier problème en introduisant non seulement une modélisation du document mais également du corpus.

Modèle "LDA' : Latent Dirichlet Analysis LDA [14] est une méthode générative pour un corpus de documents. L'idée de base est que les documents sont représentés comme des variables aléatoires sur des thèmes latents dont chacun est caractérisé par une distribution sur l'ensemble des mots. L'ensemble des paramètres β du modèle étant fixé, LDA suppose le processus génératif suivant :
– Choisir N suivant une loi de Poisson de paramètre σ.
– Choisir Θ suivant une distribution de Dirichlet de loi α.
– Pour chaque occurrence des N mots w_k :
  – Choisir un thème z_k suivant une loi multinomiale de paramètre Θ.
  – Choisir un mot w_k suivant la loi $p(w_k|z_k, \beta)$, loi multinomiale conditionnée par le thème z_k.
Dans ce modèle, la variable Θ est un vecteur de taille n qui prend ses valeurs dans le $(n-1)$ simplexe, on suppose que sa dimension est fixée et que la dimension du vecteur z, c'est-à-dire le nombre de thèmes possibles, est fixée également. La matrice β contient les probabilités des mots conditionnellement au thème, probabilités qui doivent être estimées à l'apprentissage. Les auteurs précisent que la distribution de Poisson n'est pas un élément critique de la méthode et qu'une autre loi plus réaliste peut-être utilisée. Si cette approche permet une modélisation plus crédible d'un texte que le modèle de mélange d'unigrammes en permettant de traiter des textes sémantiquement hétérogènes, la contrepartie est que l'apprentissage des paramètres devient un problème particulièrement délicat. Jordan et al. utilisent une méthode variationnelle [66] dont ils affirment qu'elles permettent d'estimer correctement les divers paramètres.

3.1.2.2 Modésentation séquentielle

Les modélisations que nous avons vu précédemment ne prennent pas en compte l'ordre des mots présents dans un document. Il peut être judicieux dans certains cas d'utiliser une modélisation plus réaliste du langage en gardant une information sur sa séquentialité. Ainsi, les modèles "n-grammes" estiment la probabilité de chaque mot conditionnnellement à la séquence des n mots précédents. Pour calculer la probabilité d'un texte, on fait l'hypothèse que chaque mot du texte ne dépend que des n mots précédents. Plus précisément, si $\{w_1, ..., w_p\}$ est une séquence de mots, on écrit (aux conditions de bord près) :

$$p(w_1, ..., w_p) = \prod_{i=1}^{p} p(w_i | w_{i-1}, w_{i-2}, .., w_{i-n})$$

L'inconvénient d'une telle modélisation est que le nombre de paramètres à estimer est potentiellement énorme. En effet, si le vocabulaire comporte N mots (N est en général de l'ordre de plusieurs dizaines de milliers), il est à priori nécessaire d'estimer N^n paramètres. Même pour un modèle bigramme ou trigramme, un tel ensemble de paramètres est très volumineux et difficile à estimer. Cependant, la méthode dite de *back-off* [71] permet de réduire considérablement le nombre de paramètres à estimer. Elle consiste à estimer la probabilité d'un mot sachant la séquence des n précédents dans le cas où cette séquence apparaît suffisamment souvent dans le corpus d'apprentissage, mais dans le cas contraire, seules les séquences des $n-1$ précédents (voire moins encore) vont être prises en compte. Ainsi, le nombre de paramètres à estimer va être grandement réduit et seuls les paramètres qui peuvent être correctement estimés à partir du corpus d'apprentissage vont être utilisés.

3.2 Construction automatique de hiérarchies

L'inférence automatique de la structure d'un lexique fait partie du domaine de la fouille textuelle. Ainsi, le programme TECHNOLANGUE, organisé en 2002 par le Ministère délégué à la recherche et aux nouvelles technologies, a comporté une évaluation sur ce thème à travers le projet CESART (Campagne d'Evaluation de Systèmes d'Acquisition de Ressources terminologiques) qui distinguait deux tâches :

- Extraction des termes pour la construction d'un référentiel terminologique dont l'application est l'enrichissement du référentiel et l'indexation libre des documents.
- Extraction de la relation de synonymie entre les termes.

Trois corpus en français des domaines spécialisés ont été construits : un corpus médical, un corpus sur le domaine de l'éducation, et un corpus politique. Les deux premiers ont été utilisés comme corpus de test, tandis que le troisième (corpus politique) a été utilisé comme corpus de masquage. Différents types de méthodes peuvent exister, faisant appel à des domaines variés (méthodes de clustering,

analyses morpho-syntaxiques, etc.), et nous nous intéressons ici seulement aux méthodes probabilistes. Avant d'exposer celles-ci, nous présentons d'abord les méthodes de sélection de modèles.

3.2.1 Méthodes de sélection de modèles

Lorsque la connaissance sur les données ne permettent pas de définir à priori un modèle dont la structure est fixée, il est nécessaire de réaliser une procédure de sélection de modèles : choix du nombre de composantes d'un mélange de lois, ordre d'une chaine de Markov, ordre d'auto-régression etc. La réponse la plus courante qui a été apportée à ce problème est de choisir comme modèle "optimal" celui qui optimise un critère pénalisé. Nous détaillons ici les trois critères les plus couramment utilisés : le Bayesian Information Criterion (BIC) [115], l'Akaike Information Criterion (AIC) [1], et le Minimum Description Length (MDL) [109].

3.2.1.1 Critère de BIC

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de variables indépendantes dont on souhaite estimer la densité f. Pour cela, on se donne une collection finie de modèles $\{M_1, \ldots, M_m\}$, chaque modèle M_i étant constitué d'une densité g_{M_i} et d'un paramètre θ_i de dimension K_i. Le BIC se place dans un cadre bayésien : θ_i et M_i sont vus comme des variables aléatoires munies d'une distribution à priori notées $P(M_i)$ et $P(\theta_i)$. Pour un modèle M_i donné, la distribution à priori du paramètre θ_i est notée $P(\theta_i|M_i)$. On note Θ_i l'espace de dimension K_i auquel appartient θ_i.
Le BIC cherche à sélectionner le modèle le plus vraisemblable au vu des données :

$$M_{BIC} = \arg\max_{M_i} P(M_i|X) \tag{3.1}$$

Or, d'après la formule de Bayes :

$$P(M_i|X) = \frac{P(X|M_i)P(M_i)}{P(X)}$$

En supposant que tous les modèles sont équiprobables, la recherche du modèle vérifiant 3.1 ne nécessite que le calcul de $P(X|M_i)$, qui s'obtient par intégration de la distribution jointe du vecteur des paramètres θ_i et des données X.

$$P(X|M_i) = \int_{\Theta_i} g_{M_i}(X, \theta_i) P(\theta_i|M_i) d\theta_i.$$

Ce résultat peut être approché en utilisant la méthode d'approximation de Laplace [80]. Celle-ci fournit, en négligeant les termes d'erreurs dans le cas d'un nombre d'échantillons très important, l'expression suivante :

$$\log(P(X|M_i)) = \log(g_{M_i}(X, \hat{\theta}_i)) - \frac{K_i}{2}\log(n)$$

.

où $\hat{\theta}_i$ est l'estimateur du maximum de vraisemblance de θ_i. Le BIC correspond à l'approximation de $-2\log P(X|M_i)$. Ainsi, il sélectionne le modèle minimisant la quantité suivante :

$$BIC_i = -2\log(g_{M_i}(X, \hat{\theta}_i)) + K_i \log(n)$$

Le critère de BIC fournit des résultats très satisfaisants pour un grand nombre d'applications telles que l'optimisation du nombre de composantes dans un modèle de mélanges [43]. Cependant, il ne faut pas oublier que le BIC est obtenu par une approximation dans un cadre asymptotique sur le nombre d'échantillons. Lorsque la taille du corpus d'apprentissage est insuffisante, son application peut fournir des résultats inexacts.

3.2.1.2 Critère AIC

L'AIC est une méthode de sélection de modèle basée sur l'information de Kullback-Leibler (K-L) [75]. Soit f la densité de probabilité constituant la "vérité terrain" et g la densité de probabilité estimant f parametrée par $\theta \in \Theta$. L'information de K-L $I(f, g)$ correspond à une distance entre f et g et s'écrit :

$$I(f, g) = \int f(x) \log \frac{f(x)}{g(x)}$$

.

Trouver la meilleure approximation de f consiste à trouver le vecteur de paramètres θ minimisant cette quantité. Cependant, f étant inconnue, il est impossible d'utiliser directement l'information de K-L "attendue" (expected K-L information). L'information de K-L peut s'écrire :

$$I(f, g) = \int f(x) \log f(x) - \int f(x) \log g(x|\theta)$$

.

Ce qui correspond à l'expression suivante :

$$I(f, g) = E_f[\log f(x)] - E_f[\log g(x|\theta)]$$

.

Cependant, le premier terme de cette différence ne dépend pas de θ. Donc, seul le deuxième terme $E_f[\log g(x|\theta)$ doit être estimé pour minimiser l'information de K-L, cette quantité s'appelle l'information de K-L "attendue". Or, Akaike a montré que, pour obtenir un critère de sélection de modèle efficace, il était nécessaire d'estimer [1] :

$$E_y[E_x[\log g(x|\hat{\theta(y)})]] \tag{3.2}$$

où la partie intérieure correspond à l'information de K-L attendue où θ a été remplacé par l'estimateur du maximum de vraisemblance sur les données y. x et y sont vus comme deux lots de données indépendants de même distribution. Akaike trouve alors que l'estimateur du maximum de vraisemblance est un estimateur biaisé de 3.2 mais dont le biais est égal la dimension du vecteur θ, que l'on note K. Ainsi, pour minimiser 3.2, il est nécessaire de minimiser $\log \mathcal{L}(\hat{\theta}|D) - K$, où D correspond aux données d'apprentissage. Akaike multiplie cette expression par -2 pour obtenir le critère AIC :

$$-2 \log \mathcal{L}(\hat{\theta}|D) + 2K$$

.

Un certain nombre de critères ont été dérivés du critère AIC. On cite ici le critère AIC_c [128] qui peut être utilisé pour améliorer les résultats lorsque la taille de la base d'apprentissage est faible, et le critère QAIC [20] qui apporte un terme correctif lorsque les données d'apprentissage comportent une forte dispersion.

3.2.1.3 Principe de la minimisation de la CS

Le principe de minimisation de la complexité stochastique consiste à dire que la structure d'une information de nature quelconque aura été d'autant mieux comprise que l'on est capable de transmettre cette information avec un nombre minimum de bits. Pour une image donnée, on dira qu'un codage plus court de l'image implique une meilleure compréhension de l'image. Il convient maintenant de préciser ce que nous entendons par le terme "complexité".

Complexité de Kolmogorov : Analysons tout d'abord la notion de complexité telle que l'a introduite Kolmogorov dans les années soixante. La complexité de Kolmogorov $C(X)$ d'une séquence de nombres $x_1, x_2, ...x_N$ est égale à la longueur du plus court programme (mesuré en bits) qui peut l'engendrer. La complexité de Kolmogorov est alors maximale lors qu'il n'existe pas de programme plus court qu'une simple énumération. Dans ce cas, si la longueur de la suite est n, nous avons alors $C(X) = n + Cte$. Cette définition semble particulièrement utile pour décrire l'information contenue dans une image. Plus une image est difficile à décrire et plus sa complexité de Kolmogorov est grande. Cependant, cette définition se heurte à des limitations importantes.
La première difficulté est directement liée au calcul de cette complexité. En effet, dans la mesure où une suite de longueur n a une complexité d'au plus $C(X) = n + Cte$, nous pourrions penser qu'afin de trouver le plus court programme, il suffit d'essayer ceux de moins de n bits et d'analyser leurs résultats. Le plus petit programme ayant engendré la suite X permettrait alors d'en déduire la complexité de Kolmogorov. Cependant, parmi les programmes testés, un certain nombre ne s'arrêteront jamais et il n'est pas possible de savoir à priori lesquels. Il est donc impossible de calculer la complexité de Kolmogorov en un temps fini. De plus, Rissanen [109] souligne une deuxième limitation quant à l'utilité de cette

complexité. En effet, notre but est de déterminer le meilleur modèle permettant de décrire la séquence X, c'est-à-dire la structure sous-jacente à cette suite. Or, il apparait difficile de déterminer cette structure à partir du programme le plus court. Il semble plus efficace à priori d'analyser la complexité de l'image au travers du modèle dans lesquels ces propriétés sont faciles à identifier.

Information de Shannon : Une façon de résoudre le problème de non calculabilité de la complexité de Kolmogorov avait déja été proposée par Shannon en 1943 [117]. L'idée sous-jacente est qu'une réalisation apporte d'autant plus d'information qu'elle est improbable. La quantité d'information d'une suite X est donc directement liée à sa probabilité d'apparition $P(X)$:

$$C(X) = -\log(P(X))$$

Même si l'approche de Shannon est très différente de celle de Kolmogorov, dont l'ambition était de déterminer la complexité d'une suite indépendamment de sa loi de probabilité, il est possible de montrer que la valeur moyenne de la complexité de Kolmogorov d'une série de réalisation de suites X issues d'une certaine loi de probabilité est en fait égale, quand le nombre de réalisations est grand, à l'espérance mathématique de leur quantité d'information (c'est-à-dire l'information de Shannon). Ceci implique donc que l'information de Shannon est une approximation du nombre de bits pour coder une suite lorsque la loi de probabilité est connue.

Complexité Stochastique L'approche de Shannon reste cependant beaucoup moins générale que celle de Kolmogorov. En effet, la quantité d'information de Shannon suppose que la loi de probabilité ayant permis d'engendrer une séquence est connue. Cela suppose que si l'on souhaite transmettre une suite X avec une approche de type Kolmogorov, il suffit de transmettre un nombre de bits égal à la complexité de Kolmogorov. En revanche, avec une approche de type Shannon, la transmission d'un nombre de bits égal à la quantité d'information de Shannon ne sera pas suffisante pour reconstruire cette suite dans la mesure où il est nécessaire de connaitre la loi de probabilité $P(X)$, c'est-à-dire le modèle, pour reconstruire cette suite. Rissanen a introduit la complexité stochastique pour pallier ce problème. Cette notion consiste à substituer à la complexité de Kolmogorov le nombre de bits qu'il faudrait pour coder la suite avec un code entropique auquel il faut ajouter le nombre de bits nécessaire pour décrire le modèle probabiliste permettant de déterminer ce code. La complexité "stochastique", dénommée ainsi par opposition à la complexité "algorithmique" de Kolmogorov, permet alors de définir une mesure de la complexité intégrant un terme relatif aux modèles sous-jacents aux données.

3.2.1.4 Minimisation de la complexité stochastique

Rissanen proposa dès le début des années 70 un principe basé sur la minimisation de la complexité, et dénommé principe de la longueur de description minimale (ou

MDL, pour l'anglais : "Minimum description length"). Même si dès le début, l'expression de cette longueur de description est analogue à la complexité stochastique dans la plupart des exemples qu'il traite, ce n'est qu'en 1989 qu'il introduit explicitement cette notion [110] dans un ouvrage de synthèse de ses principaux travaux sur ce thème. Ce principe propose ainsi un critère permettant de choisir le meilleur modèle parmi un jeu de modèles, et ceci sans connaissance à priori du véritable modèle sous-jacent.
Définissons une classe de modèles de taille finie $M = \{M_k\}_{k=1}^{K}$ où chaque modèle est défini par un vecteur de paramètres θ_k dont la taille peut varier avec k. Soit X l'échantillon que nous voulons étudier. La complexité stochastique associée au modèle M_k est la longueur de code $D(X, M_k)$ nécessaire pour décrire l'échantillon. Cette longueur de code se décompose en deux parties : la longueur de code nécessaire pour décrire les données connaissant les paramètres du modèle et la longueur du code nécessaire pour décrire les paramètres du modèle.

$$D(X, M_k) = D(X|\theta_k) + D(\theta_k)$$

Le premier terme peut être vu comme un terme d'attache aux données et le deuxième comme un terme de régularisation.
Ce principe permet de donner un choix pour le terme de régularisation dans une approche bayésienne. Une propriété intéressante est qu'il n'est pas nécessaire de fixer un terme de pondération entre le terme d'attache aux données et le terme de régularisation dans la mesure où ils sont tous les deux exprimés dans la même unité, à savoir le bit. La complexité stochastique est donc une quantité à minimiser qui ne contient aucun terme à régler de la part de l'utilisateur.
La longueur de codage est traditionnellement séparée en deux composantes [109] :

$$CS(X, M) = CS(M) + CS(X|M)$$

Le premier terme correspond à la longueur de codage du modèle, et est donc une mesure de sa complexité. Le deuxième terme correspond à la longueur de codage des données conditionnellement au modèle, et mesure donc l'attache aux données. Les lots d'images exemples étant supposées indépendantes pour chaque concept, on peut sommer la longueur de description sur les concepts :

$$C(X, M) = \sum_{c=1}^{n} [C(X_c|M) + C(M_c)] \tag{3.3}$$

.

3.2.1.5 Méthodes de construction automatique de hiérarchies

Nous présentons ici les principales méthodes de construction automatique de hiérarchies utilisant une approche probabiliste générative avec sélection de modèle. D'autres types de méthodes existent cependant, utilisant typiquement des approches *agglomerative bottom-up* [38] ou encore des approches *divide-and-conquer top-down* [60] mais ne rentrent pas dans ce cadre et ne seront pas traitées ici.

"PAH" : Probabilistic Abstraction Hierarchies Dans [116], les auteurs présentent une approche bayésienne pour apprendre des hiérarchies dites "abstraites" à partir de données de différentes sortes (les auteurs l'appliquent aux domaines du texte et de l'expression des gènes). Cette méthode utilise un modèle sous forme d'arbre dont chaque noeud est associé à un modèle probabiliste, appelé *Class-specific Probabilistic Model* (CPM). Les données ne peuvent être générées qu'à partir des feuilles de l'arbre, de sorte qu'un modèle correspond à une loi de mélanges dont les composantes sont les CPM des feuilles de l'arbre.
Plus précisement, une hiérarchie probabiliste abstraite (PAH) A est un arbre T avec des noeuds $V = \{v_1, \ldots, v_m\}$ et des arcs non-orientés E, et ayant k feuilles : $\{v_1, \ldots, v_k\}$. Chaque noeud v_i est associé à un CPM M_i, définissant une distribution sur s. Une distribution multinomiale est également définie sur les feuilles $\{v_1, \ldots, v_k\}$, dont on note θ les paramètres.
Cette modélisation est très générale et n'impose pas un type de modèle particulier pour les CPMs. Les données ne pouvant être générées que par les feuilles, une variable cachée C est introduite qui, pour chaque élément s de S, prend une valeur c comprise dans $\{1, \ldots, k\}$ qui correspond à l'index de la feuille qui a généré s. Ainsi, on écrit :

$$P(s, c|A) = P(C = c|\theta)P(s|M_c)$$

Les noeuds internes de l'arbre servent ainsi à exprimer une probabilité à priori des modèles qui seront choisis pour générer les données. La vision intuitive des auteurs est que des feuilles qui sont proches dans la hiérarchie doivent avoir des CPMs similaires. Pour obtenir ce résultat, ils introduisent une probabilité à priori sur les hiérarchies A pénalisant la dissimilarité entre des CPMs reliés par un arc dans cette hiérarchie par une fonction de distance entre distribution $\rho(M, M')$. Ainsi, le probabilité à priori sur les hiérarchies s'écrit :

$$P(A) \propto \prod_{(i,j)\in E} \exp(-\lambda\rho(M_i, M_j))$$

où λ est un paramètre pondérant l'importance de la pénalisation de la distance entre les modèles.
Étant donnée une base de données D correspondant au domaine S, la construction automatique de la hiérarchie est effectuée en déterminant la hiérarchie A qui maximise $P(A|D) \propto P(A)P(D|A)$. Cela consiste à chercher un compromis optimal entre un modèle ayant des noeuds voisins similaires et un modèle décrivant de manière correcte les données. Ce problème de maximisation est extrêmement complexe et les auteurs fournissent un algorithme itératif permettant de trouver un maximum local de $P(A|D)$. La boucle de cet algorithme comporte une étape consistant sélectionnant à modifier la structure de l'arbre, puis une étape "Expectation" et une étape "Maximisation" combinant l'algorithme *Expectation Maximization* [33] et l'algorithme *structural Expectation Maximization* [46]. Il est possible de montrer que la log-probabilité $\log P(A|D)$ augmente à chaque boucle, ce qui assure la convergence de l'algorithme vers un maximum local.

"hLDA" : Hierarchical Latent Dirichlet Analysis La hLDA [12] décrit une modélisation permettant de construire par apprentissage une hiérarchie de thèmes tout en permettant à celle-ci d'évoluer quand de nouvelles données apparaissent. Ce problème de sélection de modèle est traité en spécifiant un modèle génératif probabiliste des structures hiérarchiques et en appliquant un apprentissage de ces structures à partir de données. Le *mécanisme* probabiliste sous-jacent est une distribution de partitions d'entiers appelé le "processus du restaurant chinois", ou *Chinese restaurant process* (CRP) [3], qui est étendu au cas d'une hiérarchie de partitions dans un modèle appelé *Nested Chinese restaurant process*.
Détaillon à présent le processus que modélise le CRP. On considère un restaurant chinois contenant une infinité de tables auxquelles vont s'asseoir M clients arrivant l'un après l'autre. Le premier client s'assied à la première table. Ensuite, le consommateur m a le choix entre s'asseoir à une table déja occupée ou s'installer à une nouvelle table. La distribution est la suivante :

$$p(\text{table occupée } i|\text{clients précédents}) = \frac{m_i}{\gamma+m-1}$$
$$p(\text{table inoccupée suivante}|\text{clients précédents}) = \frac{\gamma}{\gamma+m-1}$$

où m_i est le nombre de clients précédemment installés à la table i, et γ un paramètre du modèle. Après que les M clients se soient assis, l'agencement des clients au différentes tables définit une partition de M entités discrètes. Le CRP permet ainsi de décrire la relation existant entre diverses composantes d'un mélange et des données.
Dans [12], le CRP est étendu selon une version hiérarchique, appelé *nested CRP*, pour permettre de modéliser des données qui sont associées à diverses composantes de mélange selon un chemin dans une hiérarchie. Ce modèle suppose que le nombre de restaurants chinois est infini. Cependant, un de ces restaurants est supposé être le restaurant *racine* et sur chacune de ses tables est posée une carte avec l'adresse d'un autre restaurant. Sur toutes les cartes de ces restaurants sont posées également des cartes vers d'autres restaurants, et cette structure se répète indéfiniment. Chaque restaurant est référencé de manière unique, et les restaurants sont ainsi organisés suivant un arbre comportant une infinité de branches.
Le processus se déroule alors de la manière suivante : un touriste entre dans le restaurant *racine* et choisit une table à partir de l'équation 3.2.1.5. Le lendemain, il va au restaurant correspondant à la carte qu'il a trouvé sur la table du restaurant de la veille et choisit de nouveau une table de la même manière. Il procède de la sorte pendant L jours. A la fin de séjour, il a ainsi visité L restaurants, qui définissent un chemin dans l'arbre des restaurants. Quand M touristes ont séjourné L jours dans cette ville, on obtient alors un sous-arbre de l'arbre des restaurants.
Le *nested CRP* permet alors de modéliser un corpus de M documents de la manière suivante :

- Pour chaque document $d_i \in \{d_1, \ldots, d_M\}$:
  - Soit c_1 le "restaurant racine"
  - Pour chaque niveau $l \in \{2, \ldots, L\}$
    - Une table est choisie selon l'équation 3.2.1.5. Soit c_l le restaurant correspondant à cette table.

- Chacun des L restaurants visités correspond à un thème. Un vecteur θ de dimension L de proportions de ces thèmes est choisi selon une distribution de Dirichlet.
- Pour chaque mot w_n du document, $n \in \{1, \ldots, N\}$:
  - Un thème $z \in \{1, \ldots, L\}$ est choisi selon une loi multinomiale de paramètre θ.
  - w_n est choisi selon la probabilité $p(w_n|z, \beta)$, où β est un paramètre du modèle.

Au contraire du modèle LDA dont la structure est "plate", le modèle hLDA organise donc les thèmes selon un arbre fixé de taille L et chaque document est supposé comme étant généré selon un seul chemin à travers l'arbre. La structure de l'arbre est alors inférée en utilisant un échantillonneur de Gibbs [50] en séparant l'échantillonnage en deux temps :

- Étant donné un état courant du CRP, les thèmes z sont échantillonnés pour tous les mots de chaque document selon le modèle LDA sous-jacent selon un algorithme développé dans [55].
- Étant donné les valeurs des variables cachées, les chemins parcourus dans l'arbre pour chaque documents sont alors échantillonnés selon la probabilité à priori associée.

3.3 Extraction de sémantique dans les images

La recherche d'images par le contenu a été un sujet qui a motivé beaucoup de recherches ces dernières années. Tandis que les anciennes architectures de recherche d'images utilisaient des requêtes par présentation d'images exemples, il est apparu assez rapidement comme indispensable qu'un système de recherche d'images vraiment opérationnel devait pouvoir recevoir des requêtes formulées en langage naturel. Les systèmes sont généralement annotés automatiquement par des mots-clés, ce qui permet ensuite à l'utilisateur de spécifier sa requête à travers un langage de description naturel des concepts visuels. Les deux problèmes qui sont rattachés à celui-là sont :

- L'annotation automatique d'images nouvelles.
- La recherche d'images de la base de données, basée sur une requête sémantique.

Nous présentons ici un point et une réflexion sur l'état de l'art de l'extraction de sémantique dans toutes sortes d'images, puis nous présenterons quelques descriptions usuelles de l'image avant de voir deux types d'approches pour faire de l'annotation sémantique d'images : celles faisant directement le lien entre les caractéristiques symboliques extraites dans l'image et les annotations sémantiques, et celles appliquant des méthodes textuelles à partir d'une collection de caractéristiques symboliques extraites de l'image.

3.3.1 Annotation sémantique vue comme un processus de classification

Les approches qui tentent de faire le lien directement entre le bas-niveau et le haut-niveau font une modélisation probabiliste directe en calculant le maximum à posteriori des annotations sachant les observations. Pour avoir des annotations plus précises en décrivant certaines régions plutôt que d'attacher des termes à l'ensemble de l'image, ces méthodes segmentent l'image soit par une grille régulière, soit à partir des caractéristiques de bas-niveau extraites de l'image.

3.3.1.1 Problématique d'annotation d'une image

Considérons une base d'images $I = \{I_1, ..., I_N\}$ d'images I_i et un vocabulaire sémantique $L = \{w_1, ..., w_T\}$ d'étiquettes sémantiques w_i décrivant si l'image vérifie ou non, contient ou non, un concept donné : Par exemple "extérieur" ou "intérieur", "végétation", "tigre" etc. Le but de l'annotation est d'extraire un ensemble d'étiquettes w décrivant I de façon optimale au regard d'une norme donnée. L'image est dite annotée "faiblement" si l'absence de l'étiquette w_i n'implique pas nécessairement que le concept soit absent dans l'image [135].

3.3.1.2 Etiquettage supervisé

Vu comme un étiquettage supervisé, l'étiquettage est formulé comme étant composé de T problèmes de détection déterminant la présence ou l'absence des concepts dans L. Considérons ainsi le i-ème problème de détection et la variable Y_i valant 1 si l'image considérée est étiquetée par w_i, 0 sinon. Le but est d'inférer l'état de Y_i en minimisant la probabilité d'erreur, pour tout $i \in \{1, ..., T\}$. En notant X la variable correspondant à un vecteur de caractéristiques extraites dans l'image, on peut résoudre ce problème en utilisant des résultats bien connus de théorie de la décision [4] en posant que le concept est présent si :

$$P_{X|Y_i}(X|1)P_{Y_i}(1) \geq P_{X|Y_i}(X|0)P_{Y_i}(0)$$

.

où X est un vecteur aléatoire contenant les caractéristiques de bas-niveau visuelles extraites de l'image. $P_{X|j}$ est la densité de probabilité conditionnelle sachant la classe $j \in \{0, 1\}$, et $P_{Y_i}(j)$ est la probabilité à priori de cette classe.
L'apprentissage consiste à considérer, pour tout les concepts ω_i, l'ensemble D_1 des images annotées par l'étiquette w_i et l'ensemble D_0 contenant toutes les autres images, et à utiliser une procédure d'estimation de densité pour estimer $P_{X|Y_i}(x|j)$ à partir de D_j, $j \in \{0, 1\}$. Ainsi, ce type de méthode exige l'apprentissage, pour tout i, de la classe "non concept i". Ainsi, si le concept i est présent dans certaines images mais n'est pas explicitement annoté par l'étiquette w_i correspondante, la précision de la procédure d'apprentissage s'en trouve considérablement amoindrie. De plus, il est nécessaire que le lot d'apprentissage soit particulièrement grand dans le cas où la taille du vocabulaire est importante.

Ce type de processus d'apprentissage a été abondemment utilisé dans les premiers travaux d'annotation qui se sont ainsi focalisés sur ce type d'apprentissage supervisé en prenant en compte des concepts spécifiques : différencier des scènes d'intérieur et des scènes d'extérieur [129], des peintures et des photographies [29], des êtres humains et des animaux [42], des villes et des paysages [133].

3.3.1.3 Prise en compte de la spatialité

Assez peu de travaux essaient d'introduire la spatialité des régions extraites dans l'image pour l'annotation automatique. Pourtant, premièrement, certaines zones correspondant à une même annotation sémantique peuvent contenir des régions contenant des caractéristiques différentes dont l'agencement spatial est primordial et qu'il convient de prendre en compte. Deuxièmement, la répartition spatiale des différentes zones sémantiques peut aussi être intéressante à modéliser, pour supprimer par exemple des incohérences. Nous citons ici deux travaux qui étudient respectivement ces deux points.
Le travail effectué dans [2] porte sur l'annotation d'images satellitaires et comporte deux étapes. La première est non-supervisée et apprend automatiquement à partir de l'image un certain nombre de modèles de "régions prototypes" à partir de caractéristiques spectrales, radiométriques, et texturales. Le nombre de types de ces régions est un paramètre du modèle défini à l'avance, et ces régions ne sont pas étiquetées. Des relations floues sont définies entre ces différentes régions : "entouré par", "recouvre", "adjacent", "à droite", "à gauche", "disjoint", "proche", "éloigné", "au-dessus" et "au dessous". Les auteurs proposent alors un modèle paramétrique, appelé "grammaire visuelle", permettant de calculer la probabilité d'une configuration spatiale de régions. L'utilisateur va alors définir un certain nombre de labels sémantiques, et va fournir pour chacune un lot d'images d'apprentissage. Chacune de ces images sera partitionnée en régions prototypes et un lot de paramètres associé à la grammaire visuelle pour ce label sera ainsi estimé. Cette méthode se base donc sur une inférence en deux étapes : l'une permet de passer des pixel aux régions par classification directe des pixels suivie d'un lissage permettant d'obtenir des régions homogènes, et l'autre passe d'un ensemble de régions à une scène en fonction du type des régions et de leur configuration spatiale. Le système est alors capable de reconnaître des zones particulières qui auraient été plus délicates à retrouver sans la prise en compte de la spatialité entre les régions protoypes comme "zone côtière" (le système localise la ville et la mer qui lui est adjacente) ou "nuage" (le système localise le nuage et son ombre sur le sol).
Dans les travaux décrits dans [95] cités précédemment, la modélisation des caractéristiques de Gabor est par la suite enrichie en imposant des contraintes sur les annotations à partir de la proximité des fenêtres. Ainsi, un MRF (Markov Random Field) est utilisé pour décrire une énergie d'interaction entre fenêtres voisines, qui permet d'éliminer certaines configurations incohérentes (parking entouré par des tuiles correspondant à de l'eau) et de prendre en compte la spatialité des différentes annotations dans l'image par une énergie d'interaction

entre tuiles voisines.

3.3.2 Application de techniques textuelles à l'image

Les méthodes de fouilles d'images actuelles utilisent pour la plupart des modélisations qui ont prouvé leur efficacité dans la recherche de documents textuels.

3.3.2.1 Modèles par variables latentes

L'idée fondamentale de ces méthodes [74] [79] [81] [8] est d'introduire un lot de variables latentes qui codent les états cachés dans l'image. Chaque état définit une distribution jointe sur l'espace des labels sémantiques et les descripteurs de l'image (caractéristiques calculées en certaines zones de l'image). Au cours de l'apprentissage, un lot d'annotations est fourni à chaque image, l'image est segmentée en une collection de régions, et un algorithme non supervisé traite l'ensemble de la base pour estimer la probabilité jointe des mots et des caractéristiques visuelles. Étant donnée une nouvelle image à annoter, des vecteurs de caractéristiques visuelles sont extraits, la probabilité jointe est calculée avec ces vecteurs, les variables d'état sont marginalisées et on cherche le lot de labels qui maximise la densité jointe du texte et des caractéristiques extraites dans l'image. Le modèle global est de la forme :

$$P_{X,w}(\chi, w) = \sum_{l=1}^{S} P_{X,w|L}(\chi, w|l) P_L(l)$$

où S est le nombre d'états possibles de L, χ est l'ensemble des vecteurs de caractéristiques extraits à partir de I, et w est l'étiquettage de cette image. On voit donc que le modèle est un modèle de mélange classique, où la variable latente ne contient à priori pas de sémantique. Dans un soucis de simplification, et pour éviter des problèmes dûs au couplage entre le vecteur χ qui a une valeur continue, et w qui est à valeur discrète, les caractéristiques visuelles et les étiquettes sont souvent supposées être des variables indépendantes conditionnellement à la variable latente :

$$P_{X,w|L}(\chi, w) = P_{X|L}(\chi|l) P_{W|L}(w|l)$$

Le modèle étant un modèle de mélange, l'algorithme expectation-maximization (EM) est la plupart du temps employé pour mener à bien l'apprentissage.
Dans [136], pour pallier le problème d'une segmentation préalable, les auteurs introduisent en plus d'une variable latente l associée à chaque image un mélange de 5 lois gaussiennes pour chaque image. Le poids de chaque gaussienne dans l'image correspond à la présence plus ou moins importante de chaque concept dans l'image. Ainsi, la distribution dans chaque image s'écrit :

$$P_{X|L}(x|l) = \sum_{i=1}^{4} \pi_i G(x, \mu_i^l \sigma_i^l)$$

où $\sum_{i=1}^{4} \pi_i = 1$, (μ_i^l, σ_i^l) étant la moyenne et la variance de la i-ème gaussienne de la l-ième image. Ainsi, la présence d'un concept est déterminée par l'estimation des paramètres du mélange de gaussiennes : les paramètres des gaussiennes sont comparés avec les lois des gaussiennes estimées à l'apprentissage pour chaque gaussienne en utilisant la distance de Kullback-Liebler.
Dans la thèse de Pecenovic [100], l'auteur propose une adaptation de la *Latent Semantic Analysis* à la recherche d'images par présentation d'images exemples. Utilisant différentes caractéristiques de bas-niveau extraites à partir de vecteurs de texture et d'histogrammes de couleurs, une décomposition en valeurs singulières permet de transformer le vecteur de description de l'image en un autre vecteur dans un espace de plus petite dimension. Les caractéristiques réelles des vecteurs de caractéristiques calculés dans l'image sont gardées pour créer la matrice de co-occurrence. Une manière nouvelle de passer des attributs de bas niveaux à la notion d'occurrence est introduite. Après apprentissage et calcul de la décomposition en valeurs singulières, deux matrices sont obtenues, la première permettant de faire le passage d'un vecteur de caractéristiques de bas-niveau à un vecteur dans l'espace latent, et la deuxième permettant de faire le passage d'un document au vecteur qui lui est associé dans l'espace latent. Etant donné une image requête, le vecteur latent est extrait et une distance est calculée entre chaque image de la base et l'image requête par calcul du cosinus entre les angles de leurs vecteurs latents respectifs. Les résultats obtenus sont encourageants malgré la simplicité des descripteurs d'images utilisés.
Dans [58], la démarche est tout à fait similaire, mais les documents non-annotés ne sont pas comparés aux document annotés dans l'espace latent. Les documents non-annotés sont comparés aux vecteurs correspondant aux mots d'annotations dans l'espace latent.
Pour avoir des annotations plus précises, il est préférable d'avoir des mots attachés à des régions de l'image et pas forcément à toute l'image. En effet, alors que certains concepts peuvent concerner l'ensemble de l'image ("extérieur", "intérieur","paysage"), certains concernent seulement une partie de l'image. Dans le cas d'annotation directement associées à des caractéristiques extraites sur l'ensemble de l'image, les densités conditionnelles doivent être apprises avec un nombre significatif de caractéristiques qui proviennent d'autres classes. De plus, l'utilisateur demandera parfois, dans le cas où les images sont très grandes (par exemple des images satellitaires), que le système lui renvoie uniquement des portions d'images, et non l'image entière. Dans le cas de méthodes faisant une modélisation directe entre le bas-niveau et le haut-niveau, on distingue deux approches : soit une segmentation est effectuée préalablement et une annotation est attachée à chaque région, soit on découpe arbitrairement l'image en "tuiles" plus petites auxquelles on attache une annotation.
Afin de pouvoir attacher des régions de taille non prédéfinie à des annotations, Mori et Takahashi présentent dans [92] une méthode de division progressive de l'image qui permet de faire un apprentissage de paires région/annotation à partir d'annotations attachées à l'image toute entière, ce qui évite une procédure d'annotation trop fastidieuse. À partir d'images d'apprentissage annotées, l'image

est divisée en blocs qui héritent chacun des annotations de l'image globale, ensuite une quantification vectorielles est effectuée sur les caractéristiques des sous-images. Puis, on compte les fréquences des mots sur chaque cluster et on calcule la probabilité de chaque mot étant donné un cluster. L'idée sous-jacente est de réduire de mauvaises corrélations en accumulant des régions similaires décrites par divers mots-clés. En effet, considérons une image contenant une montagne et une rivière. Après division en 2 régions, la région montagne aura 2 descriptions : "rivière et montagne" car elle hérite de toutes les annotations de l'image complète. Si, dans le lot d'apprentissage, une autre image contient de la montagne et du ciel. La zone de montagne contient 2 descriptions "ciel et montagne". Etant donné que les régions de montagne auront des caractéristiques similaires, elles seront probablement regroupées dans le même cluster et on favorisera la description par le mot clé "montagne" au détriment des mots clés "ciel" et "rivière". On peut ainsi espérer que le nombre de mauvaises descriptions va baisser au fur et à mesure.
Dans [7], le problème d'annotation automatique est vu comme un problème de traduction : à partir de données exprimées dans une certaine forme (des images, un texte écrit en français), on fait le lien avec des données dans une autre forme (des annotations, un texte écrit en anglais). En particulier, cette méthode nécessite de mettre en place un système passant d'un système de représentation à un autre. Typiquement, les lexiques sont appris à partir d'un type de données appelé "bitexte", c'est-à-dire un texte dans les deux langages où une correspondance grossière est connue, par exemple au niveau d'une phrase ou d'un paragraphe. Un ensemble d'images annotées est une forme de "bitexte" : nous avons une image segmentée en régions, et un ensemble de mots. Comme il est trop laborieux d'attacher chaque mot à une région, l'apprentissage est fait à partir d'images annotées où il n'est pas précisé quel mot est attaché à quelle région. L'algorithme EM est utilisé pour mener à bien un tel type d'apprentissage. Comme on souhaite faire le lien entre le vocabulaire de description de l'image (vecteurs de caractéristiques extraites en chaque région de l'image) et les annotations possibles, et que les caractéristiques extraites sur l'image ne sont pas discrètes, on lance un algorithme "k-means" sur les vecteurs de caractéristiques pour avoir un ensemble de m types de blobs possibles. L'algorithme permet ensuite de faire la correspondance entre les blobs et les annotations. Ensuite, étant donné une image non annotée, les blobs sont déterminés pour chaque région de l'image et ensuite le mot ayant la plus forte probabilité étant donné le blob est attaché à cette région.
Dans [64], l'image étant préalablement segmentée en régions distinctes, des blobs sont générés en utilisant un clustering effectué à partir des caractéristiques de bas-niveau extraites pour chaque région. A partir d'un lot d'images d'apprentissage annotées, des modèles probabilistes sont utilisés et permettent de générer un mot à partir des blobs d'une image non annotée. La qualité de l'annotation ainsi produite dépend beaucoup de la qualité du clustering et de la granularité qui est choisie : trop de clusters vont mener à un espace très clairsemé, tandis que trop peu de clusters vont mener à confondre des objets dans l'image.
Dans [14], les auteurs adaptent un modèle génératif hiérarchique proposé pour le texte par Hofmann [61] [62]. Ce modèle regroupe les documents dans des clusters

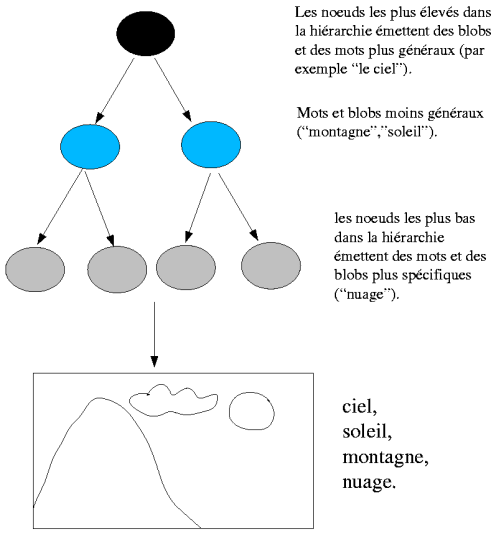


Fig. 3.1 – Illustration du processus génératif implicite au modèle statistique.

et modélise la distribution jointe des documents et des caractéristiques (modèle d'aspect). La génération des données est faite par une hiérarchie fixe de nœuds. Chaque nœud de la structure a une certaine probabilité de génération d'un mot, et a aussi une probabilité de génération d'une région de l'image avec certaines caractéristiques (voir figure 3.1). Ces probabilités sont fonctions du cluster correspondant au document, ce cluster pouvant être rapproché de la notion de variable latente mise en avant dans les modèles LDA et pLSA détaillés dans la section 3.1.2.1 de ce chapitre. Cette modélisation permet ainsi de prendre en compte une notion de généralité des concepts d'annotation. Des mots plus généraux et des descriptions d'images plus génériques se produiront à des nœuds élevés dans la hiérarchie. Le document est vu comme une séquence de mots et une séquence de régions. Le processus de génération du lot d'observations D associé au document d est décrit par la probabilité :

$$P(D|d) = \sum_c P(c) \prod_i \sum_l P(i|l,c)P(l|c,d))$$

.

où c désigne le cluster, i l'indice de l'objet discret (un mot ou une imagette), et l le niveau dans la hiérarchie. Le terme $P(l|c,d)$ est une fonction du document qu'il faut estimer face à un nouveau document. Le terme $P(i|l,c)$, dans le cas d'un mot, est simplement estimé par comptage des occurrences de ce mot au cours de l'apprentissage. Pour les caractéristiques des régions, une distribution gaussienne est utilisée, donnant des informations sur la taille, la position, la texture, la forme etc.
Dans [13], les auteurs utilisent trois méthodes hiérarchiques de génération de données annotées, adaptées de modèles traditionnellement utilisés pour générer des

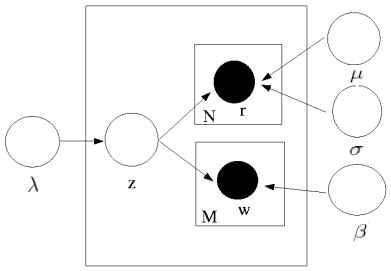


Fig. 3.2 – Modèle de mélange d'images GMM. Selon le standard graphique de représentation des modèles. Les nœuds représentent des variables aléatoires et les arcs les dépendances. Les nœuds en noir représentent des variables aléatoires observées, ceux en blanc des variables latentes. La distribution jointe peut-être obtenue en faisant le produit des distributions conditionnelles des nœuds étant donné leurs parents. Enfin, la carré entourant une variable aléatoire est une notation pour désigner la "réplication" : la boite autour de r désigne le fait que cette variable r est tirée N fois de suite pour les N mots annotant l'image.

documents textuels, tels que la LDA. Il s'agit donc de générer deux lots de données dont l'un est l'annotation de l'autre (les régions d'une image et leurs annotations, des articles et leur bibliographie, des gênes et leurs fonctions).
Le premier modèle est un modèle nommé "Gaussian multinomial mixture model" (voir figure 3.2). Etant donné un document, une seule variable latente discrète z est utilisée pour générer à la fois les N descripteurs de région r_n et les M mots d'annotation w_m. La génération des caractéristiques de chaque région sachant la variable latente z est modélisée par une gaussienne de paramètres σ et μ. La probabilité de génération d'un mot x_m sachant z est modélisée par une loi multinomiale de paramètre β. La distribution jointe du facteur caché, de l'image, et de l'annotation est exprimée de la manière suivante :

$$p(z, r, w) = p(z|\lambda) \prod_{n=1}^{N} p(r_n|z, \mu, \sigma) \prod_{m=1}^{M} p(w_m|z, \beta)$$

.

Conditionnellement au facteur latent z, les régions et les mots sont générés indépendamment et la correspondance entre les régions et les mots est ignorée. Le deuxième modèle, nommé "Gaussian multinomial LDA", essaie de pallier certaines insuffisances du premier modèle en générant les variables latentes au fur et à mesure pour un même document, de sorte que les différents mots et les différentes images peuvent provenir de variables latentes différentes. Le troisième modèle, nommé Corr-LDA, est beaucoup plus complexe et essaie de combiner la flexibilité de GMM-LDA et l'associativité de GM-mixture.

Model				
Human Annotation	sky jet plane smoke	bear polar snow tundra	water beach people sunset	buildings clothes shops street
Automatic Annotation	smoke clouds plane jet flight	polar tundra bear snow ice	sunset sun palm clouds sea	buildings street shops people skyline
Model				
Human Annotation	grass forest cat tiger	coral fish ocean reefs	mountain sky clouds tree	leaf flowers petals stems
Automatic Annotation	cat tiger plants leaf grass	reefs coral ocean fan fish	mountain valley sky clouds tree	petals leaf flowers lily stems
Model				
Human Annotation	sky jet plane smoke	sky clouds formation sunset	snow fox arctic	water boats waves
Automatic Annotation	plane jet smoke flight prop	sea sun sunset waves horizon	arctic snow polar fox ice	coast waves boats water oahu

Fig. 3.3 – résultats d'annotation d'images de la base Corel

3.3.2.2 Traitement de l'image comme une collection discrète

Primitives de l'image Dans les années 1960-1970, Julesz supposa les "textons" comme les éléments atomiques d'une perception visuelle des structures locales [68]. Dans des expériences de discrimination de textures, il trouva que le système de vision humaine détectait ces éléments de manière parallèle. Marr poursuivit les expériences de Julesz sur la notion de texton en les appelant "symbolic tokens" [86]. Un critère essentiel pour sélectionner ce dictionnaire de motifs de bas-niveau est de s'assurer qu'il comprend un vocabulaire suffisant pour représenter des images réelles, et que ces motifs ont une structure qui leur permet de se regrouper pour constituer des motifs plus complexes et de plus "haut-niveau". De nombreux travaux ont proposé des listes de motifs à partir d'une analyse statistique du signal de petites imagettes afin de traiter de grandes bases de données d'images [69].
Si, par analogie avec le langage, les textons sont les mots visuels, que sont les phrases visuelles ? Cette question est l'interrogation centrale de la théorie de la gestalt [146],[70]. On peut résumer ces travaux en disant que les relations géométriques d'alignement, de parallélisme, et de symétrie, sont les forces essentielles de groupement des parties de bas-niveau. Ces groupements peuvent s'effectuer à n'importe quelle échelle. Beaucoup correspondent à des groupements de 2 à 8 textons, mais les symétries et les parallélismes sont des groupements qui peuvent se manifester sur l'image toute entière. Les symétries, en particulier, se manifestent généralement à des échelles relativement grandes (un visage), et sont très aisément détectables par l'oeil humain.

Modélisation textuelle des mots visuels Dans [91], les auteurs étendent cette idée, qui consiste à appliquer à l'annotation 'images des méthodes qui ont

montré certains succès en fouille textuelle. L'image est segmentée grossièrement en 3 régions : le centre, la partie haute et la partie basse. Pour chaque région, un histogramme de couleurs est calculé et l'image est représentée par un vecteur concaténant l'histogramme de chaque région ainsi que les occurrences des termes de l'annotation, si l'image est annotée, ce qui est le cas des images d'apprentissage. A partir de cette description en vecteurs des images, les auteurs étudient deux approches introduites pour la recherche de documents textuels : la LSA et la PLSA.

Dans [123], les auteurs exposent une approche pour indexer des scènes cinématographiques en définissant un "vocabulaire visuel". Ce vocabulaire visuel est tout d'abord construit à partir d'une base d'apprentissage sur lesquelles sont extraites des "régions d'intérêt". Des descripteurs SIFT sont calculés sur ces régions de façon à avoir une caractérisation robuste et informative de ces régions. Ces vecteurs SIFT sont ensuite clusterisés par un "k-means" de façon à obtenir les "mots" du vocabulaire visuel. Ces mots ne contiennent pas de sémantique, les auteurs affirment cependant qu'ils constituent une description pertinente des scènes de film qu'ils souhaitent indexer. Etant donné une nouvelle scène, les régions d'intérêt sont extraites, les descripteurs sont extraits de chaque région et ensuite quantifiés en étant associés au codeword le plus proche. Ainsi, chaque scène est représentée par un histogramme des mots visuels. Ce "sac de mots" est ensuite normalisé par la pondération *tf-idf*. Pour comparer deux documents, le produit scalaire est utilisé comme mesure de ressemblance entre deux documents.

Un problème majeur soulevé par un grand nombre de ces travaux employant des techniques textuelles est la perte d'information préjudiciable entrainée par le fait de ne pas prendre en compte la configuration spatiale des mots visuels. Un certain nombre de travaux ont cherché à compenser ces problèmes.

Dans [121], le modèle pLSA est utilisé pour classifier des images non annotées. L'image est traité comme une collection discrète de mots visuels contenant des descripteurs de régions de type SIFT quantifiés, mais des *doublets*, mots codant des paires de mots visuels apparaissant dans un même voisinage de taille fixée. L'image est ensuite traitée comme un sac de mots, mais celui-ci contient, à travers ces *doublets*, des informations sur la répartition spatiale des mots visuels au sein de l'image. Dans [112], l'information sur la répartition spatiale des mots visuels est traitée en utilisant des segmentations multiples permettant de regrouper ces mots visuels. Etant donnée une image, l'algorithme Normalized Cut est appliqué en faisant varier deux paramètres de l'algorithme afin d'obtenir plusieurs segmentations de cette image. Chaque région de chaque segmentation est représentée comme un histogramme de mots visuels (des SIFT quantifiés) sur lesquelles sont appliquées les modélisation pLSA et LDA afin de détecter des objets et des scènes.

D'autres travaux tentent de résoudre ce problème en modifiant le modèle LDA pour prendre explicitement en compte la spatialité des mots visuels. Dans [140], le modèle SLDA (Spatial Latent Dirichlet Allocation) est proposé afin d'améliorer les performances obtenues par LDA pour la recherche d'objets dans les images. Dans ce modèle, la répartition spatiale des mots visuels dans le document est codée dans

une variable cachée du modèle. Le modèle génératif permet ainsi de grouper des mots visuels proches dans l'image au sein d'un même *document*. Dans le cas où un seul document est présent dans une image, SLDA se confond avec le modèle LDA. Dans [101], les auteurs proposent un autre modèle, nommé modèle GLDA (Geometric Latent Dirichlet Allocation), qui prend en compte la position des mots visuels. Le modèle GLDA ajoute au modèle LDA une variable latente comprenant la position d'un mot visuel. Une image est alors généré comme l'image des positions des mots visuels à travers une homographie. Dans [122], une hiérarchie est introduite selon l'apparence visuelle des objets en appliquant au domaine visuel le modèle hLDA (voir section 3.2.1.5). L'apprentissage du modèle hLDA sur un lot d'images non supervisées permet de faire apparaitre automatiquement une structure hiérarchique en groupant des objets selon un arbre à plusieurs couches.

3.3.2.3 Apprentissage interactif

Beaucoup de systèmes utilisent un retour de pertinence pour améliorer les résultats d'annotation, mais il ne constitue pas nécessairement le coeur de la méthode. C'est pourquoi nous ne détaillerons pas ici ce point. Nous citons cependant l'approche décrite dans [98] où les auteurs proposent une méthode d'*apprentissage interactif* pour relier les concepts subjectifs qui intéressent l'utilisateur aux valeurs symboliques calculées de façon non supervisée dans l'image. L'information extraite de l'image est organisée en 5 couches reliées l'une à l'autre par inférence bayésienne et représentant chacune un niveau d'abstraction différent (voir figure 3.4). Le niveau le plus bas correspond aux pixels de l'image (niveau 0). Des modèles stochastiques sont appliqués sur toutes sortes de caractéristiques (spectrales, texturales, etc.) et sont obtenus en estimant le maximum à posteriori du vecteur de paramètres $\theta_M = arg_\theta(max(p(\theta|D, M)))$. Ensuite, de nouvelles caractéristiques sont calculées en utilisant le maximum à posteriori sur un ensemble de modèles, pondérés par le facteur d'Occam qui agit comme une pénalité pour éviter des modèles excessivement compliqués. Ces "caractéristiques de caractéristiques" constituent le niveau 2 de la hiérarchie de description de l'image. À partir des caractéristiques du niveau 1 et des "méta caractéristiques" du niveau 2, un ensemble caractéristique de classes est cherché dans les espaces de paramètres des différents modèles, et doit refléter les structures existantes dans les différents espaces de caractéristiques. Ces classes sont obtenus par une quantification non supervisée, en utilisant une classification bayésienne, ou un algorithme *k-means*. Les niveaux 1 à 3 sont obtenus par une caractérisation complètement non-supervisée de l'image. À partir de cette description objective de l'image, il reste à définir les concepts qui intéressent l'utilisateur (niveau 4). Nous notons ces éléments subjectifs A_μ et les relions aux éléments objectifs ω_i en utilisant les probabilités $p(\omega_i|A_\mu)$. Au moment de la création du niveau 3, un vocabulaire de classes est créé pour chaque type de caractéristiques, étant donné que l'on ne sait pas quelles caractéristiques devront être combinées avec quelles autres. Ainsi, le vocabulaire total est décomposé en "sous-vocabulaire" :

$$\omega_{jk} = \omega_{sp,j}\omega_{tx,k}$$

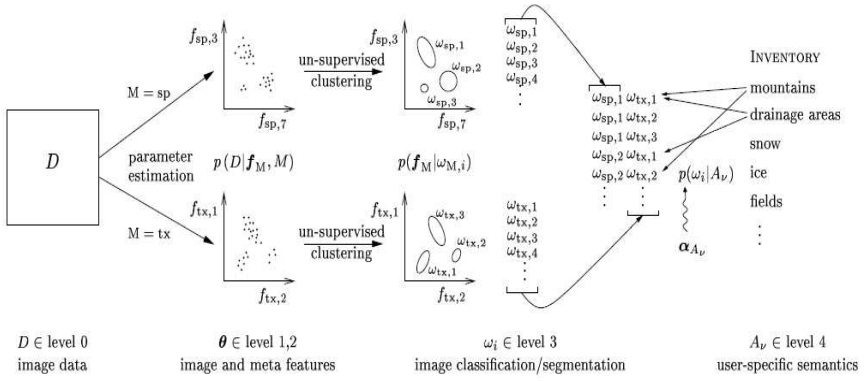


Fig. 3.4 – Schéma du système KIM

L'exemple donné ici utilise une combinaison de caractéristiques spectrales et de caractéristiques de texture mais on peut utiliser toutes sortes de modèles. Le système doit alors apprendre les vraisemblances $p(\omega_{jk}|A_\mu)$ à partir d'exemples fournis par l'utilisateur. L'indépendance conditionnelle est supposée entre les éléments du vocabulaire, nous avons donc l'expression suivante :

$$P(\omega_{jk}|A_\mu) = P(\omega_{sp,j}|A_\mu)P(\omega_{tx,k}|A_\mu)$$

Un mécanisme d'inférence bayésienne est utilisé pour apprendre ces probabilités. S'il y a r probabilités $P(\omega_i|A_\mu)$ à calculer, on suppose que l'on a un lot d'apprentissage T fournit par l'utilisateur $N = N_1, ..., N_r$ où N_i est le nombre d'occurrences de ω_i dans T. Etant donné que ω_i est une variable à r états, le vecteur N a une distribution multinomiale, le vecteur de paramètre $\theta = \theta_1, ..., \theta_r$ est ainsi introduit pour chaque concept A_μ pour avoir une représentation paramétrique : $P(\omega_i|A_\mu, \theta)$. Après observation du lot d'apprentissage, la probabilité à posteriori est :

$$P(\theta|T) = P(T|\theta)P(\theta)/P(T) = Dir(\theta|1 + N_1, ..., 1 + N_r)$$

Les paramètres peuvent ainsi être mis à jour au fur et à mesure en recalculant les hyper-paramètres $\alpha_i = 1 + N_i$ à partir d'images d'exemples et de contre exemples qui augmentent ou diminuent les valeurs des N_i.

3.3.3 Analyse syntaxique de l'image

Les travaux d'analyse syntaxique de l'image visent à extraire la sémantique en analysant les relations entre différentes primitives de l'image. En effet, beaucoup de motifs complexes sont composés d'un petit nombre de primitives liées par des relations simples. Ceci est totalement similaire au langage où un grand nombre de phrases complexes peuvent être générées à partir d'un vocabulaire limité et de règles de grammaire d'une manière hiérarchique : mot, syntagme et phrase.
Les premiers travaux d'analyse syntaxique de l'image apparurent dans les années 1970 avec les travaux d'Ohta et Kanade[99]. Mais on peut dire que ces travaux

étaient en avance sur leur temps et firent face rapidement à des difficultés qui étaient insurmontables pour l'époque :
- Une grande complexité de calcul : Les images réelles contiennent toujours un nombre important d'objets. Il s'agit de mettre au point un système qui peut traiter un nombre suffisant de catégories à détecter et qui peut coordonner les procédures "bottom-up" et "top-down".
- Le "fossé sémantique" entre les pixels et les motifs élémentaires à détecter. La nécessité de franchir cet écart entre les pixels et une description symbolique de l'image a motivé de nombreux travaux sur la reconnaissance d'apparence [94], les pyramides d'images [120], les ondelettes [36], et les méthodes d'apprentissage [113] [45], les méthodes de boosting [44].

Après un nombre important d'avancées dans ces domaines, des travaux d'analyse syntaxique d'images commencent à apparaître de nouveau dans la littérature [131] [51] [65] [139] [56] [24] [25] [132]. Ces travaux ont également bénéficié d'un certain nombre de progrès accomplis depuis les années 70, notamment un cadre mathématique et statistique efficace, comme les grammaires stochastiques [26]. Nous ne détaillons ici que les grammaires stochastiques sans contexte.

3.3.3.1 Grammaires stochastiques sans contexte.

Les grammaires stochastiques reprennent le cadre des grammaires formelles mais ajoutent un lot de probabilité P comme cinquième composante. Ainsi, étant donné $G = \{V_N, V_T, R, S, P\}$ une grammaire stochastique, et étant donné un symbole non-terminal A, un certain nombre de règles de réécriture γ_i sont possibles :

$$\gamma_i : A -> \beta_i$$

Chaque règle γ_i est associée à une probabilité $P(\gamma_i) = P(A -> \beta_i)$ telle que : $\sum_{i=1}^{n(A)} P(\gamma_i) = 1$, où $n(A)$ est le nombre de réécritures possibles pour A.
La probabilité de l'arbre syntaxique $\mathbf{pt}(\omega)$ ("parsing tree") s'écrit alors :

$$P(\mathbf{pt}(\omega)) = \prod_{j=1}^{n(\omega)} p(\gamma_j)$$

L'ensemble de toutes les phrases (pour le langage), ou configurations (pour l'image) qu'il est possible d'obtenir à partir d'une grammaire G est appelé langage et est noté $L(G)$.
La probabilité d'une phrase ou d'une configuration $\omega \in L(G)$ sécrit :

$$P(\omega) = \sum_{\mathbf{pt}(\omega)} p(\mathbf{pt}(\omega))$$

La grammaire stochastique est dite cohérente si la grammaire vérifie la condition suivante :

$$\sum_{\omega \in L(G)} P(\omega) = 1$$

3.3.3.2 Différences entre l'analyse syntaxique d'images et l'analyse textuelle.

Passer de grammaires de langages en une dimension à des grammaires pour l'image en deux dimensions n'est pas trivial. On peut relever trois difficultés fondamentales :

- La perte de l'ordre naturel gauche-droite du langage. Dans le langage, chaque règle de production $A- > \beta$ est supposée générer une séquence de nœuds ordonnés. Et, en appliquant ces règles jusqu'aux feuilles, une séquence de mots terminaux ordonnée linéairement est ainsi obtenue. En image, les liens implicites de voisin de gauche et de droite sont perdus et remplacés par des liens plus complexes de graphe d'adjacence de régions. Certaines idées pour faire face à la perte de l'organisation naturelle gauche-droite du langage ont été proposées par Fu sous les noms de "web grammar" et "plex grammar" [47], par Grenander [54], et plus récemment dans des graphes de grammaire pour l'interprétation de diagrammes [107]
- L'échelle d'un objet qui peut être quelconque. On ne peut pas lire une langue à différentes échelles, mais une grammaire d'image doit avoir une représentation en multi-résolution.
- L'irrégularité des motifs qui est plus grande que dans le langage. Les images peuvent comprendre des occlusions et des zones texturées dont les règles de production seront fortement stochastiques.

3.4 Application des réseaux sémantiques pour la fouille d'images satellitaires

Les réseaux sémantiques sont un outil très important pour la fouille d'images satellitaires. Les images satellitaires constituent en effet un monde "fermé" au sens où il est possible de déterminer un ensemble de concepts de taille raisonnable décrivant tous les objets et nommant toutes les différentes zones susceptibles d'être trouvées dans une base d'images. Ainsi, la définition d'une ontologie parait particulièrement importante pour clarifier les termes employés et les relations entre ces concepts.
Ainsi, dans [77], la base de donnée topographique ATKIS (Amtliches Topographisch Kartographisches InformationsSystem), créee par l'administration allemande, est modélisée par le réseau ERNEST (Elanger Semantisches Netzwerksystem [76]). Trois différents types de liens ("part-of", "specialization-of", et "concrete-on") sont présents dans ce réseau, structuré hiérarchiquement en différents niveaux. Le plus haut niveau de la hiérarchie contient 7 classes générales ("point fixe", "habitation", "trafic", "végétation", "eau", relief", et "zone"). Ces

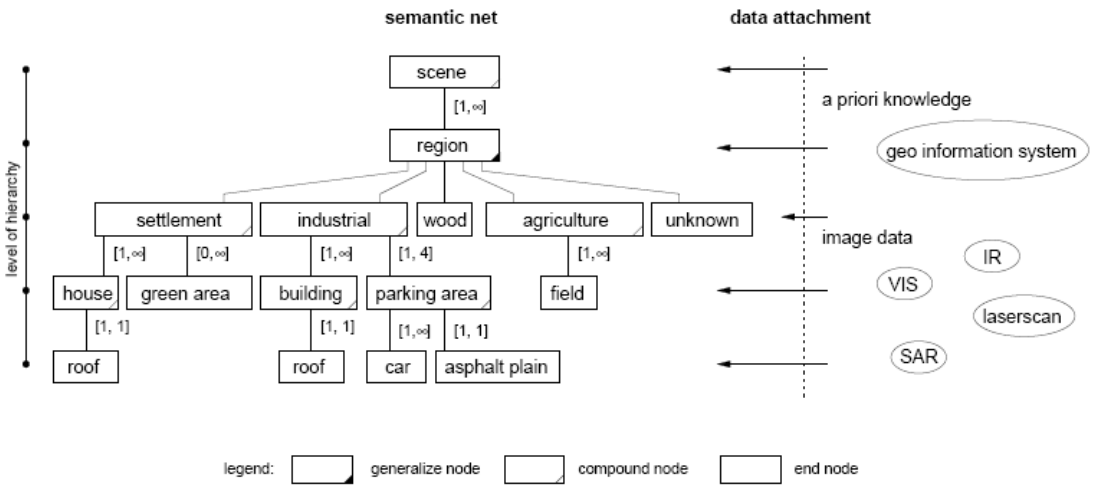


Fig. 3.5 – Réseau sémantique utilisé par le système GeoAIDA

classes sont subdivisées en sous-classes plus spécifiques, et les sous-classes sont divisés en objets ATKIS. Le réseau sémantique est utilisé dans un but de vérification de la classification de régions segmentées dans une image qui est effectuée à partir de caractéristiques de bas niveau extraites dans l'image.
Le système GeoAIDA [19], utilisé pour l'interprétation d'images satellitaires, utilise une représentation explicite de connaissances sur l'image apportées par des photo-interprètes à travers un réseau sémantique hiérarchique (voir figure 3.4). Ce réseau sémantique contient deux types de nœud : les nœuds de généralisation (relation sémantique d'hyponymie) et les nœuds de composition (relation sémantique de méronymie). Chaque nœud comporte une information sur le type de zone qu'il représente et possède des attributs qui font également partie des connaissances apportées par les photo-interprètes. La stratégie d'interprétation se base sur un certain nombre de règles fixées à priori.

3.5 Conclusion

Au cours de la dernière décennie, les techniques d'annotation sémantique d'images se sont concentrées sur des méthodes à variables latentes qui sont inspirées de techniques de fouilles de données utilisées pour les documents textuels. Des caractéristiques sont généralement extraites et discrétisées de manière à obtenir un vocabulaire discret contenant des éléments supposés être l'analogue des mots dans les textes. C'est pourquoi ces unités élémentaires sont souvent appelés *mots visuels*. Cependant, une limitation de ces méthodes statistiques d'annotation est l'absence de prise en compte des relations sémantiques entre les différents concepts. Or, la sémantique se base sur l'hypothèse que les concepts vivent dans un espace

qui leur est propre et qui est structuré. De plus, les concepts peuvent corréspondre à des zones très variables en terme d'échelle et de diversité de paysage qu'il convient de prendre en compte. Les méthodes d'annotation utilisant un réseau sémantique permettent une prise en compte des liens entre les différents concepts. Cependant, ces méthodes sont très rigides étant donné que la constitution d'un réseau sémantique condense l'information donnée par des photo-interprètes. Ainsi, l'ajout d'un nouveau concept nécessite une adaptation du réseau et l'intervention d'un expert.
Nous proposons dans cette thèse un modèle permettant d'exploiter la richesse sémantique d'un réseau sémantique, tout en conservant la flexibilité des techniques statistiques basées sur un apprentissage. L'introduction du modèle hiérarchique explicité et évalué dans le reste de ce rapport s'est faite en plusieurs étapes au cours du travail de thèse avant d'arriver à maturité sous sa forme finale. Citons au moins ici deux types de modèles qui ont été étudiés et mis en œuvre séparément. Une première modélisation de type 'part-of' a été mise en place où chaque concept d'annotation est associé à un modèle qui peut comprendre plusieurs couches en fonction de la complexité du type de région auquel correspond le concept. Chaque couche contient des "sous-modèles" correspondant à certains types de sous-régions. Le nombre de couches, le nombre de modèles de chaque couche et leurs paramètres sont déterminés à l'apprentissage par minimisation de la complexité stochastique. Ainsi, la région annotée par le concept d'intérêt est décomposée en sous-régions modélisées par les "sous-modèles" correspondant. Cependant, dans cette modélisation, les différents concepts n'ont pas de liens sémantiques qui les relient. De plus, les "sous-modèles" présents dans les couches intermédiaires des modèles associés à chaque concept ne contiennent pas de sémantique. La représentation sémantique dans cette modélisation n'est donc pas satisfaisante.
Ensuite, un autre type de modélisation de type 'kind-of' a été explicitée et mise en œuvre dans laquelle chaque concept correspond à un modèle de mélange sur un lot de distributions de probabilité. Le modèle de mélange permet de prendre en compte la diversité des paysages auxquels peut correspondre chaque concept et la complexité du mélange, à savoir le nombre de distributions, est fixé par minimisation de la complexité stochastique. Ainsi, un modèle général tel que "végétation" se verra attribuer plusieurs distributions associées à différents types de paysages. Cependant, ces différents types de paysages ne sont pas associés à des concepts sémantiques et les relations sémantiques ne sont ainsi que faiblement prises en compte. De plus, les expérimentations qui ont été effectuées amènent à constater une disparité importante entre les complexités des différents concepts utilisés pour la tâche d'annotation. En effet, si certains concepts peuvent être inferés efficacement par une modélisation directe des caractéristiques de bas-niveau, on remarque que pour d'autres concepts, le *fossé sémantique* qui les sépare du signal est trop important pour pouvoir être franchi en une seule étape. On en vient donc à introduire un "niveau de sémantique" qui diffère selon les concepts et qui doit être explicité et pris en compte pour une inférence efficace de la sémantique dans les bases d'images satellitaires.

Chapitre 4

Modélisation stochastique associée à un réseau sémantique

L'objectif de ce travail de thèse est de développer une méthode annotant des images satellitaires avec un vocabulaire dont les éléments appartiennent au langage naturel. Or, tous les linguistes s'accordent aujourd'hui sur le fait que les mots ne sont pas présents de façon isolée dans notre esprit et que le lexique possède une organisation, tout comme son contenu : les significations. Pourtant, on remarque que jusqu'à présent, cette organisation n'a pas pu être mise en évidence de manière complète et définitive par aucune méthode exhaustive. Cependant, lorsque le lexique concerne un domaine restreint à un domaine de connaissance bien délimité, les réseaux sémantiques permettent d'en capturer un certain niveau de structuration. La méthode ASP, dont le principe et la procédure d'apprentissage sont détaillés dans ce chapitre, consiste à prendre en compte l'organisation d'un lexique d'annotation d'images satellitaires, représentée par un réseau sémantique, en mettant ce réseau en relation avec un modèle stochastique. Les relations sémantiques prises en compte sont les relations d'hyperonymie/hyponymie, correspondant à une structuration du lexique dans le sens d'une généralité croissante/décroissante, de méronymie/holonymie, correspondant à une structuration du lexique dans le sens de termes annotant des zone de complexité croissante/décroissante, et de synonymie.

Le méthode ASP se positionne donc à la confluence entre les méthodes statistiques par apprentissage, souples mais traitant les labels d'annotation de manière indépendante, et les méthodes par règles utilisant des réseaux sémantiques, permettant une meilleure description de l'image mais souffrant d'une trop grande rigidité.

En effet, tant la structure que les paramètres du modèle stochastique utilisé par la méthode ASP sont estimés lors de la phase d'apprentissage. La structure du réseau sémantique intégrant le vocabulaire d'annotation en est alors déduit de manière déterministe. De plus, si un modèle et son réseau sémantique dual ont déja été appris sur une base d'images annotées, des labels supplémentaires (associé à des images exemples) peuvent Ã[a]tre intégrés automatiquement à la structure existante sans effectuer une procédure d'apprentissage complète sur la base d'images

annotées élargie. La méthode ASP, dont tous les paramètres sont estimés automatiquement, dispose donc d'une grande souplesse dans l'introduction des labels d'annotation et ne requiert pas l'intervention d'un expert. Par ailleurs, la plus-value apportée par l'utilisation de cette dualité entre un réseau sémantique et un modèle stochastique est double :

– Elle permet aux labels occupant des places élevées dans l'une ou l'autre des hiérarchies de mettre à profit le pouvoir de description acquis par les labels localisés en dessous d'eux. Ainsi, les modèles associés à des labels correspondant à des zones complexes peuvent-ils rester relativement simples à condition qu'ils puissent s'appuyer sur un ensemble suffisamment riche de labels de plus bas-niveau. Cela revient à dire que lorsque le *fossé sémantique* est trop important pour un label, son franchissement est décomposé en plusieurs étapes.
– Elle permet à une image de test d'être annotée selon plusieurs niveaux de complexité et de généralité. La description de l'image prend ainsi en compte la structure du lexique d'annotation. Ce point sera traité dans le chapitre 5, consacré au processus d'annotation.

4.1 Stratégie de franchissement du fossé sémantique adoptée

4.1.1 Précisions sur le vocabulaire employé

La méthode ASP se fonde sur une mise en correspondance entre un réseau sémantique et un modèle stochastique qu'on appelera modèle stochastique *global*. En effet, ce modèle stochastique global est composé d'un ensemble de modèles, attachés chacun à un label d'annotation. Chacun de ces modèles est composé d'un ensemble de parmètres et d'une place dans la structure du modèle global. Notons que le réseau sémantique et le modèle stochastique sont deux objets de nature différente (voir figure 4.1.1). Le réseau sémantique est en effet un graphe de représentation de la structure du lexique d'annotation. Il est composé de noeuds qui sont les labels d'annotation et d'arcs qui sont les relations sémantiques entre ces labels. Le modèle stochastique global est quant à lui un objet mathématique qui permet de calculer la probabilité jointe d'un label d'annotation et d'une image. Cette probabilité jointe ne peut pas toujours être calculée uniquement avec le modèle associé au label d'annotation car celui-ci peut dépendre d'autre modèles présents dans le modèle global. Par ailleurs, on appelera *modélisation* un certain type de modèle génératif tel que le modèle bayésien naïf, le modèle de mélange, etc.

4.1.2 Relations sémantiques et modélisations génératives

La sémantique lexicale considère principalement 4 types principaux de relations sémantiques (cf 2.2) : la synonymie, l'antonymie, l'hyponymie et la méronymie. Afin de mettre en relation un modèle stochastique global à un réseau sémantique, la méthode ASP attache une modélisation générative particulière à chaque relation

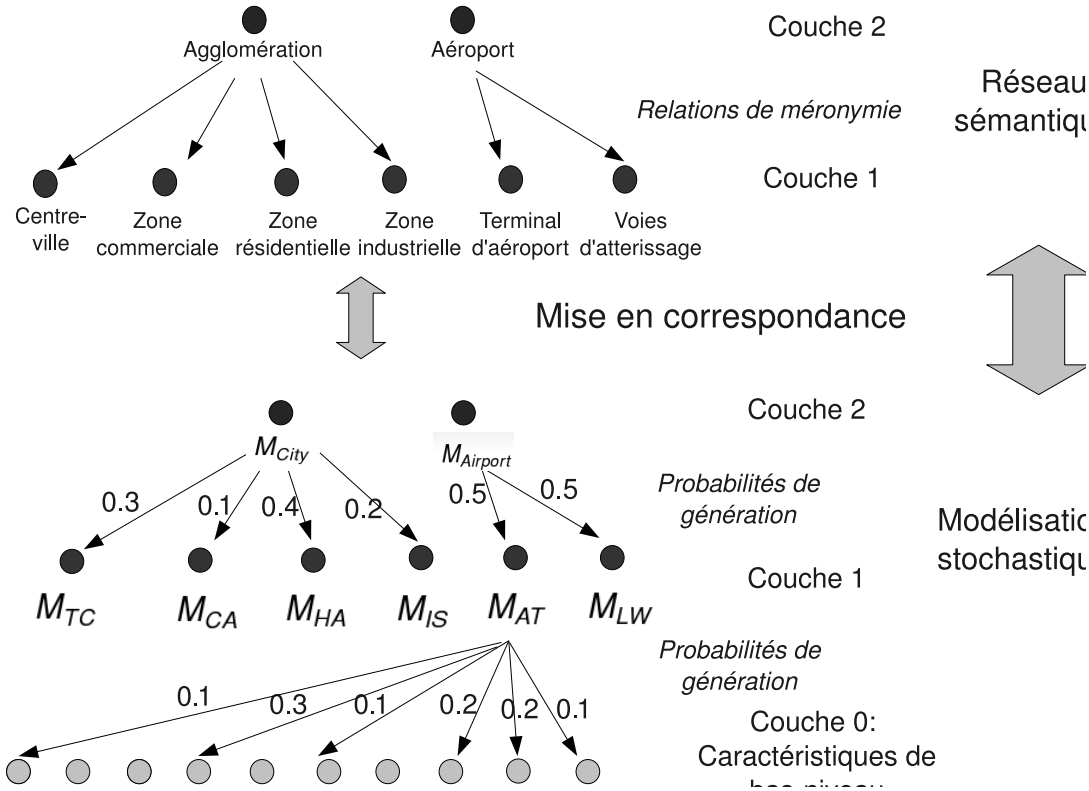

Agglomération
Aéroport
Couche 2
Relations de méronymie
Réseau sémantiqu
Couche 1
Centre-ville
Zone commerciale
Zone résidentielle
Zone industrielle
Terminal d'aéroport
Voies d'atterissage
Mise en correspondance
Couche 2
M_{City}
$M_{Airport}$
0.3
0.1
0.4
0.2
0.5
0.5
Probabilités de génération
Modélisatic stochastiqu
Couche 1
M_{TC}
M_{CA}
M_{HA}
M_{IS}
M_{AT}
M_{LW}
Probabilités de génération
0.1
0.3
0.1
0.2
0.2
0.1
Couche 0:
Caractéristiques de

Fig. 4.1 –

sémantique. Ainsi, la relation d'hyperonymie dans le réseau sémantique est associé à un modèle de mélange dans le modèle stochastique : la loi correspondant au label général s'exprime comme un mélange des lois correspondant aux labels plus spécifiques. La relation de méronymie, quant à elle, correspond à un modèle partitionnant une zone annotée par le label de haut-niveau en régions annotées par des labels de plus bas-niveau. La relation de synonymie entre deux labels correspond au fait que 2 labels sont associés au même modèle génératif.

L'antonymie est la seule relation sémantique lexicale qui n'a pas été prise en compte dans ce travail. Notons pourtant qu'une relation d'antonymie contradictoire du type "*urbain* est antonyme de *végétation*" pourrait avoir un sens pour relier des labels annotant des images satellitaires. Cette relation correspondrait alors à une relation de *OU exclusif* : "A est antonyme de B" implique que toute région d'une image est annotée par un et un seul de ces labels. Cependant, la méthode ASP a la prétention de pouvoir inférer automatiquement les relations sémantiques reliant les concepts à partir d'une base d'images annotées. Etant donné la diversité des types de zones qui peuvent être présentes dans des images satellitaires, déterminer automatiquement à partir d'images exemple qu'un type de zone A correspond à "tout sauf le type de zone B" nécessiterait un nombre d'images exemple prohibitif. Ainsi, la méthode ASP exposée dans ce travail se limite à la prise en compte des 3 autres types de relations.

En organisant les labels dans un réseau sémantique comportant une structure en couches, ceux-ci sont organisés en des niveaux distincts de généralité et de complexité. La méthode ASP peut alors adapter sa stratégie d'inférence en fonction de ces niveaux.

Un label correspondant à des zones au signal très homogène sera inféré directement à partir des caractéristiques de bas-niveau extraites dans l'image et occupera la couche la plus basse du réseau sémantique. Les régions annotées par des labels appartenant à cette couche, en contact direct avec les caractéristiques de bas-niveau, peuvent être rapprochées de ce qui est appelé dans certains travaux des *régions prototypes* [2]. Pour des labels correspondant à des zones plus générales ou plus structurées, l'inférence sera effectuée à partir des régions prototypes. Sous réserve de disposer d'un vocabulaire de labels suffisants, la méthode ASP décompose ainsi le franchissement du fossé sémantique en un certain nombre d'étapes.

4.2 Structure générale du système

Cette hiérarchie entre concepts exposée dans la partie précédente peut être représentée et formalisée de façon naturelle et efficace par des réseaux sémantiques hiérarchiques. Cette section présente la structure générale des réseaux sémantiques qui nous considérons ainsi que les deux relations sémantiques qui seront utilisés par la suite.

4.2.1 Réseaux sémantiques "kind-of" et "part-of"

Dans ce chapitre, on considère des réseaux sémantiques où les nœuds sont hiérarchisés en couches et où les concepts ne peuvent être reliés entre eux par un lien sémantique qu'entre couches successives. Chaque nœud d'une couche donnée doit être relié à au moins un nœud de la couche supérieure età au moins un nœud de la couche inférieure, cette dernière contrainte ne s'appliquant pas aux nœuds de la couche la plus basse qui ne peuvent être reliés qu'à des nœuds plus hauts dans la hiérarchie.
Nous construisons ici deux réseaux sémantiques distincts utilisant respectivement deux liens sémantiques qui sont à présent introduits.

Réseau sémantique avec lien "générique/spécifique" : Soit le lien sémantique $G(.,.)$, que l'on appelera lien "kind-of" (sorte de) tel que, si c est un concept d'une couche donnée et $\{c_1, ..., c_k\}$ un ensemble de concepts de la couche inférieure à celle de c :
– $G(c, \{c_1, ..., c_k\})$ signifie que : c_1, c_2,..., et c_k sont spécifiques de c (lien paradigmatique d'hyponymie), et réciproquement que c est générique de c_1, c_2,..., et c_k (lien paradigmatique d'hyperonymie).
On définit un réseau sémantique dont la structure est celle détaillée ci-dessus et où le seul lien sémantique existant est le lien G. Les concepts c et les concepts c_1, c_2,..., et c_k sont reliés si et seulement si la relation $G(c, \{c_1, ..., c_k\})$ est vraie. Nous définissons ici ce lien dans le réseau comme une relation "ou exclusif". En effet, nous supposons qu'une image est annotée par le concept c si et seulement si elle est annotée par le concept $c1$, ou par le concept $c2$, ... ou par le concept c_k. Un exemple d'un tel réseau est montré sur la figure 4.2.1. Ainsi, une image annotée par le concept "Végétation" est aussi annotée soit par le concept "Toundra", soit par le concept "Forêt". Les nœuds de ce type de réseau sont d'autant plus généraux qu'ils appartiennent à une couche élevée dans la hiérarchie.

Lien "partie/composé de" : Soit le lien sémantique $P(.,.)$, que l'on appelera lien "part-of" (partie de) tel que, si c est un concept d'une couche donnée et $\{c_1, ..., c_k\}$ un ensemble de concepts de la couche inférieure à celle de c :
– $P(c, \{c_1, ..., c_k\})$ signifie que : c_1, c_2,..., et c_k sont des parties de c (lien paradigmatique de méronymie), et réciproquement que c est composé de c_1, c_2,..., et c_k (lien paradigmatique d'holonymie).
Le lien de méronymie est par définition paradigmatique et induit une structuration verticale du lexique. Cependant, on peut voir aussi ce lien comme étant syntagmatique. Une relation syntagmatique consiste combine en effet des mots pour aboutir à une nouvelle signification généralement plus complexe. Ainsi, dans l'image, par analogie avec le texte, une relation syntagmatique regroupe plusieurs régions d'une image annotées par des concepts pour former une région correspondant à un concept plus abstrait et plus complexe. L'analogie d'une séquence de mots pour un ensemble de régions n'est cependant pas unique. Nous décidons ici qu'elle correspond à un ensemble de régions formant une zone

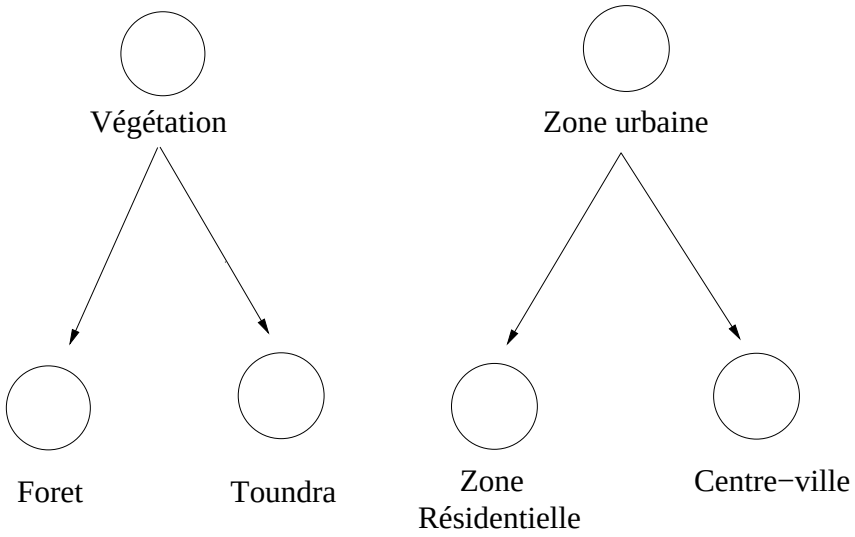


Fig. 4.2 – Exemple de réseau sémantique hiérarchisé par la relation *kind-of*. Une image est annotée par le concept "Toundra" ou "Forêt" sera aussi annotée par le concept "Végétation". Une image annotée par le concept "Zone urbaine" est soit annotée par le concept "Zone résidentielle", soit par le concept "Centre-ville".

4-connexe de l'image. On définit un réseau sémantique dont la structure est celle détaillée ci-dessus et où le seul lien sémantique existant est le lien P. Les concepts c et les concepts c_1, c_2,..., et c_k sont reliés si et seulement si la relation $P(c, \{c_1, ..., c_k\})$ est vraie. $P(c, \{c_1, ..., c_k\})$ signifie alors que, dans la base d'images que l'on considère, si une région R 4-connexe est annotée par le concept c, R est partitionnée en une ou plusieurs sous-régions elles-mêmes 4-connexes et annotées par des concepts appartenant à l'ensemble $\{c_1, ..., c_k\}$. Plusieurs sous-régions de R peuvent être annotées par un même concept. De même, tous les concepts $\{c_1, ..., c_k\}$ ne sont pas obligatoirement présents dans R.
Des concepts appartenant à des couches élevées dans la hiérarchie d'un tel réseau annotent ainsi des régions fortement structurées.

4.2.2 Dualité réseau sémantique/modélisation probabiliste

Notre approche est fondée sur la mise en correspondance entre un réseau sémantique, d'une part, et un modèle probabiliste, d'autre part, dont la structure est liée à celle du réseau sémantique, et qui permet d'exprimer la vraisemblance de la base de données d'images exemples. L'hypothèse de base que nous faisons dans cette section est que le meilleur réseau sémantique est celui qui correspond à la meilleure modélisation de la base de données. En effet, les capacités d'interprétation d'une scène par l'homme, à savoir la mise en correspondance de cette scène avec des concepts sémantiques, sont considérées comme excellentes et peuvent être prises comme modèle pour les systèmes informatiques d'interprétation d'image. Nous supposons donc qu'un réseau sémantique pertinent correspond à un bon modèle dans l'espace des modélisation possibles de la base de données. La

différence entre ce type de modélisation et les modèles graphiques traditionnels réside dans le fait que les liens existants entre les noeuds du réseau ne correspondent pas à des indépendances entre variables mais à des dépendences plus complexes entre les modèles pour exprimer la vraisemblance d'une image. Cependant, les structures possibles des réseaux sémantiques, même dans le domaine restreint que nous considérons, sont trop nombreuses. Ainsi nous nous limitons ici à deux types de relations sémantiques entre les concepts : "part-of" et "kind-of", et nous supposons une structure hiérarchisée en couches dans lesquels des liens sémantiques ne peuvent exister qu'entre 2 couches successives et selon des contraintes précises qui seront explicitées en 4.6.2. Nous proposons dans ce chapitre deux réseaux sémantiques correspondant à chacun de ces deux liens sémantiques, qui seront associés à deux modélisations probabilistes différentes du signal de l'image.

4.2.3 Couche de bas-niveau

La première étape de la méthode présentée ici est d'extraire des caractéristiques de bas-niveau dans l'image. Nous choisissons par la suite de quantifier ces caractéristiques pour travailler avec un vocabulaire discret, permettant une modélisation simple. Nous avons expérimenté deux types d'images pour lesquels nous avons choisi respectivement deux familles de descripteurs distincts :

- Des images SPOT5 à 2,5m de résolution : la texture semble être le meilleur outil pour décrire des images satellitaires à cette résolution.
- Des images Quickbird à 70cm de résolution : ce type d'images comporte beaucoup d'informations géométriques et les meilleures caractéristiques de bas-niveau pour décrire ce type d'images ne sont pas encore connues. Pour décrire ces images, nous avons choisi une méthode de classification bayésienne de patchs utilisant une modélisation probabiliste générative de textons présents dans ces patchs.

4.2.4 Description de bas-niveau d'images SPOT5 : caractéristiques de Haralick

4.2.4.1 Caractéristiques de Haralick

Les caractéristiques de Haralick se basent sur les matrices de co-occurrence [57]. L'idée principale de cette méthode est que toutes les informations de texture peuvent être exprimées par un ensemble de matrices (les matrices de cooccurrence) de dépendance spatiale des niveaux pour différents angles. Cependant, les matrices de cooccurrence sont très volumineuses, redondantes, et incertaines, car elles sont calculées à partir d'un nombre très faible d'occurrences. Elles sont donc représentées à partir d'un petit nombre de caractéristiques qui résument bien leur comportement (homogénéité, contraste etc.).
Des expérimentations ont été faites pour évaluer la performance de ces caractéristiques pour la classification d'imagettes de taille 64*64 extraites d'images

satellitaires SPOT5 à 2,5m de résolution sur 33 scènes différentes correspondant à des sites très variés. L'apprentissage à été fait sur une base d'images annotées manuellement pour classifier 7 classes différentes : champs, villes, montagnes, mer, forêt, nuages, neige. Les résultats de classification avec les coefficients de Haralick sont de l'ordre de 98%, ce qui peut être considérés comme très satisfaisant sur des images de cette résolution [22]. Les caractéristiques de Haralick ont également montré qu'elles étaient très robustes aux changements de contraste et de luminosité.

4.2.4.2 Clustering des caractéristiques de Haralick

Afin de déterminer la taille optimale du codebook à utiliser pour effectuer le clustering des caractéristiques de Haralick sur la base d'images SPOT5 que nous considérons, nous utilisons l'approche proposée dans [63]. Un modèle de mélange de gaussiennes est utilisé pour modéliser cet ensemble de caractéristiques et le critère de "minimum description length" est appliqué pour déterminer la complexité optimale du modèle, c'est-à-dire la taille optimale, notée ici N, du codebook. Les codewords ayant été calculés, toutes les caractéristiques du corpus d'images sont quantifiées. Nous avons donc un nouveau lot d'images dont les valeurs de pixels sont les indices associés à chaque vecteur de caractéristiques, ces valeurs sont donc comprises dans l'ensemble $\{1, ..., N\}$ (voir figure 4.3).

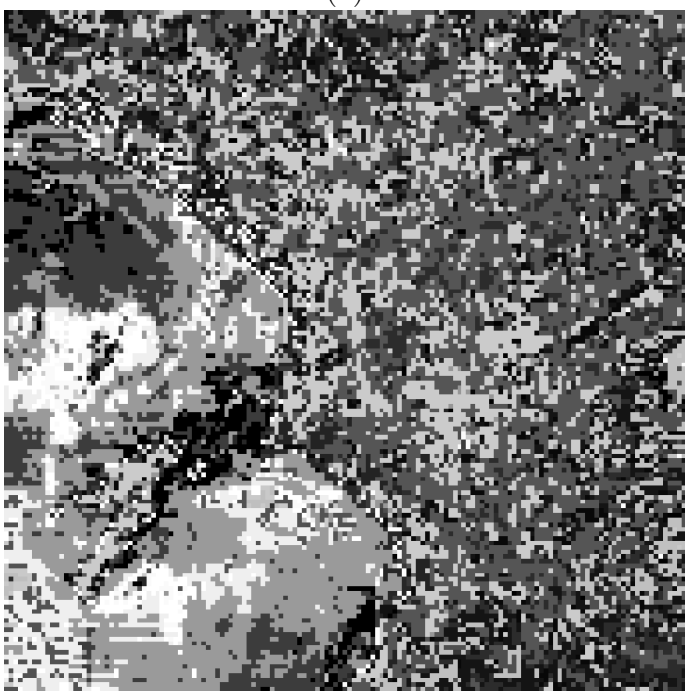
4.2.5 Description de bas-niveau d'images Quickbird

La méthode d'extraction de caractéristiques de bas-niveau utilisée pour décrire notre base d'image Quickbird est détaillée en annexe A. Elle est fondée sur un modèle probabiliste génératif de "mots-visuels" qui sont ici des descripteurs SIFT quantifiés. Une classification bayésienne par maximum de vraisemblance est ensuite appliquée.

4.2.6 Notations et formalisme employé

De façon formelle, soit un ensemble de concepts $\Omega = \{\omega_1, ..., \omega_n\}$ où chaque concept ω_i est lié à un lot d'images exemples X_i fournies par l'utilisateur. La base de données globale est notée $X = \{X_1, ..., X_n\}$, et on note X_{ij} la j-ème image d'apprentissage du lot d'image associé au concept i. Soit S_Ω un réseau sémantique construit à partir de Ω. S_Ω est mis en relation avec un modèle probabiliste général M_Ω permettant de calculer la vraisemblance $P(X|M_\Omega)$ de la base de données X sachant M_Ω. La structure de M_Ω correspond à la structure de RS_Ω d'une manière qui sera détaillée dans les sections suivantes. Nous considérons ici que chaque concept est associé à un modèle stochastique M_{ω_i}, celui-ci contenant une place dans la structure de M_Ω ainsi qu'un lot de paramètres θ_i. On note $C_1(M_\Omega)$ l'ensemble des modèles placés en première couche, $C_2(M_\Omega)$ l'ensemble des modèles disposés dans la seconde couche.

(a)



(b)

Fig. 4.3 – Résultat de quantification des descripteurs de Haralick. L'image (a) est une image 6000×6000 de Marseille. Les caractéristiques de Haralick ont été calculées sur une fenêtre glissante 64*64 avec un pas de 40 textons. L'image (b) de taille 148×148 correspond à la quantification de ces caractéristiques : la valeur de chaque texton est l'indice du cluster dans lequel a été classé le vecteur correspondant à cette fenêtre. ©CNES

L'ensemble des réseaux sémantiques S_Ω est mis, d'une manière qui sera détaillée dans la section suivante, en bijection avec l'ensemble des modélisations M_Ω. Ainsi, en déterminant la modélisation *optimale* de la base de données X, on détermine l'unique réseau sémantique qui est mis en relation avec M_Ω. Le critère que nous utilisons pour déterminer le meilleur modèle parmi l'ensemble des modèles possibles est le critère de minimisation de la complexité stochastique. Nous supposons donc que le meilleur réseau sémantique est celui qui correspond au modèle probabiliste codant la base de données X avec un nombre de bits minimal.

On définit une région dans une image comme un sous-ensemble 4-connexe de textons de cette image. On suppose de plus, conformément au travail effectué jusqu'ici, que des caractéristiques de bas-niveau discrètes pouvant prendre n_0 valeurs ont été préalablement extraites dans l'image suivant une grille régulière. Ces caractéristiques de bas-niveau étant discrètisées et étant extraites selon une grille régulière, elles forment les textons d'une nouvelle image sur lesquelles s'appliquent les algorithmes d'inférence de sémantique mis en œuvre dans ce travail. On appelera textons ces caractéristiques de bas-niveau. Les modélisations statistiques d'une région utilisées par la méthode ASP faisant l'hypothèse de l'interchangeabilité des textons au sein de la région (car inspirées des méthodes à variable latente utilisées pour les documents textuels 3.1.2.1) ne s'intéresse ici qu'à l'histogramme de ces textons dans une région donnée, qu'on notera pour une région quelconque $x = (x_1, ..., x_{n_0})$.

4.3 Relation de type "kind-of"

On considère ici des réseaux sémantiques hiérarchisés en couches où le seul lien sémantique existant est le lien sémantique "kind-of" liant deux concepts entre deux couches successives. La structure employée pour ces réseaux sémantiques est celle explicitée en 1.2 en limitant pour simplifier les notations à 2 le nombre maximal de couches possibles. Nous n'imposons pas de contrainte sur le nombre de nœuds de la couche 2 auquel peut-être relié un nœud de la couche 1. Par contre tout nœud de la couche 2 doit être relié à au moins un nœud de la couche 1.

Pour alléger les notations, on écrira M_i pour M_{ω_i}, modèle associé au concept i. Le nombre de concepts de la couche i sera noté n_i. Par définition, la couche 0 contient les textons. Les seules régions qui sont considérées ici sont celles définies par l'ensemble des textons de l'image. Chaque image X_{ij} de l'ensemble d'apprentissage du concept i est vue comme une région représentée par l'histogramme des valeurs des textons qui y sont présents.

4.3.1 Modélisation associée à la relation de type "kind-of"

Dans cette section, la mise en parallèle qui est faite entre le lien sémantique de généralisation/spécification et le modèle statistique de loi de mélange est détaillée.

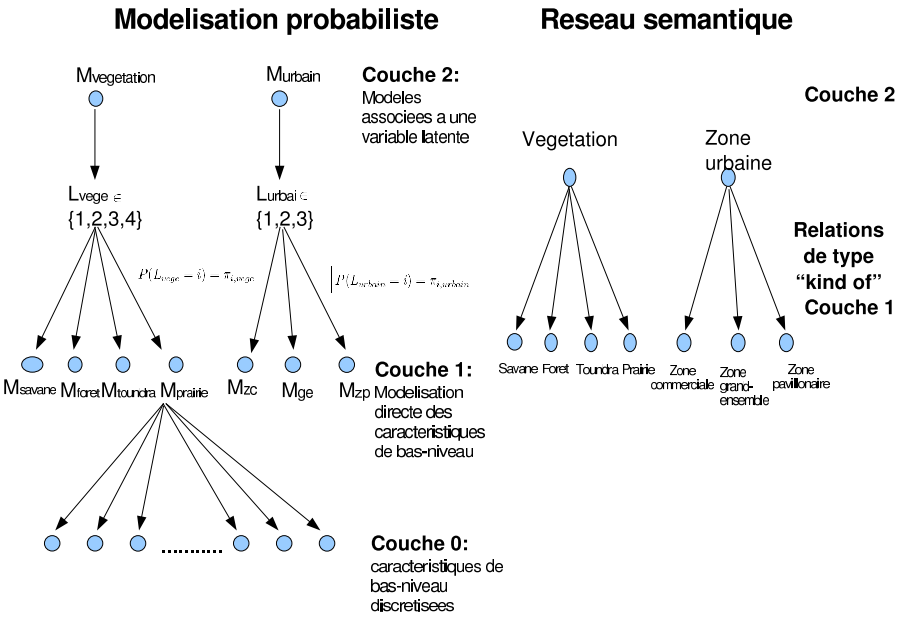


Fig. 4.4 – Dualité modélisation probabiliste/Réseau sémantique avec relations de type "kind-of"

4.3.1.1 Nœuds de la première couche

Les concepts de la couche 1 sont associés à un modélisation directe sur les caractéristiques de bas-niveau. Ainsi, si c est un concept de la première couche, la vraisemblance d'une image x annotée par le concept c ne dépend que du modèle M_c :

$$P(x|M) = P(x|M_c) \tag{4.1}$$

Cette vraisemblance est exprimée selon un modèle de type bayésien naif ou le nombre de textons dans l'image est codé par une loi de Poisson, et où chaque texton est tiré indépendemment avec probabilité θ_{cj} si sa valeur est j. Ainsi, la vraisemblance de l'image x sachant M_c s'écrit :

$$P(x|M_c) = Poiss_{\lambda_c}(\sum_{j=1}^{n_0} x_j) \prod_{j=1}^{n_0} \theta_{cj}^{x_j} \tag{4.2}$$

4.3.1.2 Nœuds de la deuxième couche

Le fait que des concepts $c_1, ..., c_k$ sont reliés à un concept c par la relation "kind-of" signifie d'un point de vue sémantique que les concepts $c_1, ..., c_k$ constituent des spécifications du concept c. Nous modélisons cette structure sémantique par le fait que la loi de probabilité de c est un mélange des probabilités P_{c_i} :

$$P(x|M_c) = \sum_{i=1}^{k} \pi_i P(x|M_i) \tag{4.3}$$

.
où x est une image de la base de données représentée par son histogramme $x = (x_1, x_2, ..., x_{n_0})$ des valeurs des textons, et où $\forall i \in \{1, ..., k\}, \pi_i \in [0, 1]$ et $\sum_{i=1}^{k} \pi_i = 1$. Le fait que chaque concept de la couche 2 soit relié à au moins un nœud de la couche 1 garantit qu'une fonction de probabilité soit bien définie pour chacun de ces concepts.
Cette modélisation revient à dire que c est associé à une variable latente L_c qui peut prendre ses valeurs dans $\{1, .., k\}$ avec probabilité $P(L_c = i) = \pi_i$ et que la probabilité de génération du vecteur x est exprimée avec le vecteur de paramètre θ_i si $L_c = i$.

4.3.1.3 Expression de la probabilité globale

Les différents lots d'apprentissage X_i sont supposés être indépendants conditionnellement au modèle global M. on écrit ainsi :

$$P(X_1, X_2, ..., X_n|M) = \prod_{i=1}^{n} P(X_i|M) \tag{4.4}$$

.
Comme vu précédemment, si $M_c \in C1(M)$, $P(X_c|M) = P(X_c|M_c)$. On écrit ainsi l'équation 4.4 de la manière suivante :

$$P(X_1, X_2, ..., X_n|M) = \prod_{M_i \in C1(M)} P(X_c|M_c) \prod_{M_i \in C2(M)} P(X_c|M) \tag{4.5}$$

.

4.3.1.4 Propriété d'extensivité du modèle

Propriété d'extensivité Soit M un modèle dont les paramètres et la structure ont été optimisés par maximum de vraisemblance de la base d'apprentissage $X = \{X_1, ..., X_n\}$ parmi l'ensemble **M** des modèles possibles :

$$M = \arg\max_{\boldsymbol{M}} P(X|M)$$

.

Supposons qu'on ajoute à présent un nouveau concept c_{n+1}, et soit X' la base d'apprentissage constituée par l'ajout d'un lot d'apprentissage X_{n+1} associé au concept c_{n+1}.

$$X' = X \cup X_{n+1}$$

.

Soit alors M_2 le modèle estimé par maximum de vraisemblance de la base d'apprentissage X' parmi l'ensemble des modèles **M'** :

$$M' = \arg\max_{M' \in \mathbf{M'}} P(X'|M')$$

.

Nous voulons à présent démontrer la propriété suivante :

$$P(X|M') \geq P(X|M)$$

.
Nous appelons ici cette propriété "extensivité" par emprunt à la thermodynamique. Un système thermodynamique étant défini, une variable d'état est dite extensive si elle croit linéairement avec la taille du système [6] (Comme le volume, le nombre de particules). Ici, il n'y a pas de propriété de linéarité de la vraisemblance par rapport au nombre de concepts présents dans le modèle global, mais la vraisemblance ne peut qu'augmenter par l'ajout d'un nouveau concept, d'où cette analogie.

Preuve : Soit $i \in \{1, ..., n\}$, $\mathcal{M}_i$ et $\mathcal{M}'_i$ les deux ensembles de modèles décrivant le lot d'apprentissage X_i pour chacun des deux cas de figure. Le réseau étant contraint à un nombre maximum de couches, les modèles M_i et M'_i peuvent se situer en première ou en deuxième couche du réseau.
Les modèles de la couche 1 correspondants étant définis par une modélisation directe sur les textons de l'image et leur expression n'est pas reliée aux autres modèles du réseau (voir équation 4.1). On a donc l'égalité suivante :

$$C1(\mathcal{M}_\rangle) = C1(\mathcal{M}'_\rangle)$$

.

En revanche, en ce qui concerne la deuxième couche, on peut écrire :

$$\{C2(\mathcal{M}'_\rangle)/\pi_{i,n+1} = 0\} = C2(\mathcal{M}_i) \tag{4.6}$$

.
Ainsi, on a la relation d'inclusion :

$$\forall i, \mathcal{M}_\rangle \supset \mathcal{M}'_\rangle$$

.

Donc

$$\mathcal{M}' \supset \mathcal{M}$$

. On en déduit la propriété à démontrer :

$$\max_{\mathcal{M}'} P(X|M') \geq \max_{\mathcal{M}} P(X|M)$$

.

Commentaires Cette propriété est très intuitive. En effet, les modèles de la deuxième couche s'expriment à partir des modèles de la couche 1 suivant un modèle de mélange. Le fait d'ajouter un nouveau modèle qui est susceptible d'occuper la couche 1 donne un "degré de liberté" suppémentaire qui ne peut ainsi que faire augmenter la vraisemblance globale.
Ainsi, plus un nombre important de concepts est défini, meilleure est la description de la base de données globale. Contrairement à la modélisation explicitée au chapitre 3, le système tire ici profit de la connaissance fournie par l'utilisateur lorsque il annote un groupe d'images par un concept.

4.3.2 Codage des différents modèles

Le principe de minimisation de la complexité stochastique a été introduit par Rissanen en 1978 [109]. Basé sur des concepts issus à la fois de l'estimation statistique et de la théorie de l'information, il apparait, à un niveau intuitif, relativement naturel.

Couche de niveau 1 Les modèles de la couche 1 sont supposés générer directement les textons des images exemples qui leur sont associés. Par conséquent : $C(X_c|M) = C(X_c|M_c)$. Ainsi, si X_{cj} est la j-ème image d'apprentissage du concept c, on utilise la formule proposée par Shannon [117] liant directement la longueur de codage $CS(X_c|M_c)$ à sa probabilité d'apparition :

$$CS(X_c|M_c) = -\log P(X_c|M_c)$$

.
où la loi P est la probabilité définie par l'équation 4.2. Les N_c images X_{cj} de la base d'images fournies pour le concept c étant supposées indépendantes, l'équation précédente s'écrit :

$$CS(X_c|M_c) = -\sum_{j=1}^{N_c} logP(X_{cj}|M_c)$$

.

Soit, en introduisant la formule 4.2 dans l'équation précédente :

$$CS(X_c|M_c) = -\sum_{j=1}^{N_c} log[Poiss_{\lambda_c}(\sum_{k=1}^{n_0} x_{cjk}) \prod_{k=1}^{n_0} \theta_{ck}^{x_{cjk}})] \tag{4.7}$$

.

En développant cette expression, on obtient donc :

$$CS(X_c|M_c) = -\sum_{j=1}^{N_c}\sum_{k=1}^{n_0} x_{cjk} log\lambda_{ck} + \lambda_c - N_c log\lambda_c + \sum_{j=1}^{N_c} logj \tag{4.8}$$

.

Cette dernière quantité est la longueur de code nécessaire pour coder un par un les textons de l'image.
Pour coder le modèle M_c, il faut tout d'abord coder le numéro de la couche auquel il appartient. Ici, nous ne considérons qu'un modèle possédant au maximum deux couches, nous codons donc le numéro de couche par un bit à valeur 0 ou 1. Le seul lien existant dans ce modèle étant la relation *Kind-of*, il n'est pas nécessaire de coder la nature des liens le reliant aux concepts de la couche 1. Il reste ainsi à coder les paramètres de génération des textons θ_c, et le paramètre de taille Λ_c. Pour cela nous utilisons la formule introduite par Rissanen [109] qui attribue au codage d'un vecteur de paramètres de taille T estimé avec N_{ech} la longueur de codage :

$$\frac{T}{2}\log N_{ech} \tag{4.9}$$

.

Le vecteur θ_c est de taille n_0, et le nombre d'échantillons avec lequel il est estimé est égal au nombre total de textons de la base X_c. Le vecteur Λ_c est de taille 1 et le nombre d'échantillons avec lequel il est estimé est égal au nombre total d'images de la base, soit par définition N_c.

$$CS(M_c) = \frac{n_0}{2}\log(\sum_{j=1}^{N_c}\sum_{k=1}^{n_0} x_{cjk}) + \frac{1}{2}\log(N_1) \tag{4.10}$$

.

Ainsi, la complexité globale $CS(X_c, M_c)$ s'écrit :

$$CS(X_c, M_c) = -\sum_{j=1}^{N_i}\sum_{k=1}^{n_0} x_{cjk}\log\theta_{ck} + \frac{n_0}{2}\log(\sum_{j}^{N_c}\sum_{k}^{n_0} x_{cjk}) + \frac{1}{2}\log(N_1) + 1 \tag{4.11}$$

.

Couche de niveau supérieur à 2 c étant un concept d'une couche supérieure à 2, le terme $C(X_c|M)$ s'écrit :

$$C(X_c|M) = -\log P(X_c|M)$$

.

Soit, en utilisant l'hypothèse d'indépendance des images de la base X_c et en introduisant l'expression de $P(X_c|M_c)$ écrite en 4.3 dans la dernière équation, on obtient l'expression :

$$CS(X_c|M) = -\sum_{j=1}^{N_c}\log(\sum_{i=1}^{k}\pi_i Poiss_{\lambda_c}(\sum_{k=1}^{n_0} x_k)\prod_{k=1}^{n_0}\theta_c^{x_k})$$

.

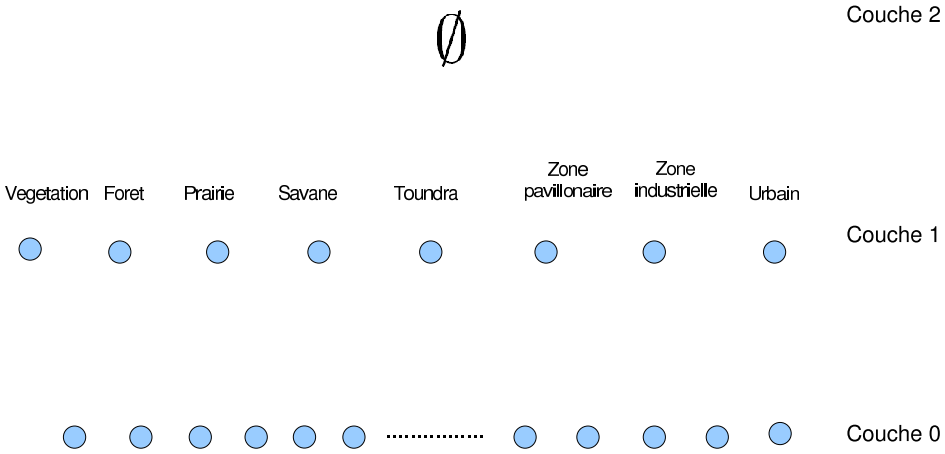


Fig. 4.5 – Initialisation de l'algorithme d'optimisation : tous les concepts sont mis dans la couche 1.

Il est nécessaire de coder les paramètres de génération π_c des concepts de la couche 1. Pour chaque modèle, on code ainsi le numéro de la couche auquel il appartient et le vecteur de paramètres π_c de la loi de la probabilité de la variable latente. π_{cj} n'est non nul que pour les concepts de la couche inférieure reliés par un lien sémantique de type "kind-of" au concept c. π_c a un nombre de paramètres égal au nombre de nœuds de la couche précédente, et qui est estimé avec le nombre d'images de la base, soit par définition N_c. En utilisant la formule 4.9 et en supposant que c appartient à la couche 2, on obtient pour $CS(M_c)$ l'expression :

$$CS(M_c) = \frac{n_1}{2} \log(N_c) \tag{4.12}$$

.

4.3.3 Algorithme d'optimisation utilisé

Dans un cas où le nombre de couche du modèle est 2 et où l'on a n concepts, le nombre de dispositions possibles des nœuds au sein des deux couches est de 2^n, ce qui fait un nombre de configurations trop important pour qu'elles puissent être explorées intégralement. Nous proposons ici un algorithme glouton itératif qui choisit à chaque étape la structure minimisant localement la complexité stochastique $CS(X, M)$.

État initial La configuration de départ de l'algorithme est celle pour laquelle tous les modèles sont tous situés sur la couche 1 et où la couche 2 est donc vide.

Évolution À chaque étape, et pour chaque concept c de la couche 1, on calcule la complexité stochastique associée à la configuration dans laquelle le concept c est mis dans la couche 2. Les paramètres de modèles de la première couche sont tout d'abord estimés en utilisant le maximum de vraisemblance.

On a donc, pour tout modèle c de la première couche les formules suivantes :

$$\forall j \in \{1, ..., n_0\}, \theta_{cj} = \frac{occ_{X_c}(j)}{card(X_c)} \tag{4.13}$$

.

$$\lambda_c = \frac{1}{N_c} \sum_{j=1}^{N_c} card(X_{cj}) \tag{4.14}$$

.

où $card(X_{cj})$ est le nombre de textons dans l'image X_{cj},
$card(X_c) = \sum_{j=1}^{N_c} card(X_{cj})$. $occ_{X_c}(L_c = j)$ est le nombre d'occurrences du texton de valeur j dans la base X_c.
On a donc, pour tout modèle c de la deuxième couche les formules suivantes :

$$\forall j \in \{1, ..., n_1\}, \pi_{cj} = \frac{\sum_{i=1}^{N_c} P_{c_j}(x_i)}{\sum_{i=1}^{N_c} \sum_{k=1}^{n} P_{c_k}(x_i)} \tag{4.15}$$

.

Les paramètres du modèle étant ainsi estimés, la complexité stochastique globale est alors calculée avec la formule 3.3. Le modèle qui minimise la complexité est retenu et est mis dans la couche 1 si la complexité correspondante est inférieure à la complexité obtenue lors du calcul de l'étape précédente.

Condition d'arrêt de l'algorithme L'algorithme s'arrête lorsque la complexité stochastique augmente. Le dernier modèle ainsi obtenu est pris comme le modèle optimal. On obtient ainsi le réseau sémantique résultat en gardant la disposition des concepts obtenue dans les différentes couches et en créant un lien de type "kind-of" entre un concept a de la couche 2 et un concept b de la couche 1 si la probabilité que la variable latente L_a prenne la valeur b soit supérieure à un seuil fixé arbitrairement :

$$P(L_a = b) > s$$

Dans tout ce travail, le seuil s égal à 0.

4.3.4 Analyse de l'algorithme d'optimisation

À chaque itération, la complexité globale du modèle est calculée et un concept passe de la couche 1 à la couche 2 si cette complexité est plus faible que la complexité calculée à l'itération précédente. Ainsi, l'algorithme comporte au plus N itérations. De plus, à l'étape k, la couche 1 comporte $N - k$ concepts et $N - k$ configurations sont donc à étudier. L'algorithme est donc de complexité N^2.
Notons par ailleurs que rien n'empêche le fait que l'algorithme s'arrête dès la première itération, car la couche 2 peut être vide.

4.3.5 Remarque

Dans le cas d'un où l'on ne souhaite pas construire automatiquement la structure du réseau (car un réseau sémantique est déja à disposition), il est juste nécessaire d'estimer les paramètres des modèles. Ainsi, il est uniquement nécessaire d'appliquer les formules 4.13 et 4.14 pour les modèles de la couche 1, et 4.15 pour les modèles de la couche 2.

Discussion sur la condition d'arrêt La condition d'arrêt de l'algorithme d'optimisation est que la CS remonte. Prouvons que si la CS augmente à une étape e_1 donnée de l'algorithme, il n'est pas possible qu'elle redescende lors d'une étape $e_2 \geq e_1$ lors de l'ajout d'un concept c dans la couche 2. En effet, soit M_{e_1-1} le modèle à l'état $e_1 - 1$, qui est la dernière étape avant que la CS augmente. Par hypothèse, séléctionner le concept c pour le faire passer dans la deuxième couche à l'étape e_2 fait diminuer la CS. Comparons la variation de CS entraînée par le fait de sélectionner le concept c à l'étape e_1 par rapport au fait de le séléctionner à l'étape e_2.
Analysons le fait que faire passer un modèle M_c de la couche 1 à la couche 2 a deux impacts : un impact sur le terme de complexité $CS(X_c|M_c)$ qui va s'en trouver modifié, et un impact sur les termes de complexité des modèles déja présents dans la couche 2. Remarquons que ce dernier est nécessairement négatif : le fait d'avoir moins de valeurs possibles pour la variable latente associée à chaque modèle de la couche 2 ne peut en effet faire qu'augmenter la complexité.
Or, à l'étape e_2, la couche C_1 contient $e_2 - e_1$ concepts en moins. La diminution du terme $CS(X_c|M)$ entrainée par le passage du concept c sur la couche 2 sera donc nécessairement moins important en e_2 qu'en e_1 car le nombre de modèles à disposition pour exprimer le modèle de mélange sera plus faible. De plus, étant donné qu'il y a un plus grand nombre de concepts sur la couche 2 à l'étape e_2 qu'en e_1, l'augmentation induite par le passage du concept c à l'étape e_2 sera plus importante qu'en e_1 car le nombre de modèles de mélange impliquant le modèle c est plus important.
Ainsi, si Δ_c^e est la variation de CS entrainée par le passage du concept c dans la couche 2 à l'étape e de l'algorithme. Il vient d'être démontré que

$$\Delta_c^{e_1} < \Delta_c^{e_2}$$

Par hypothèse, on a

$$\Delta_c^{e_2} < 0$$

.
On en déduit d'après ces deux inégalités que :

$$\Delta_c^{e_1} < 0$$

.

Ceci entre en contradiction avec le fait que la CS augmente à l'étape e_1 au cours de l'algorithme. Car séléctionner le concept c aurait fait diminuer la complexité.

Discussion sur l'heuristique Cependant, nous essayons ici de justifier l'heuristique sur laquelle il repose. Cette justification n'a cependant pas valeur de démonstration mathématique.
Remarquons premièrement que la complexité $CS(M)$ augmente de façon logarithmique avec le nombre de textons dans la base X (voir équations 4.9 et 4.28), tandis que la complexité associée à l'attache aux données $CS(X|M)$ augmente linéairement (voir équation 4.21). Ainsi, nous nous plaçons ici dans un cas où l'on suppose que la base de données est suffisamment grande pour que

$$CS(M) << CS(X|M)$$

Nous ne prenons donc en compte ici que $CS(X|M)$.
Raisonnons ici par l'absurde et supposons un cas où le modèle trouvé par l'algorithme soit différent du modèle optimal. Soit ainsi M_{opt} le modèle correspondant au minimum global de la complexité stochastique, et M_{loc} un modèle correspondant à un minimum local fourni par l'algorithme.
Supposons que $C1(M_{loc}) \supset C1(M_{opt})$, ce qui signifie que la couche 1 du modèle optimal est un sous-ensemble de la couche 1 du modèle trouvé par l'algorithme. Par définition, l'algorithme d'optimisation proposé passe un nœud de la couche 1 à la couche 2 à chaque étape tant que la complexité diminue. Ainsi, chacun des nœuds de l'ensemble $C1(M_{loc}) \cap C1(M_{opt})$ a été passé dans la couche 2 en diminuant la complexité. Ainsi, en partant de la configuration $C1(M_{opt})$, faire passer un de ces nœuds dans la couche 2 diminuerait vraisemblablement la complexité, ce qui est contradictoire avec l'hypothèse de modèle optimal.
Supposons donc que $C2(M_{loc}) \supset C2(M_{opt})$, ce qui signifie que la couche 2 du modèle optimal est un sous-ensemble de la couche 2 du modèle trouvé par l'algorithme. Soit $n_{opt} = card(C2(M_{opt}))$ le nombre de nœuds de la couche 2 du modèle optimal, et $n_{loc} = card(C2(M_{loc}))$ le nombre de nœuds de la couche 2 du modèle renvoyé par l'algorithme. A l'étape n_{loc}, faire passer n'importe quel nœud de la couche 1 à la couche 2 ne fait qu'augmenter la complexité. Partant de la configuration M_{loc}, ajoutons itérativement des nœuds appartenant à $C2(M_{loc}) \cap C2(M_{opt})$ à la couche 2. En notant C_j la complexité de l'algorithme à l'étape j, étant donné que, par hypothèse, $CS(M_{opt}) < CS(M_{loc})$
$\exists j \in \{n_{loc}+1, ..., n_{opt}\} \backslash C_j < C_{j-1}$
Ainsi, nécessairement, on a $(C1(M_{loc}) \cup C1(M_{opt}) - (C1(M_{loc}) \cap C1(M_{opt}))) \neq \emptyset$ et $(C2(M_{loc}) \cup C2(M_{opt}) - (C2(M_{loc}) \cap C2(M_{opt}))) \neq \emptyset$. Considérons l'ensemble non vide $Int = C2(M_{loc}) \cup C2(M_{opt}) - (C2(M_{loc}) \cap C2(M_{opt}))$. Si les nœuds appartenant à $Int \cap C1(M_{opt}$ correspondent à un minimum global de complexité, cela implique par définition que les modèles associés à chacun de ces nœuds décrivent bien la base de donnée qui leur est associée sous forme d'un modèle de mélange des modèles de la couche 0. Considérons à présent le modèle M_u tel que $C1(M_u) = C1(M_{opt}) \cap C1(M_{loc})$ et $C2(M_u) = C2(M_{opt}) \cap C2(M_{loc})$. La différence de complexité par rapport au modèle M_{opt} peut s'exprimer sous la forme de deux termes :

$$C(M_u) - C(M_{opt}) = \Delta_1 + \Delta_2$$

.

Cette différence est positive par définition de M_{opt} comme modèle optimal. Δ_1 correspond à la modification de la complexité associé aux nœuds de Int, et Δ_2 à la modification de complexité des nœuds de $C2(M_{opt})$ résultant du passage des nœuds de Int en première couche.
Si Δ_1 a augmenté, cela veut dire que les nœuds de $Int \cap C2(M_{opt}))$ interviennent de façon très forte dans le modèle de mélange des nœuds de $Int \cap C2(M_{loc}))$ dans le modèle renvoyé par l'algorithme. Or, si les nœuds de $Int \cap C2(M_{opt}))$ ont été mis en deuxième couche dans le modèle M_{opt}, c'est qu'ils s'expriment de façon efficace comme un mélange des nœuds de la première couche, soit par une combinaison linéaire des lois de probabilité des modèles de la première couche. Ainsi, les nœuds de $Int \cap C2(M_{loc}))$ devraient également s'exprimer comme une combinaison linéaire des modèles de la première couche, et donc permettre une diminution de la complexité.
Le raisonnement est totalement symétrique pour le terme Δ_2. Ainsi, le terme $\Delta_1 + \Delta_2$ et, selon notre raisonnement le modèle M_u devrait avoir une complexité plus faible que celle de M_{opt}, ce qui est contradictoire.

4.4 Modélisation associée à la relation de type "part-of"

Dans cette section, on considère des réseaux sémantiques hiérarchisés en couches où le seul lien sémantique existant est "part-of" qui lie deux concepts entre deux couches successives. Le fait que les concepts $c_1, ..., c_k$ soient reliés au concept a par la relation "part-of" signifie d'un point de vue sémantique que les concepts $c_1, ..., c_k$ constituent des sous-parties du concept a (voir Section 4.2.1). La modélisation que l'on associe à cette structure sémantique se traduit par le fait que le modèle a correspond à la modélisation hiérarchique détaillée au chapitre 3. Ainsi, chaque image d'index i de la base de données associée à a est partitionnée en régions annotées $\{R_{i1}, ..., R_{im_{ai}}\}$. On suppose que chaque région R_{ij} correspond à un sous-ensemble 4-connexe de textons de l'image d'index i et est annotée par un concept $c \in \{c_1, ..., c_k\}$. On notera m_{ai} le nombre de régions trouvées dans l'image i, $c(R_j)$ le concept annotant la région R_j et $x(R_j)$ l'histogramme des textons à l'intérieur de cette région.
Par souci de simplification, nous supposons que le nombre maximal de couches du modèle global est de deux. Cette modélisation qui est détaillée ici est généralisable à un nombre de couches quelconque.

4.4.1 Nœuds de la première couche

Les concepts de la couche 1 sont, comme dans la section précédente, associés à une modélisation directe des textons de l'image. Ainsi, la propriété 4.1 est toujours vérifiée et si $c \in C_1(M)$, on écrit :

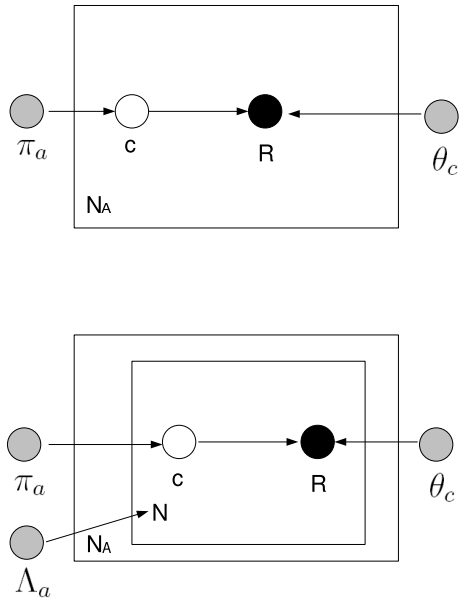


Fig. 4.6 – Représentation des modèles"kind-of" (en haut), et "part-of" (en bas). Les carrés correspondent au tirage successif et indépendant des variables aléatoires qui sont à l'intérieur avec un nombre de tirages égal au nombre inscrit en bas à gauche. Un disque coloré correspond à une variable aléatoire observable, un disque non coloré à une variable aléatoire non observable, et un disque grisé à un lot de paramètres. Le carré extérieur correspond aux N_A images exemples. Dans le modèle "part-of", le carré intérieur correspond aux différentes régions annotées.

$$P(x|M_c) = Poiss_{\lambda_c}(\sum_{j=1}^{n_0} x_j) \prod_{j=1}^{n_0} \theta_c^{x_j} \tag{4.16}$$

.

4.4.2 Nœuds de la deuxième couche

Soient a est un concept appartenant à la deuxième couche et i l'index de l'image. Le modèle génératif est le suivant :
- m_{ai} est choisi avec la loi $Poiss_{\Lambda_a}$.
- une partition de l'image $\{R_1, R_2, ..., R_{m_{ai}}\}$ est choisie avec une loi uniforme.
- Pour j variant de 1 à m_{ai}, un concept $c(R_j)$ est choisi parmi $\{1, ..., n\}$ avec probabilité $\{\pi_{a1}, ..., \pi_{an}\}$ et la probabilité de l'histogramme des textons à l'intérieur de la région est calculée conditionnellement au concept c : $P(R_j|c(R_j))$.

Par conséquent, on écrit la vraisemblance de l'image de la manière suivante :

$$\begin{aligned} P(X_{ai}, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(X_{ai}|M_a, \{R_1, R_2, ..., R_{mai}\}, \{c(R_1), c(R_2), .., c(R_m)\}) \\ P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) \end{aligned} \tag{4.17}$$

.

Le premier terme de ce produit s'écrit :

$$\begin{aligned} P(X_{ai}|M_a, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}) = \\ Poiss_{\Lambda_a}(m_{ai}) \prod_{j=1}^{m_{ai}} P(x(R_j)|c(R_j)) \end{aligned} \tag{4.18}$$

.

où le terme $P(x(R_j)|c(R_j))$ s'écrit avec la formule 4.16, $c(R_j)$ étant un concept de la première couche.
On suppose une indépendance entre les annotations et le choix de la partition de l'image en régions conditionnellement au modèle M_a. Par conséquent, on écrit le deuxième terme de l'expression 4.17 de la manière suivante :

$$\begin{aligned} P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(\{R_1, R_2, ..., R_m\}|M_a) P(\{c(R_1), c(R_2), .., c(R_m)\}|M_a) \end{aligned} \tag{4.19}$$

.

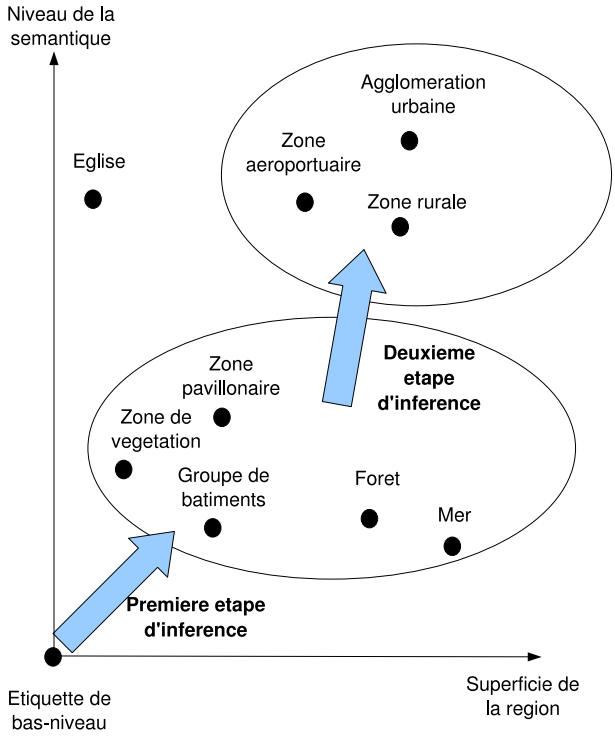


Fig. 4.7 – Illustration du niveau de sémantique extraite à chaque étape d'inférence en utilisant un réseau sémantique avec relation de type "part-of".

Les concepts sont supposés indépendants conditionnellement au modèle M_a, on a donc : $P(\{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \prod_{j=1}^{m} P(c(R_j)|M_a)$. Et par définition $P(c(R_i) = c_k) = \pi_{ak}$ est un paramètre du concept a. On pose une loi uniforme sur l'ensemble des partitions de l'image. On a donc $P(\{R_1, R_2, ..., R_m\}|M_a) = K$, où K est égal à l'inverse du nombre de partitions possibles dans l'image avec des régions 4-connexes, nombre dépendant de l'image et que nous ne cherchons pas ici à calculer. Λ_a est un paramètre portant sur le nombre de régions annotées avec les concepts de la couche 1 que l'on trouve dans la base d'images de a.

4.4.3 Expression de la probabilité globale

Comme dans le réseau avec lien de type kind-of, les différents lots d'apprentissage X_i sont supposés être indépendants conditionnellement au modèle global M. L'expression 4.5 démontrée dans la section précédente s'applique toujours et la vraisemblance de la base de données s'écrit donc :

$$P(X_1, X_2, ..., X_n|M) = \prod_{c \backslash M_c \in C1(M)} P(X_c|M_c) \prod_{c \backslash M_c \in C2(M)} P(X_c|M) \qquad (4.20)$$

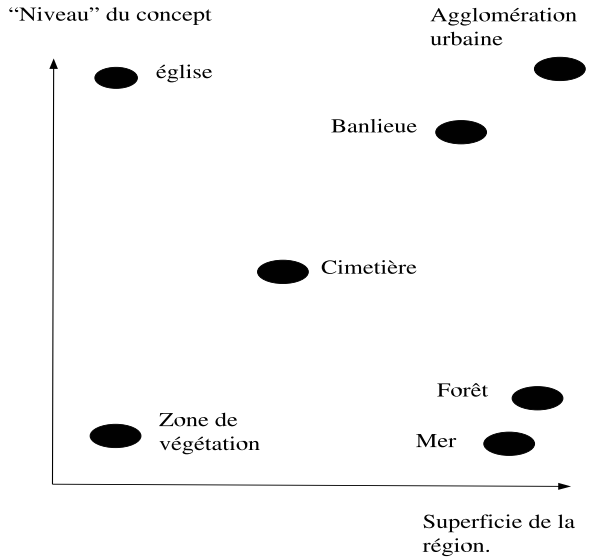


Fig. 4.8 – Représentation de différents concepts que l'on attache à des superficies et des "niveaux de sémantique" différents

4.4.4 Analyse de la modélisation

Comparaison des réseaux sémantiques avec lien "part-of" et "kind-of". Avec la modélisation "part-of", chaque image est partitionnée en régions et la vraisemblance de chaque région est exprimée par une des lois d'un modèle de la couche inférieure. La relation "kind-of" est, quant à elle, modélisée par une loi pour laquelle chaque image de la base de données est générée intégralement et de façon pondérée par chacune des lois du mélange. Cette différence fondamentale entre ces deux modélisation est schématisée figure 4.6.

Propriété d'extensivité Comme dans la section précédente, la propriété dite d'extensivité est vraie également avec la modélisation de type "part-of". La preuve effectuée en 4.3.1.4 est rigoureusement valable et peut être réutilisée telle quelle car l'argument d'inclusion explicité dans l'équation 4.6 s'applique encore.

4.5 Optimisation de la complexité stochastique pour le réseau sémantique avec lien de type "part-of"

Comme en section 4.3.2, la longueur de codage nécessaire pour coder la base de données est également séparée en deux termes : $CS(X, M) = CS(X|M) + CS(M)$. Les bases d'images étant supposées indépendantes pour chaque concept, on somme la longueur de description de la base de données associée à chaque concept : $CS(X, M) = \sum_{c=1}^{n} [CS(X_c|M) + CS(M_c)]$.

4.5.1 Codage de la couche de niveau 1

Les modèles de la couche 1 sont supposés générer directement les textons des images exemples qui leur sont associés. Ainsi, $C(X_c|M) = C(X_c|M_c)$ et si X_{cj} est la j-ème image exemple associée au concept c, en utilisant la formule de Shannon [117], le terme $CS(X_c|M_c)$ s'écrit :

$$CS(X_c|M_c) = -\log P(X_c|M_c)$$

.
avec la probabilité définie en 4.16. Les images X_{cj} de la base d'images fournies pour le concept c étant supposées indépendantes, l'équation précédente s'écrit :

$$CS(X_c|M_c) = -\sum_j \log P(X_{cj}|M_c)$$

.

Soit, en posant $card(X_{cj}) = \sum_{j=1}^{n_0} x_{cj}$ le nombre total de textons dans l'image X_{cj}, et en introduisant la formule 4.2 dans l'equation précédente :

$$CS(X_c|M_c) = \lambda_c - card(X_{cj}) \log \lambda_c + \sum_{j=1}^{card(X_{cj})s} \log(j) - \sum_{j=1}^{n_0} x_{cj} \log(\theta_c) \qquad (4.21)$$

.

Pour coder le modèle M_c, il faut tout d'abord coder le numéro de la couche auquel il appartient. Ici, nous ne considérons qu'un modèle possédant au maximum deux couches, nous codons donc le numéro de couche par un bit à valeur 0 ou 1. Il reste ainsi à coder les paramètres de génération des textons θ_c, et le paramètre de taille Λ_c. Pour cela nous utilisons la formule introduite par Rissanen [109] qui attribue au codage d'un vecteur de paramètres de taille T estimé avec N_{ech} la longueur de codage :

$$\frac{T}{2} log N_{ech} \qquad (4.22)$$

.

Le vecteur θ_c est de taille n_0 et le nombre d'échantillons avec lequel il est estimé est égal au nombre total de textons de la base X_c. Le vecteur Λ_c est de taille 1 et le nombre d'échantillons avec lequel il est estimé est égal au nombre total d'images de la base, soit par définition N_c.

$$CS(M_c) = \frac{n_0}{2} \log(\sum_{j=1}^{N_c} \sum_{k=1}^{n_0} x_{cjk}) + \frac{1}{2} \log(N_1) \qquad (4.23)$$

.

Ainsi, la complexité globale $CS(X_c|M_c)$ s'écrit :

$$CS(X_c|M_c) = \lambda_c - card(X_{cj})log\lambda_c + \sum_{j=1}^{card(X_{cj})s} logj - \sum_{j=1}^{n_0} x_j \log(\theta_c) + \frac{n_0}{2}\log(\sum_{j}^{N_c}\sum_{k}^{n_0} x_{cjk}) + \frac{1}{2}\log N_1 + 1 \quad (4.24)$$

.

4.5.2 Codage de la couche de niveau 2

c étant un concept d'une couche de niveau supérieur à 2, i l'index d'une image dans la base de données X_c et $P_i = \{R_{i1}, R_{i2}, ..., R_{im_{ci}}\}$ une partition en régions annotées de cette image, le terme $C(X_c|M, P_i)$ s'écrit :

$$C(X_{ci}|M, P_i) = -logP(X_{ci}|M, P_i)$$

.

On exprime cette probabilité en effectuant une sommation sur la probabilité jointe de l'image et des concepts sur toutes les annotations possibles de l'image, étant donné une partition :

$$P(X_{ci}|M, P_i) = \sum_{\{c(R_1),c(R_2),...,c(R_{m_{ci}})\}} P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i)$$

.
Or, le terme $P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i)$ se décompose de la façon suivante :

$$P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i) = P(X_{ci}|\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}, M, P_i)P(\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M) \quad (4.25)$$

.

Le premier terme de ce produit est exprimé par la formule 4.18, et le deuxième par la formule 4.19. On écrit donc :

$$P(X_{ci}|M, P_i) = \sum_{\{c(R_1),c(R_2),...,c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (4.26)$$

.

Cependant, cette expression est difficile à estimer pour un nombre de régions élevé. Certains algorithmes d'approximations peuvent être utilisés pour trouver une borne inférieure aussi proche que possible de cette expression. Cependant, dans le présent travail, nous nous contentons pour l'instant d'une borne inférieure très grossière, à savoir :

$$P(X_{ci}|M,P_i) = max_{\{c(R_1),c(R_2),...,c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (4.27)$$

.

Il est nécessaire de coder les paramètres de génération π_c des concepts de la couche 1. Pour chaque modèle, on code ainsi le numéro de la couche auquel il appartient et le vecteur de paramètres π_c de la loi de la probabilité de la variable latente. π_{cj} n'est non nul que pour les concepts de la couche inférieure reliés par un lien sémantique de type "part-of" au concept c. π_c a un nombre de paramètres égal au nombre de nœuds de la couche précédente, et qui est estimé avec le nombre d'images de la base, soit par définition N_c. En utilisant la formule 4.9 et en supposant que c appartient à la couche 2 :

$$CS(M_i) = \frac{n_1}{2} \log(N_c) \quad (4.28)$$

.

Il s'agit pour les concepts appartenant à la couche 2, comme pour les concepts de la première couche, de coder les valeurs des textons des images sans notion d'ordre ni de relations spatiales avec une longueur de code minimale. Ainsi, la partition P_i pour chaque image d'index i n'est pas codée et la longueur de code $C(X,M)$ est exprimée comme la partition permettant une complexité stochastique minimale :

$$C(X,M) = min_{P,M}(C(X|M,P) + C(M))$$

.

4.5.3 Algorithme d'optimisation utilisé

Le même algorithme que celui détaillé en 4.3.4 est utilisé, seules les formules d'estimation des paramètres sont différentes pour les modèles de la couche 2. Etant donné un modèle c de la deuxième couche, et i un index d'une image, on note $P(X_{cj} = \{R_{1j}, ..., R_{m_j j}\}$ l'annotation optimale qui a été trouvée de cette image, et $P(X_c) = \{P(X_{c1}), P(X_{c1}), ..., P(X_{cN_c})\}$.
On a donc, pour tout modèle c de la deuxième couche les formules suivantes :

$$\forall j \in \{1, ..., n_1\}, \pi_{cj} = \frac{occ_{P(X_c)}(c(R_i) = c_j)}{card(P(X_c))} \quad (4.29)$$

.

$$\Lambda_c = \frac{1}{N_c} \sum_{j=1}^{N_c} card(P(X_{cj})) \tag{4.30}$$

.
$occ_{P(X_c)}(c(R_i) = c_j)$ est par définition le nombre d'images qui ont été annotées avec le concept c_j dans l'ensemble des régions annotées de la base d'images X_c.
$card(P(X_{cj}))$ est par définition le nombre de régions annotées dans la partition $P(X_{cj})$.
Comme pour l'algorithme du chapitre 2, le modèle résultat est pris comme modèle optimal. On obtient ainsi le réseau sémantique résultat en gardant la disposition des concepts obtenue dans les différentes couches et en créant un lien de type "kind-of" entre un concept a de la couche 2 et un concept b de la couche 1 si la probabilité que la variable latente L_a prenne la valeur b soit supérieure à un seuil fixé arbitrairement :

$$P(L_a = b) > seuil$$

Ici, on prend le seuil égal à 0 et au lien créé est ajouté une valeur, qui correspond à la probabilité correspondante.

4.5.4 Remarque

Dans le cas d'un où l'on ne souhaite pas construire automatiquement la structure du réseau (car un réseau sémantique est déja à disposition), il est juste nécessaire d'estimer les paramètres des modèles. Ainsi, il est uniquement nécessaire d'appliquer les formules 4.13 et 4.14 pour les modèles de la couche 1, et 4.29 et 4.30 pour les modèles de la couche 2.

4.6 Réseau sémantique intégrant méronymie, synonymie et hyponymie

Jusqu'à présent, seuls des réseaux sémantiques ne pouvant contenir qu'un seul type de relation sémantique ont été considérés. Cependant, pour représenter de manière pertinente des structures sémantiques plus complexes et intégrer des concepts plus variés, il convient de spécifier une structure générale de modèle sémantique hiérarchique intégrant des liens de type "kind-of" et des liens de type "part-of". Cependant, définir une telle structure n'a rien d'évident car la méronymie, relation sémantique qui correspond au lien "part-of", et l'hyponymie, qui correspond au lien de type "kind-of", introduisent des hiérarchies de nature totalement différente. La méronymie introduit en effet une hiérarchie de type "tout/partie de" tandis que l'hyponymie introduit une hiérarchie de type "général/spécifique". Il n'y a donc pas une hiérarchie naturelle qui s'impose quant à la manière d'intégrer ces deux types de lien dans un même réseau. Cependant, afin de pouvoir appliquer une

modélisation qui soit aussi simple que possible, nous souhaitons imposer des contraintes assez strictes sur la structure des réseaux que nous prenons en compte. Dans cette section, nous commençons par détailler comment la relation de synonymie est prise en compte, ensuite nous expliquons la structure globale du réseau que nous définissons pour intégrer les relations de synonymie, méronymie et hyponymie.

4.6.1 Relation de synonymie

Soit un vocabulaire $\Omega = \{c_1, ..., c_n\}$ et un réseau sémantique S_Ω dont les nœuds sont les éléments de Ω. On nomme M_Ω un modèle statistique qui est mis en relation avec S_Ω. Soient deux concepts c_i et c_j, le fait que ces concepts soient synonymes dans S_Ω est traduit par le fait qu'ils sont attachés au même modèle dans M_Ω. Ainsi, si M_k est le modèle rattaché à c_i et c_j, l'ensemble des concepts rattachés à M_k s'écrit : $c(M_k) = \{c_i, c_j\}$.
La relation de synonymie peut être inférée, comme les relations d'hyponymie et de méronymie, par un algorithme de sélection de modèles. En effet, supposons que les deux concepts c_i et c_j appartiennent à la première couche du réseau sémantique et notons X_i et X_j les deux bases de données d'images exemples qui leur sont associés. Les modèles M_i et M_j appartenant à la première couche de M_Ω, ils contiennent respectivement les paramètres θ_i et θ_j de génération des textons estimés sur les bases X_i et X_j (voir équation 4.2). Si les deux concepts ne sont pas synonymes, la complexité stochastique de chacun des modèles M_i et M_j qui leur sont associés s'écrit :

$$C(M_i, X_i) = C(M_i) + C(X_i|\theta_i)$$

.

$$C(M_j, X_j) = C(M_j) + C(X_j|\theta_j)$$

.

Démontrons que, pour des bases de données suffisamment grandes, l'algorithme de minimisation de la complexité stochastique peut mettre en évidence la relation de synonymie.

preuve : Nous avons montré précédemment que $C(M_i) \sim log(|X_i|)$ (4.10) car la complexité stochastique nécessaire pour coder le modèle est proportionnelle à la taille du vecteur de paramètres à coder et est aussi proportionnelle au logarithme du nombre d'échantillons avec lesquels sont estimés les paramètres. Ainsi, écrivons la différence des complexités stochastiques dans un cas où la vraisemblance de la base X_i est calculée avec le modèle M_j :

$$C(X_i|\theta_i) - C(X_i|\theta_j) = -\log P(X_i|\theta_i) + \log(P(X_i|\theta_j))$$

.

Par indépendance des images de la base :

$$C(X_i|\theta_i) - C(X_i|\theta_j) = \sum_{k=1}^{n_i} [\log P(X_{ik}|\theta_j) - \log(P(X_{ik}|\theta_i))]$$

.

On pose à présent $Y_k = \log P(X_{ik}|\theta_j) - \log(P(X_{ik}|\theta_i))$. Étant donné que les concepts c_i et c_j sont synonymes, on suppose que les variables Y_k sont de moyenne nulle , de variance K, et sont de même loi.
Ainsi, le théorème des valeurs centrales s'applique et permet de dire que $\frac{C(X_i|\theta_i)-C(X_i|\theta_j)}{\sqrt{n_i}}$ converge en loi vers une loi gaussienne de moyenne nulle et de variance K. Or, les concepts c_i et c_j étant synonymes, on peut supposer que la variance K tend vers 0 avec la taille de la base de données. De plus, étant donné que $C(M_i) \sim log(|X_i|)$, la probabilité que $\frac{C(X_i|\theta_i)-C(X_i|\theta_j)}{C(M_i)} < 1$ tend vers 1 avec la taille de la base de données.
Ainsi, si cette inégalité est vérifiée, on a :

$$C(M_i) + C(X_i|\theta_i) + C(X_j|\theta_i) < C(M_i) + C(X_i|\theta_i) + C(M_j) + C(X_j|\theta_j)$$

.

Ce qui signifie que la complexité stochastique est plus faible en associant les deux concepts au même modèle plutôt qu'en apprenant des modèles distincts pour chaque concept.
Ainsi, une procédure de sélection de modèle permet d'inférer une relation de synonymie entre deux concepts. On remarque que contrairement aux relations de méronymie et d'hyponymie, qui sont mises en évidence par le terme d'attache aux données, c'est le terme de codage du modèles qui permet de déterminer la présence d'une relation de synonymie.

Discussion Comme on l'a vu précédemment, la diminution de la complexité stochastique entrainée par l'introduction d'un lien de synonymie entre deux concepts provient du terme de codage des paramètres. Ainsi, dans le cas où la taille de la base d'apprentissage est insuffisante, deux concepts peuvent être reliés par un lien de synonymie lors de la procédure de sélection de modèles même si ces concepts ne correspondent pas au même type de régions. En effet, si la taille de la base de données associée à un concept est trop réduite, le terme d'attache aux données est très faible par rapport au coût de codage des paramètres et lier ce concept avec un autre, même si celui-ci correspond à un type de régions très différent, fera baisser la complexité stochastique.
Pour mettre en évidence ce phénomène, prenons l'exemple de deux concepts correspondant à deux modèles stochastiques M_1 et M_2 modélisant des textons pouvant prendre n valeurs différentes. Le modèle M_1 génère avec probabilité 1 le texton de valeur 0, et le modèle M_2 génère avec probabilité 1 le texton de valeur 1. Si ces deux concepts sont considérés comme synonymes et que les bases de données X_1 et X_2 qui leur sont associés comportent le même nombre de textons, le modèle

M' qui est estimé sur la base $X_1 \cup X_2$ génère avec probabilité 0.5 le texton 0 et avec probabilité 0.5 le texton 1. Écrivons la différence des complexités stochastiques correspondant respectivement au cas où les concepts sont considérés comme synonymes et au cas où les concepts ne sont pas considérés comme synonymes :

$$C(X, M') - C(X, M) = C(X|M') + C(M') - C(X|M) - C(M)$$

.

$$C(X, M') - C(X, M) = C(X|M') + C(M') - C(X_1|M_1) - C(M_1) - C(X_2|M_2) - C(M_2)$$

.

Les modèles M_1 et M_2 étant purs, les termes d'attaches aux données sont nuls, on obtient ainsi l'expression suivante :

$$C(X, M') - C(X, M) = C(X|M') + C(M') - C(M_1) - C(M_2)$$

.

En remplaçant les termes de codage des paramètres par leurs expressions (voir équations 4.9 et 4.10), on obtient :

$$C(X, M') - C(X, M) = N \log(2) - \frac{n}{2} \log(N) - \frac{n}{2} \log(2)$$

.

Posons $f_n(N) = N \log(2) - \frac{n}{2} \log(N) - \frac{n}{2} \log(2)$ définie pour $N \geq 1$
Quand N vaut 1, la fonction $f_n(N)$ est négative, ce qui signifie que la procédure de sélection de modèle entrainera la création d'un lien de synonymie entre les deux concepts, alors même que ces concepts correspondent à des fonctions de probabilité respectives totalement différentes. Quand N tend vers l'infini, la fonction $f_n(N)$ tend vers l'infini, ce qui signifie que la procédure de sélection de modèle n'entrainera pas la création d'un lien de synonymie (voir figure 4.9). Ainsi, la fonction $f_n(N)$ étant continue, elle passe nécessairement une fois par 0. Le passage de la fonction f_n par 0 signifie que $C(X, M') = C(X, M)$. Ceci veut dire que la taille d'apprentissage critique pour laquelle le coût de codage est le même avec un modèle commun pour les deux lots d'apprentissage que pour deux modèles estimés séparement pour chaque lot. Il est possible de démontrer que $f_n(N)$ passe une et une fois seulement par 0 pour une valeur N_0 strictement supérieure à 0. Il est donc nécessaire d'estimer le modèle avec une base d'apprentissage de taille très grande devant N_0 pour pouvoir avoir une sélection de modèle fiable. En effet, le cas $N < N_0$ correspond au cas où une relation de synonymie est faussement détectée à cause d'une base d'apprentissage trop réduite.
La valeur $N_0(n)$ des zéros de la fonction f_n en fonction du paramètre n sont illustrés figure 4.10. Léquation $f_n(N) = 0$ n'admet pas de solutions analytiques mais $N_0(n)$ est une fonction croissante qui semble évoluer linéairement. La taille de la base d'apprentissage doit être donc d'autant plus importante que le nombre de textons est élevé.

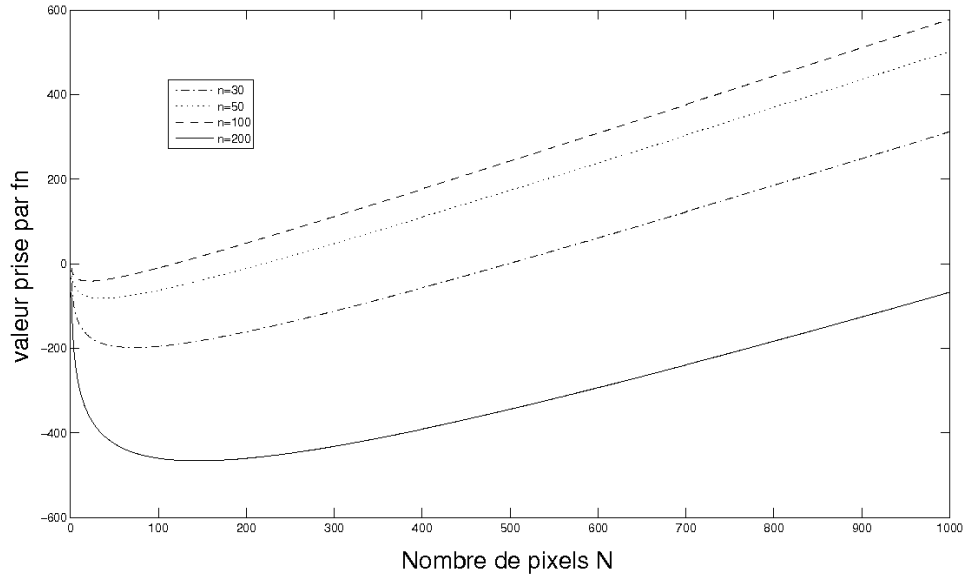


Fig. 4.9 – Graphe de la fonction $f_n(N)$ pour différentes valeurs de n. Le passage par zéro de chaque courbe correspond à la taille limite d'apprentissage nécessaire pour que la relation de synonymie puisse être inférée avec fiabilité.

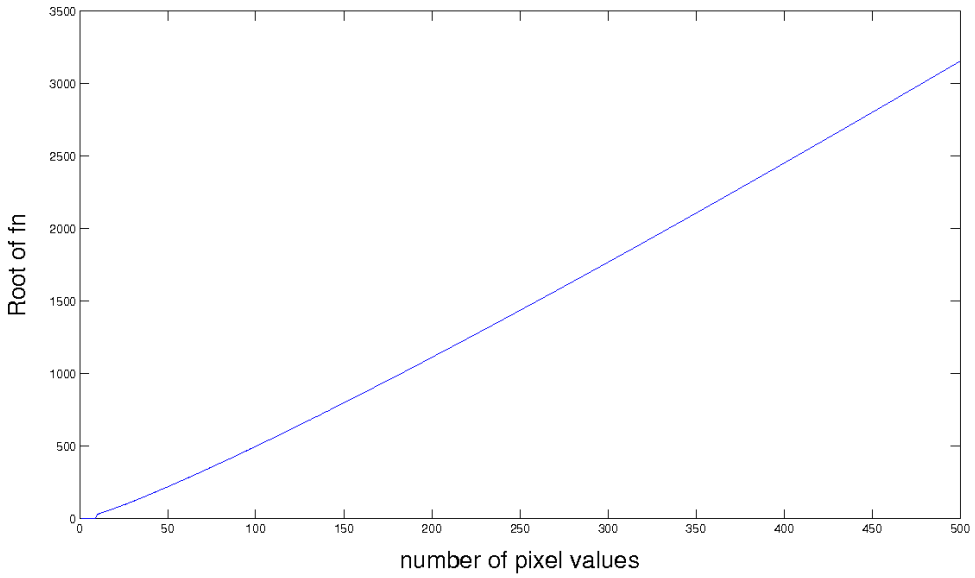


Fig. 4.10 – Zeros de la fonction $f_n(N)$ pour $N \geq 1$ en fonction du paramètre n. Ces valeurs correspondent, en fonction du nombre de valeurs possibles des textons, à une valeur plancher de la taille d'apprentissage nécéssaire à un concept pour en faire l'apprentissage.

4.6.2 Structure globale du réseau

La prise en compte du lien de synonymie est faite de façon très simple en associant plusieurs concepts sémantiques à un même modèle stochastique. Par contre, intégrer une hiérarchie de type méronymie et une hiérarchie de type hyponymie dans un même réseau n'a rien d'évident. Pour éviter des structures anarchiques qui seraient trop lourdes à modéliser, il est nécessaire d'imposer une contrainte très forte sur la structure du réseau et certains types de structures intégrant ces deux liens ont été proposées, telles que les structure et/ou (graphes "and/or", voir les articles [90] [142]). La stratégie que nous proposons consiste à considérer tout d'abord un réseau sémantique de type "part-of" sur lequel est superposé un réseau de type "kind-of". Ainsi, la première couche du réseau de type "kind-of" est constituée de l'ensemble des concepts du réseau de type "part-of".

Soit Ω un vocabulaire de concepts, le réseau sémantique global S_Ω^{tot} dont les nœuds sont les concepts de Ω contient deux sous-réseaux S_Ω^{ko} et S_Ω^{po}, correspondant respectivement au réseau de type "kind-of" et au réseau de type "part-of". La structure du premier vérifie les contraintes du réseau de type "kind-of" détaillé en 6.3, la structure du deuxiéme vérifie les contraintes du réseau de type "part-of" détaillé en 6.4. On suppose que chacun de ces réseaux ne peut avoir plus de deux couches. Pour combiner ces deux structures, on fait les 2 hypothèses suivantes :

– Les concepts placés en première couche de S_Ω^{ko} coincident avec l'ensemble des concepts contenus dans S_Ω^{po} :

$$C1(S_\Omega^{ko}) = C1(S_\Omega^{po}) \cup C2(S_\Omega^{po})$$

.

– Les ensembles de concepts placés dans la couche 2 de S_Ω^{ko} et dans la couche 2 de S_Ω^{po} sont disjoints :

$$C2(S_\Omega^{ko}) \cap C2(S_\Omega^{po}) = \emptyset$$

.

On fait donc le choix de ne pas imbriquer les structures "kind-of" et "part-of" car ces relations correspondent à des hiérarchies de nature totalement différente. On met ainsi la deuxième couche du réseau sémantique de type "kind-of" au dessus des deux couches du réseau de type "part-of". Ce choix est arbitraire mais la raison fondamentale part du constat que les couches supérieures du réseau de type part-of correspondent à des zones de plus grande complexité. Or, les concepts résidant dans les couches supérieures du réseau de type "kind-of" correspondant à des zones générales, elles décrivent des zones de complexité variable et nécessitent donc de s'appuyer sur des concepts appartenant à la couche supérieure du réseau de type "part-of". Ainsi, le concept général "urbain" pourra aussi bien correspondre à une zone de petite échelle comme "zone résidentielle pavillonnaire" qu'à une zone large et complexe comme "agglomération" (voir un exemple d'un tel réseau en figure 4.11).

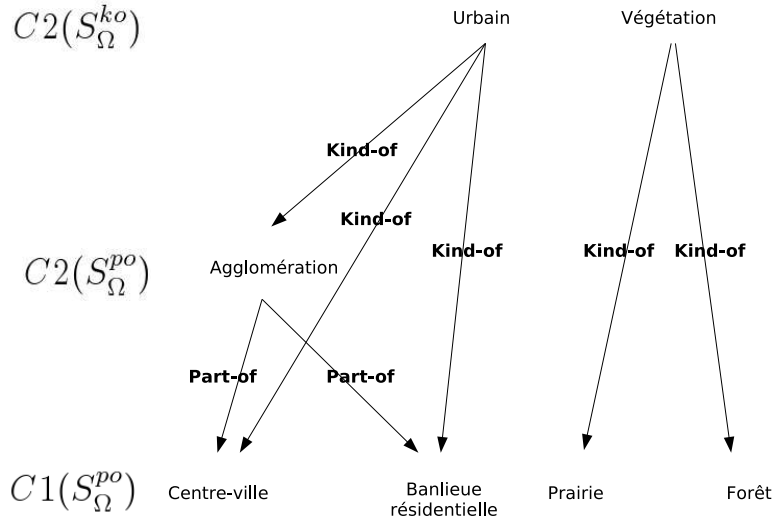


Fig. 4.11 – Exemple de structure admissible pour un réseau sémantique dans le cadre de la modélisation proposée. Une hiérarchie est imposée entre les différentes relations au sens où des relations d'hyponymie peuvent exister entre la couche $C2(S_\Omega^{ko})$ et $C2(S_\Omega^{po})$, mais, réciproquement, aucune relation de méronymie ne peut exister entre ces deux couches

4.6.3 Construction automatique du réseau

Soit n la taille totale du vocabulaire Ω de concepts considérés. Nous proposons ici un algorithme glouton itératif qui choisit à chaque étape la structure minimisant localement la complexité stochastique $CS(X, M_\Omega)$.

État initial La configuration de départ de l'algorithme est celle pour laquelle les modèles sont tous situés sur une couche. Ainsi, dans cette configuration initiale, les réseaux S_Ω^{po} sont confondus en une seule couche. Les paramètres de modèles de la première couche sont alors estimés en utilisant le maximum de vraisemblance : $M_\Omega^{initial} = argmax P(X|M_\Omega)$.

Évolution

- Pour chaque concept C_j de la couche du réseau S_Ω^{po}, on calcule la complexité stochastique associée à la configuration dans laquelle le concept C_j est mis dans la couche 2 du réseau S_Ω^{po} avec la formule 3.3.
- Pour chaque concept C_j du réseau S_Ω^{ko}, on calcule la complexité stochastique associée à la configuration dans laquelle le concept C_j est mis dans la couche 2 du réseau S_Ω^{ko} avec la formule 3.3.
- Pour tout couple de concepts du réseau S_Ω^{tot}, on calcule la complexité stochastique associée à la configuration dans laquelle ces concepts sont synonymes avec la formule 3.3.

Parmi toutes les configurations ainsi explorées, on retient celle minimisant la complexité stochastique. Si la condition d'arrêt de l'algorithme n'est pas vérifiée (voir paragraphe suivant), on retourne au début de la phase d'évolution avec le modèle correspondant à la configuration obtenue.

Condition d'arrêt de l'algorithme L'algorithme s'arrête lorsque la complexité stochastique augmente. Le dernier modèle ainsi obtenu est pris comme étant le modèle optimal. On obtient ainsi le réseau sémantique résultat en gardant la disposition des concepts obtenus dans les différentes couches et en créant un lien entre un concept a de la couche 2 et un concept b de la couche 1 si la probabilité que la variable latente L_a prenne la valeur b est supérieure à 0. La valeur associée au lien est alors $P(L_a = b)$.

4.7 Expériences

Des expériences ont été effectuées pour vérifier l'applicabilité de la mise en œuvre de construction automatique du réseau sémantique.

4.7.1 Données synthétiques

4.7.1.1 Relation de synonymie

Une fonction de probabilité gaussienne discrète de paramètres (λ, σ) $g_{\lambda,\sigma}(x)$ est définie et complétée sur les bords de manière à ce que $\sum_{i=1}^{256} g(i) = 1$. Deux lots de données X_1 et X_2 contenant chacun un nombre N images 200×200 sont générés à partir de cette même distribution et correspondent à deux concepts c_1 et c_2. Deux modèles M et M' sont estimés pour deux cas correspondant à deux structures différentes :

- Cas où c_1 et c_2 sont supposés ne pas être synonymes et appartiennent tous deux à $C_1(S_{ko})$. On a donc $M = \{M_1, M_2\}$ où M_1 est estimé par maximum de vraisemblance sur X_1 et M_2 sur X_2.
- Cas où c_1 et c_2 sont supposés être synonymes et appartiennent tous deux à $C_1(S_{ko})$. On a donc $M' = \{M'_1\}$ où M'_1 est estimé par maximum de vraisemblance sur $X = X_1 \cup X_2$.

La complexité stochastique est calculée dans chacun de ces cas :

$$C(X, M) = -logP(X_1|M_1) - logP(X_2|M_2) + C(M_1) + C(M_2)$$

.

$$C(X, M') = -logP(X_1|M') + C(M')$$

.

La figure 4.12 montre le rapport $R_{syn} = \frac{C(X,M)}{C(X,M')}$ en fonction de la taille de la base de données.

On constate sur la courbe 4.12 que R_{syn} tend asymptotiquement vers 1, ce qui est tout à fait cohérent. Rappelons en effet, comme il a été détaillé en 4.6.1, que les expressions de $C(X, M)$ et de $C(X, M')$ diffèrent simplement par l'expression du codage du modèle. $C(M)$ correspond au codage de deux modèles, tandis que dans le cas de M', les deux modèles sont supposés être synonymes et donc un seul modèle est codé. Or, le terme $C(M)$ évolue logarithmiquement avec la taille de la base de données, tandis que le terme $C(X|M)$ évolue linéairement. Ainsi, le rapport $\frac{C(X,M)}{C(X,M')}$ tend vers 1 asymptotiquement.

4.7.1.2 Relation d'hyponymie/hyperonymie

On considère k distributions gaussiennes g_i de moyenne $\sigma_i = \frac{i}{256}, i \in \{1, \ldots, 256\}$, et complétées sur les bords de manière à ce que $\sum_{j=1}^{256} g_i(j) = 1$. k lots de données X_i associés chacun à des concepts c_i sont générés chacun à partir de la distribution g_i contenant chacun N images de taille 200×200. Un lot X_{k+1} associé à un concept c_{k+1} est généré de la manière suivante (voir figure 4.7.1.2) :

- Pour $i \in \{1, \ldots, N\}$
  - Un nombre entier j est tiré avec probabilité uniforme dans $\{1, \ldots, 256\}$.
  - L'image $X_{(k+1)i}$ est générée en tirant indépendemment les textons avec probabilité g_i.

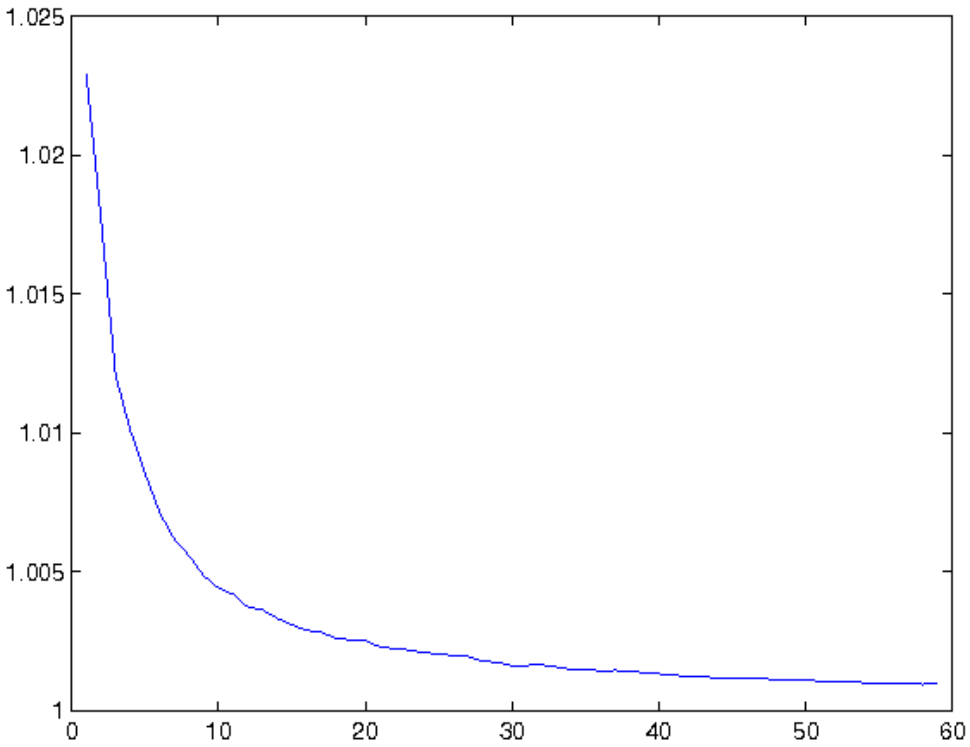


Fig. 4.12 – Rapport $\frac{C(X,M)}{C(X,M')}$ en fonction du nombre N d'images présentes dans X_1 et X_2

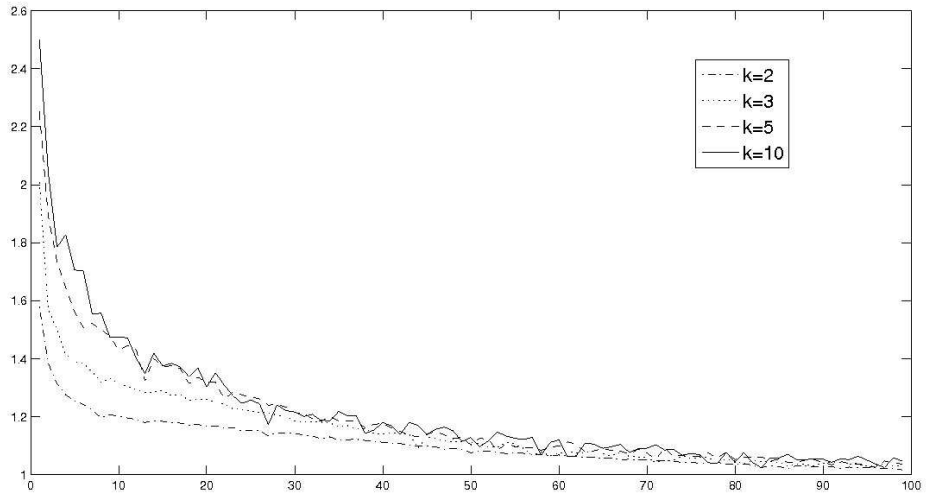
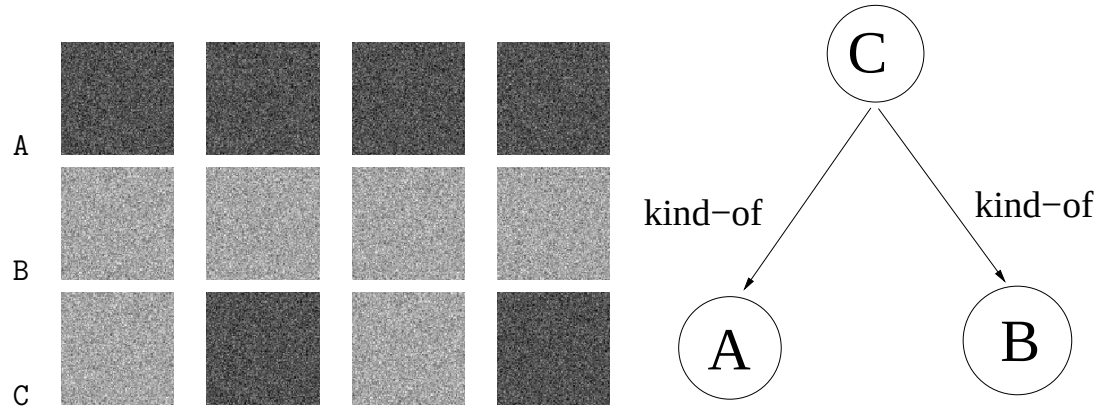


Fig. 4.13 – Rapport $\frac{C(X,M)}{C(X,M')}$ en fonction de σ et pour différentes valeurs de k



Deux modèles M et M' sont estimés pour deux cas correspondant respectivement à deux structures différentes :

- Cas où $\{c1,...,c_k\}$ et c_{k+1} sont tous situés sur la même couche. Le modèle M_{k+1} est estimé par maximum de vraisemblance sur X_{k+1} au même titre que les autres modèles.
- Cas où $\{c1,...,c_k\}$ sont supposés être hyponymes de c_{k+1}. Le modèle M_{k+1} est donc construit comme un modèle de mélange, et le vecteur de paramètres λ_{k+1} est estimé à partir de X_{k+1}.

La figure 4.13 montre le rapport $R_{syn} = \frac{C(X,M)}{C(X,M')}$ en fonction de l'écart type des gaussiennes. On voit que le rapport R_{syn} tend vers 1, ce qui signifie que la relation d'hyponymie est moins bien reconnue par le modèle. En effet, augmenter l'écart type revient à dire que les caractéristiques sont moins discriminantes. Moins les caractéristiques sont discriminantes, et moins l'introduction d'un lien d'hyponymie apporte une diminution de la complexité stochastique.

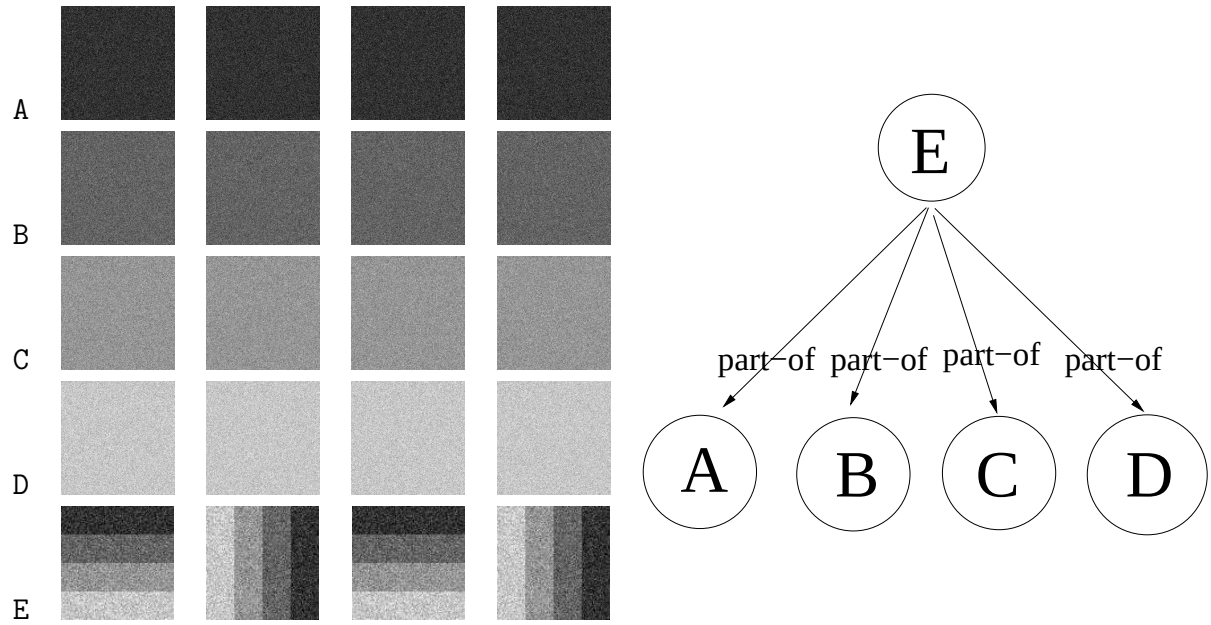


Fig. 4.14 – Illustration des expériences de validation du processus de construction automatique d'un réseau de type "part-of" à partir d'une base d'images synthétiques.

4.7.1.3 Relation de méronymie/holonymie

On considère k distributions gaussiennes g_i de moyenne $\sigma_i = \frac{i}{256}, i \in \{1, \ldots, 256\}$, et complétées sur les bords de manière à ce que $\sum_{j=1}^{256} g_i(j) = 1$. k lots de données X_i associés chacun à des concepts c_i sont générés chacun à partir de la distribution g_i contenant chacun N images de taille 200×200. Un lot X_{k+1} associé à un concept c_{k+1} est généré de la manière suivante (voir figure 4.7.1.3) :

- Pour $i \in \{1, \ldots, N\}$, l'image X_i, de taille $200 \times (200 * k)$ est décomposée en k régions $R_1(X_i), ..., R_k(X_i)$ de taille 200×200
  - Pour $i \in \{1, \ldots, k\}$
  - Chaque région $R_j(X_i)$ est générée en tirant indépendemment les textons avec probabilité g_j.

Deux modèles M et M' sont estimés pour deux cas correspondant respectivement à deux structures différentes :

- Cas où $\{c1,..., c_k\}$ et c_{k+1} sont tous situés sur la même couche. Le modèle M_{k+1} est estimé par maximum de vraisemblance sur X_{k+1} au même titre que les autres modèles.
- Cas où $\{c1,..., c_k\}$ sont supposés être méronymes de c_{k+1}. Le modèle M_{k+1} est donc construit comme un modèle de mélange, et le vecteur de paramètres λ_{k+1} est estimé à partir de X_{k+1}.

La figure 4.15 montre le rapport $R_{syn} = \frac{C(X,M}{C(X,M')}$ en fonction de l'écart type des gaussiennes. On voit que le rapport R_{syn} tend vers 1 avec la variance des gaussiennes. Ainsi, comme dans le cas de l'hyponymie, moins les caractéristiques sont discriminantes, et moins l'introduction d'un lien de méronymie apporte une diminution de la complexité stochastique.

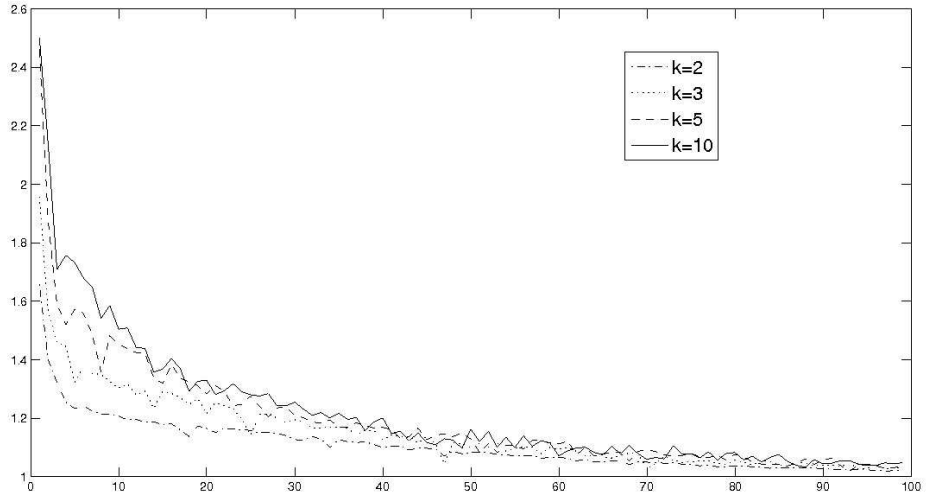


Fig. 4.15 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction de σ et pour différentes valeurs de k

Concepts	Forêt	Centre-Ville	Montagne	Zone résidentielle	Mer
Gain $\frac{C(X,M)}{C(X,M')}$	1.00031	1.00042	1.00061	1.00012	1.0064

Fig. 4.16 – Gains en Complexité stochastique obtenus pour différents concepts d'annotation. On constate que ce rapport est toujours supérieur à 1, ce qui signifie que le lien de synonymie est bien mis en évidence sur ces concepts

4.7.2 Données réelles

L'applicabilité de la mise en œuvre de construction automatique du réseau sémantique a été testée sur une base d'images SPOT5 à 2,5m de résolution. Cette base de données est constituée d'images exemples associées à différents concepts listés tableau 4.20.

4.7.2.1 Relation de synonymie

Pour évaluer la diminution de la complexité stochastique liée à l'introduction du lien de synonymie dans le cas où deux annotations sont introduites pour décrire un type de région similaire, nous effectuons le protocole exprimental suivant :
- Pour tout concept c du vocabulaire d'annotation.
  - La base d'images X_c associée au concept c est scindée en deux sous-bases X_c^1 et X_c^2 de tailles similaires et que l'on annote par deux concepts c_1 et c_2 et qui peuvent correspondre à "concept version 1" et "concept version 2".
  - La complexité stochastique est calculée dans le cas où c_1 et c_2 sont supposés ne pas être synonymes. Cette complexité stochastique est notée $C(X_c, M)$.
  - La complexité stochastique est calculée dans le cas où c_1 et c_2 sont supposés être synonymes. Cette complexité stochastique est notée $C(X_c, M')$.

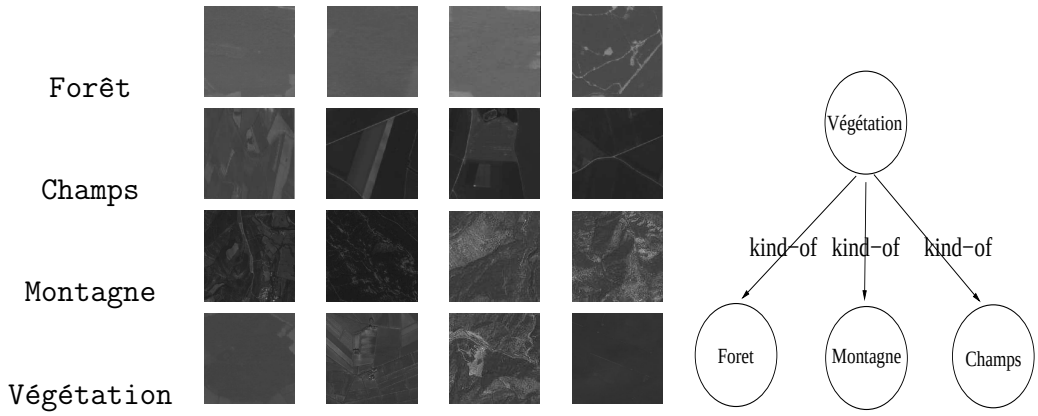


Fig. 4.17 – Des images exemple de forêt, de champs et de montagne sont fournies pour l'apprentissage et constituent des concepts correspondant à un seul type de zone. La classe de végétation contient par contre des zones de champs, de montagne, et de forêt fournis en proportions égales. On constate alors que l'ajout de la relation d'hyponymie et l'ajout d'un lien kind-of entre le concept de végétation et les trois autres concepts permet une diminution significative de la complexité stochastique. Cependant, ce gain dépend du pouvoir de description des caractéristiques de bas-niveau employées. ©CNES

Le rapport $\frac{C(X_c,M)}{C(X_c,M')}$ obtenu pour différents concepts est listé sur le tableau 4.7.2.1. On constate que ce rapport est supérieur à 1 pour tous ces concepts, ce qui prouve que la méthode proposée met en évidence un lien de synonymie lorsque plusieurs concepts sont introduits alors qu'ils correspondent à un même type de région.

4.7.2.2 Relation d'hyponymie/hyperonymie

Nous évaluons ici la diminution de la complexité stochastique liée à l'introduction du lien d'hyponymie/hyperonymie dans le cas où une annotation correspond à la généralisation d'un ensemble d'autres annotations présent dans le vocabulaire d'annotation.
Nous considérons k concepts $\{c_1, \ldots, c_k\}$, associés respectivement à des bases d'images $X_1, \ldots, X_k$. Un concept c est ensuite introduit et est associé à une base d'images exemples X_{k+1} qui contient une mélange à proportions égales d'images correspondant aux concepts $c_1, \ldots, c_k$. La complexité stochastique $C(X, M)$ de la base $X = \{X_1, \ldots, X_k, X_{k+1}\}$ est calculée avec un modèle M correspondant à une structure où les concepts $\{c_1, \ldots, c_k, c\}$ sont mis intégralement sur une même couche. la complexité stochastique $C(X, M')$ est ensuite calculée avec un modèle M' qui correspond à une structure où le concept c est en relation hyperonymique avec les concepts $\{c_1, \ldots, c_k\}$.
Une expérience a été menée dans le cas du concept *végétation*, qui correspond à une généralisation des concepts : *montagne*, *champs* et *forêt* (voir figure 4.7.2.2). Pour tester le gain $\frac{C(X,M)}{C(X,M')}$ en fonction du pouvoir de discriminance des caractéristiques de bas-niveau, on rajoute un bruit sur les textons de la manière suivante : étant donné b un pourcentage de bruit, pour tous les textons, avec une

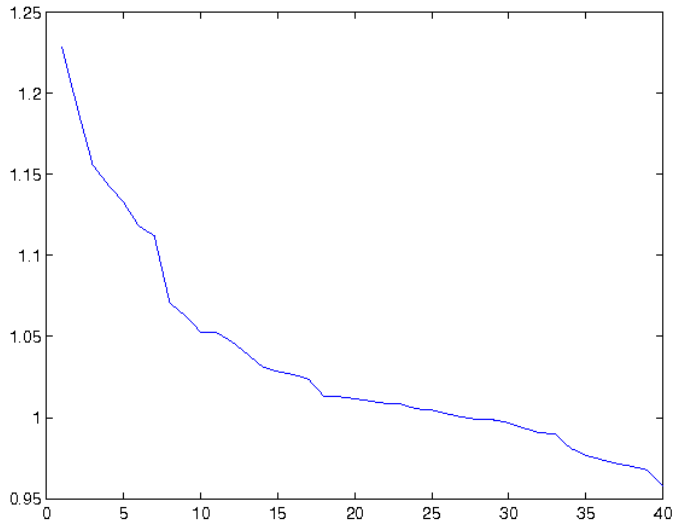


Fig. 4.18 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction du pourcentage de bruit montrant la diminution de complexité stochastique entrainée par l'ajout d'un lien d'homonyie enre le concept "végétation" et les concepts : "montagne", "champs", et "forêt".

probabilité b, la valeur du texton va être changée en une nouvelle valeur tirée avec une probabilité uniforme sur l'ensemble des valeurs possibles du texton. Le rapport $\frac{C(X,M)}{C(X,M')}$ est calculé pour différentes valeurs de b. La courbe de résultat est affichée en 4.18. On observe une diminution du gain relativement comparable aux résultats obtenus en 4.7.1.2 sur données synthétiques. Aux alentours de 20% de bruit, le rapport $\frac{C(X,M)}{C(X,M')}$ passe en dessous de 1, ce qui signifie que la relation d'hyponymie ne peut plus être identifiée par le système. Le système de construction du réseau sémantique nécessite donc une certaine discriminance des caractéristiques de bas-niveau. Et le gain apporté par l'introduction du lien d'hyponymie quant à la diminution de la complexité stochastique est lié directement au pouvoir de description des données des caractéristiques de bas-niveau.

4.7.2.3 Relation de méronymie/holonymie

Pour évaluer la diminution de la complexité stochastique liée à l'introduction du lien de méronymie/holonymie dans le cas où un concept est lié à un ensemble d'autres concepts par un lien de type "tout/partie de", une expérience a été menée dans le cas du concept *zone rurale*, lié par un lien "tout/partie-de" avec les concepts *habitations éparses*, *champs*, et *zone résidentielle*. La procédure est identique à celle qui a été menée dans la section 4.7.2.2, les résultats sont montrés figure 4.19 et les conclusions qui peuvent en être tirées sont très similaires à ce qui a été obtenu dans le cas de la relation d'hyponymie.

4.7.2.4 Construction d'un réseau sémantique complet

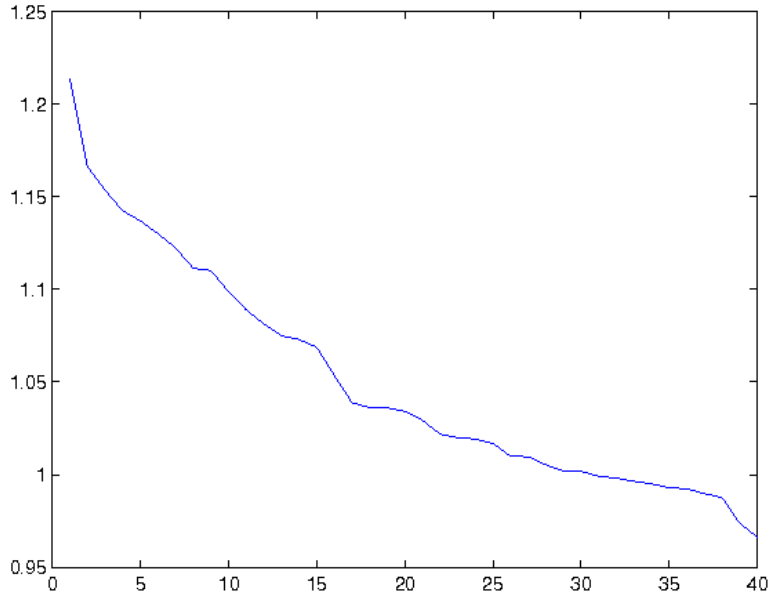


Fig. 4.19 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction du pourcentage de bruit montrant la diminution de complexité stochastique entrainée par l'ajout d'un lien de méronymie enre le concept "zone rurale" et les concepts : "habitations éparses", "champs", et "zone résidentielle".

Expérience sur images SPOT5 Les expériences ont été effectuées sur une base d'images SPOT5 à 2,5m de résolution prises sur les villes de Marseille, Nimes, Angers et Nice. La base d'apprentissage comprend 15 concepts, le nombre d'images exemples dans chaque lot d'apprentissage variant en fonction des concepts (voir tableau 4.20). Les caractéristiques de bas-niveau à partir desquelles est faite la modélisation sont, comme dans les chapitres précédents, des caractéristiques de Haralick qui sont quantifiées en un vocabulaire de taille 90 (voir Section 4.2.4). Les caractéristiques de la base d'apprentissage sont détaillées dans le tableau ci-dessous, ainsi que la complexité stochastique $C(X_i|M_i)$ à l'état initial correspondant au cas où le modèle M_i est placé dans la première couche.
La construction du réseau sémantique est limitée à un réseau "part-of" ayant au maximum deux couches.
La complexité stochastique diminue pendant 4 itérations puis remonte (voir figure 4.22). En 4.3.5, il a été démontré que la CS ne peut pas rediminuer après avoir augmenté une première fois, il n'est pas nécessaire de continuer à itérer l'algorithme. Le réseau sémantique résultat contient cinq concepts sur sa deuxième couche : Zone maritime, Banlieue industrielle, Zone montagneuse, Agglomération et Zone rurale.

Expérience sur images Quickbird Les expériences ont été effectuées sur une base d'images d'apprentissage Quickbird à 0,7m de résolution prises sur la ville de

Concept	**Nombre d'images exemples**	**Complexité stochastique**
Zone d'activité industrielle	15	3182
Zone résidentielle	14	2601
Champs	15	5054
Zone montagneuse	3	19846
Bois	15	2497
Eau	5	6832
Habitations éparses	15	15632
Centre ville	15	6804
Zone rurale	3	32668
Raffinerie	2	2093
Agglomération	3	27851
Banlieue industrielle	3	29434
Cimetière	2	4860
Carrière	2	3487
Montagne	12	4539
Aéroport	4	11349
Zone maritime	4	9234

Fig. 4.20 – Présentation de la base de données utilisée pour faire l'apprentissage du réseau sémantique

Concepts de la deuxième couche	**Concepts de la première couche**
Zone Rurale	Carrière, Bois, Habitations éparses, Champs, Zone résidentielle
Agglomération	Centre ville, Zone résidentielle, Cimetière
Banlieue industrielle	Raffinerie, Zone résidentielle, Zone d'activité
Zone montagneuse	Montagne, Carrière, Bois , Habitations éparses
Zone maritime	Eau, Zone résidentielle, Bois, Zone industrielle

Fig. 4.21 – Relations de type "part-of" inférées automatiquement par le système entre les concepts de la deuxième et de la première couche à partir de la base d'apprentissage d'images SPOT5 utilisée dans l'expérience.

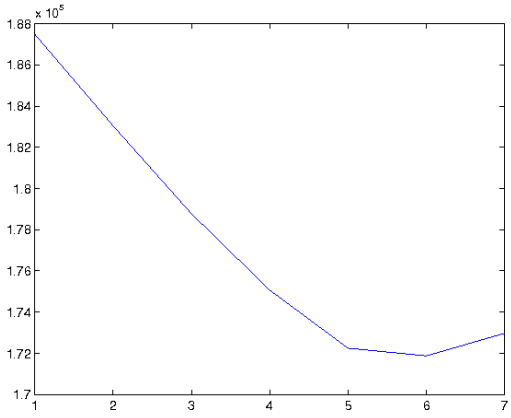


Fig. 4.22 – Evolution de la complexité stochastique en fonction du nombre d'itérations de l'algorithme d'apprentissage du réseau sémantique sur la base d'images SPOT5

Pékin. La base d'apprentissage comprend 32 concepts listés dans le tableau 4.23. La construction du réseau sémantique est limitée à un réseau "part-of" ayant au maximum deux couches. Le seuil sur la probabilité de génération d'un concept pour la création d'un concept sémantique est imposée à 0.01. La complexité stochastique diminue pendant 5 itérations puis remonte (voir figure 4.24). On remarque que l'allure de la courbe que l'on obtient est très semblable de celle obtenue lors de la construction du réseau sémantique sur la base d'image SPOT5. Les labels qui sont mis en deuxième couche sont : Aéroport, Zone rurale, Complexe industriel, Jardin public, Zone urbaine éparse, Zone résidentielle, Zone pavillonaire. Ces labels correspondent en effet à des zones de plus grande étendue et de plus grande complexité et leur emplacement en deuxième couche semble parfaitement justifiée.

4.8 Conclusion

Dans ce chapitre, une approche mettant en relation un espace de réseaux sémantiques vérifiant certaines contraintes et un espace de modèles stochastiques est développée. Cette méthode prend en compte les relations de synonymie, méronymie et d'hyponymie qui peuvent relier les concepts entre eux et de les associer à différent types de modélisations sur les caractéristiques de bas-niveau. Une structure sémantique étant déterminée, un modèle dont la structure est analogue au réseau sémantique en termes de dépendance des modèles permet alors de calculer la vraisemblance de la base d'images. Les couches supérieures de type "kind-of" contiennent des termes généraux, tandis que les couches supérieures de type "part-of" contiennent des concepts correspondant à des régions fortement structurées, de grande échelle et qui véhiculent donc une information sémantique de niveau élevé. Ces deux types de hiérarchie sont essentielles pour parvenir à annoter de façon pertinente des grandes bases d'images. Le critère de minimisation

Jardin	Zone commerciale
Grandes tours	Zone résidentielle
Serres	Zone pavillonaire
Champs	Hutong (Habition pékinoise traditionnelle)
Maisons individuelles	Bâtiments résidentiels intermédiaires
Bidonville	Grands bâtiments résidentiels
Chantier	Aéroport
Grande cour	Complexe industriel
Usine	Zone rurale
Entrepots	Jardin public
Petit jardin	Piste d'atterrissage
Hangars	Bois
Zone d'activité	Terminal d'aéroport
Installations sportives	Zone urbaine éparse
Lac	Prairie
Parking	Colline

Fig. 4.23 – Présentation des concepts utilisés pour faire l'apprentissage du réseau sémantique

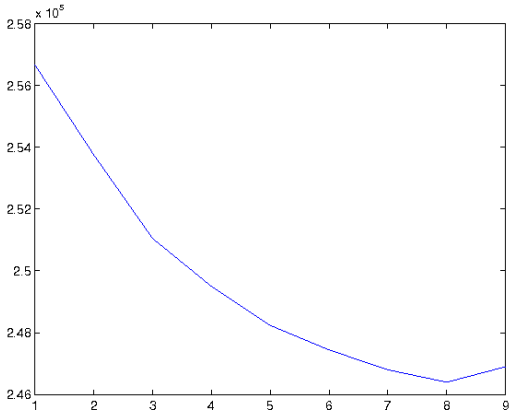


Fig. 4.24 – Evolution de la complexité stochastique en fonction du nombre d'itérations de l'algorithme d'apprentissage du réseau sémantique sur la base d'images Quickbird. ©LIAMA

de la complexité stochastique permet, étant donné une base de données annotée par un lot de concepts, de déterminer automatiquement la structure du réseau sémantique contenant ces concepts.

Chapitre 5

Annotation d'images tests

Nous étudions dans ce chapitre comment obtenir une annotation pertinente d'images test en utilisant la modélisation stochastique détaillée au chapitre précédent. En effet, la distribution des labels selon différentes couches va permettre d'annoter une image de test selon plusieurs niveaux de généralité et de complexité. Afin de pouvoir effectuer des requêtes dans une base d'images, des relations spatiales sont définies et sont rajoutées après le partitionnement d'une image de test en régions annotées. Des expériences sont ensuites menées sur des images SPOT5 et des images Quickbird pour comparer la qualité des annotations ainsi produites avec une méthode markovienne.

5.1 Méthode d'annotation sémantique d'une image test

Avant de décrire le mécanisme d'annotation sémantique que nous allons mettre en œuvre, rapprochons la de la démarche détaillée dans le chapitre précédent. Pour cela, rappelons l'approche qui a été détaillée jusqu'à présent. Nous avons considéré et modélisé trois différents types de liens sémantiques : la synonymie, la méronymie et l'hyponymie. La notion de synonymie est vue simplement comme le fait d'attacher plusieurs concepts à un même modèle. Les deux autres relations sont par contre mises en relation avec des modélisations statistiques du signal bien précises et différenciées. La relation d'hyponymie, qui introduit une relation de spécification entre deux concepts, est associée à un modèle de mélange d'unigrammes du signal. La relation de méronymie introduit une notion hiérarchique de type "tout/partie de", et une image annotée par un concept pourra être décomposée en régions annotées par d'autres concepts si ceux-ci sont reliés au premier par une relation de type "part-of".
Étant donnée à présent une image de test que l'on suppose de taille caractéristique supérieure aux images de la base d'apprentissage, l'objectif est de décomposer cette image en régions annotées par des concepts appris par le système. On est donc dans une situation où l'image peut être vue comme un tout et où l'on cherche des parties de l'image. La modélisation statistique sera donc naturellement proche

de celle employée pour la méronymie. Ainsi, il semble en effet pertinent de voir l'image comme étant annotée par un concept virtuel inconnu qui résiderait dans une troisième couche du modèle sémantique "part-of". C'est la raison pour laquelle le processus d'annotation mis en œuvre commencera par décomposer l'image avec les concepts présents dans le sous-réseau sémantique avec liens de type "part-of". Dans un deuxième temps, les concepts synonymes et les concepts résidant dans les couches supérieures du réseau sémantique de type kind-of seront ensuite surajoutés comme un complément d'annotation.

5.1.1 Modélisation d'une image de test

Soient I une image de test, $\{R_1, R_2, ..., R_m\}$ une partition annotée de cette image en régions, et S_Ω un réseau sémantique. On note $c(R_i)$ le label annotant la région R_i, où $c(R_i)$ appartient à l'ensemble des labels constituant la couche de plus haut-niveau du réseau de S_Ω^{po}. Comme dit précédemment, I est vue comme une image annotée par un label inconnu appartenant à une couche située à un niveau immédiatement supérieur de la couche de plus haut niveau du réseau S_Ω^{po} 5.1. La modèle génératif d'une image test est donc similaire à celui employé pour les images annotées par un label appartenant à la deuxième couche d'un réseau "part-of" 4.4.2 :

$$\begin{aligned} P(I, \{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M) = \\ P(I|\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}, M) \\ P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M) \end{aligned} \quad (5.1)$$

.

Le premier terme du produit s'écrit :

$$\begin{aligned} P(I|\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}, M) = \\ Poiss_\Lambda(m) \prod_{j=1}^{m} P(x(R_j)|c(R_j), M) \end{aligned} \quad (5.2)$$

.

où Λ est un paramètre qui doit être fixé manuellement. En pratique, le terme $Poiss_\Lambda(m)$ a peu d'influence sur le résultat de l'annotation et apparaît ici par cohérence avec la modélisation utilisée pour le réseau de type "part-of" (cf 4.4.2). La vraisemblance $P(x(R_j)|c(R_j), M)$ de chaque région R_j est exprimé en fonction de la couche à laquelle appartient le label $c(R_j)$. Ainsi, si $c(R_j) \in C1(S_\Omega^{po})$, $P(x(R_j)|c(R_j), M)$ est exprimée selon l'équation 4.16, et si $c(R_j) \in C2(S_\Omega^{po})$, $P(x(R_j)|c(R_j), M)$ est exprimée selon l'équation 4.17.

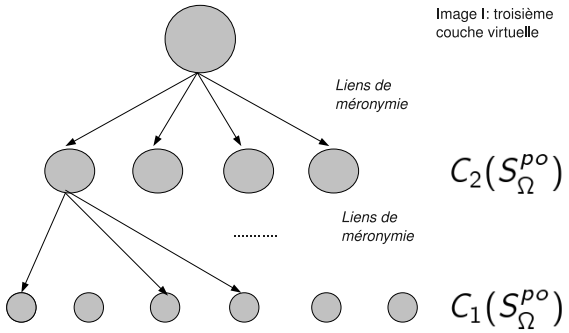


Fig. 5.1 – Représentation de la modélisation utilisée l'annotation d'une image de test. Dans le cas où le réseau sémantique "part-of" comporte deux couches, l'image de test I est vue comme une image annotée par un label appartenant à la couche 3 du réseau "part-of".

On suppose une indépendance entre les annotations et le choix de la partition de l'image en régions conditionnellement au modèle M. Par conséquent, le deuxième terme de l'expression 5.1 s'écrit de la manière suivante :

$$P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M) = \\ P(\{R_1, R_2, ..., R_m\}|M)P(\{c(R_1), c(R_2), .., c(R_m)\}|M) \quad (5.3)$$

.
Les labels $c(R_j)$ sont supposés indépendants conditionnellement au modèle M, on a donc : $P(\{c(R_1), c(R_2), .., c(R_m)\}|M) = \prod_{j=1}^{m} P(c(R_j)|M)$. Le label annotant I étant inconnu, tous les $c(R_j)$ sont supposés équiprobables : $P(c(R_j)|M) = \frac{1}{n}$.
Comme pour ce qui est fait en 4.4.2, une loi uniforme est définie sur l'ensemble des partitions de l'image. On a donc $P(\{R_1, R_2, ..., R_m\}|M_a) = K$, où K est égal à l'inverse du nombre de partitions possibles dans l'image avec des régions 4-connexes, nombre que nous ne cherchons pas ici à calculer.
La vraisemblance totale 5.1 de l'image I s'écrit donc :

$$P(I, \{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M) = K(\frac{1}{n})^m Poiss_\Lambda(m) \prod_{j=1}^{m} P(x(R_j)|c(R_j), M) \quad (5.4$$

5.1.2 Algorithme d'inférence

Le but de l'inférence est de trouver les partitions annotées les plus vraisemblables de l'image I. Dans le cas où S_Ω^{po} ne comporte qu'une seule couche, une seule

partition annotée G_1 doit être inférée. Ainsi, on choisit la partition maximisant la vraisemblance de l'image :

$$max_{G_1} P(I, G_1|M)$$

.

Dans le cas où S_Ω^{po} comporte deux couches, il est nécessaire de trouver deux partitions annotées G_1 et G_2 de l'image, G_1 étant une partition de l'image où chaque région est annotée avec des concepts de $C1(S_\Omega^{po})$, et G_2 avec des concepts de $C2(S_\Omega^{po})$. Etant donné une image I, on souhaite trouver les partitions annotées G_1 et G_2 maximisant la probabilité

$$max_{G_1,G_2} P(I, G_1, G_2|M)$$

.

Les modèles associés aux concepts de la couche 2 s'exprimant en fonction des modèles de la couche 1 (cf 4.4.2), on décompose ce terme comme le produit des deux vraisemblances :

$$max_{G_1,G_2} P(G_1, I|M)P(G_2, I|G_1, M)$$

.

L'ensemble des configurations G_1 et G_2 étant beaucoup trop grand, on pratique deux optimisations locales en déterminant tout d'abord $G_{1,opt}$ maximisant $P(G_1, I|M)$. Puis, on optimise le terme $P(G_2, I|G_{1,opt}, M)$.
L'image est tout d'abord annotée avec les concepts de la première couche selon l'algorithme décrit dans la section 5.1.2. Ensuite, l'image est annotée avec les concepts de la deuxième couche avec un algorithme dont le principe reste exactement le même que celui détaillé en section 5.1.2. Dans le cas où S_Ω^{po} ne comporte qu'une seule couche, seule la première étape d'inférence explicitée ci-après est nécessaire.

Première étape d'inférence

– Initialisation de l'algorithme :
  Une partition annotée initiale est créée en utilisant les modèles M_c associés aux modèles $c \in C_1(S_\Omega^{po})$ de la manière suivante :
  Pour chaque texton de l'image de coordonnées (k, l), l'histogramme de taille n_0 $U(k, l)$ est calculé :

$$U(k,l) = \sum_{(i,j)\in I} E_{I(k,l)} g_{k,l,\sigma}(i,j)$$

  .

où E_i correspond au $i$ème vecteur de base, $g_{m_1,m_2,\sigma}(x,y)$ est la fonction gaussienne 2D de moyenne (m_1, m_2) et de variance σ^2. σ est un paramètre de l'algorithme. Le vecteur $U(k,l)$ donne une caractérisation du voisinage autour du texton (k,l).
Ainsi, pour $i \in \{1,\ldots,n_1\}$, les probabilités suivantes sont calculées :

$$P(U_(k,l)|\theta_i) = \sum_{j=1}^{n_0} p_{ij}^{U_{(k,l)}}(j)$$

.
Ensuite, le texton (k,l) est annoté par le label c vérifiant

$$P(U_{(k,l)}|\theta_c) = \min_{i\in\{1,\ldots,n_1\}} P(U_{(k,l)}|\theta_i)$$

Une partition annotée G_1^0 est ensuite créée en construisant une région annotée par le concept c pour chaque zone 4-connexe de textons qui sont reliés au concept c durant l'étape précédente.
Soit i le nombre d'itérations effectuées dans la boucle, tant que le nombre de régions contenues dans G_1^i est supérieur à 1 :
– Pour toutes les paires de régions adjacentes :
  – On fusionne les deux régions adjacentes. Pour les n_1 annotations possibles de cette nouvelle région, on calcule la vraisemblance de la partition annotée qui en résulte.
– La configuration maximisant la vraisemblance est conservée et notée G_1^i
La partition annotée finale G_1^{opt} est la configuration vérifiant :

$$P(I, G_1^{opt}|M) = \max_i P(I, G_1^i|M)$$

.
À chaque passage dans la boucle, deux régions sont fusionnées. Par conséquent, l'algorithme termine en un nombre fini d'itérations. Plus σ est grand, plus le nombre de régions présentes dans G_1^0 est faible et plus l'algorithme termine rapidement.
La deuxième étape d'inférence est similaire à la première. Un chemin est exploré dans l'espace des configurations possibles en créant une partition initiale G_2^0 et en fusionnant itérativement des régions jusqu'à obtenir seulement une seule région dans l'image.

Deuxième étape d'inférence
– Initialisation de l'algorithme :
  Une partition annotée initiale est créée en utilisant les modèles M_c associés aux modèles $c \in C_2(S_\Omega^{po})$ de la manière suivante :
  Pour chaque région R_k de G_1^0, l'histogramme suivant de taille n_1 est calculé

$$U(R_k) = \sum_{j/adj(R_j,R_k)} E_{c(R_j)}$$

.

Ensuite, pour $i \in \{1, \ldots, n_2\}$, les probabilités suivantes sont calculées :

$$P(U(R_k)|\theta_i) = \sum_{j=1}^{n_1} p_{kj}^{U_{(R_k)}}(j)$$

La région R_k est alors annotée par le concept c vérifiant

$$P(R_k|\theta_c) = \min_{i \in \{1,\ldots,n_2\}} P(R_k|\theta_i)$$

.

Une partition G_2^0 est ensuite créée en construisant une région annotée par le concept c pour chaque zone 4-connexe de textons qui ont été associés au concept c durant l'étape précédente.

Soit i le nombre d'itérations effectuées dans la boucle. Tant que le nombre de régions de la partition est supérieur à 1 :

- Pour toutes les paires de régions adjacentes :
  - On fusionne les deux régions adjacentes. Les n_2 annotations possibles de cette nouvelle région sont considérées et pour chaque cas la vraisemblance de la partition annotée qui en résulte est calculée.
- La partition annotée maximisant la vraisemblance pour toutes les partitions envisagées dans la boucle est conservée et notée G_2^i

La partition annotée finale est notée G_2^{opt} et est celle qui vérifie :

$$P(G_2^{opt}|G_1^{opt}) = \max_i P(G_2^i|G_1^{opt})$$

.

À chaque passage dans la boucle, deux régions sont fusionnées. Par conséquent, l'algorithme termine en un nombre fini d'itérations.

5.1.3 Représentation sémantique de l'image

On suppose ici que S_Ω^{po} comporte deux couches. L'algorithme détaillé précédement fournit alors deux partitions annotées de l'image de test. On place toutes ces régions dans un seul ensemble de régions que l'on notera : $P_{po} = \{G_1, G_2\}$. Cet ensemble forme la base de ce qui sera la représentation sémantique de l'image. Cependant, cette représentation est enrichie de la manière suivante :

- Pour toute région R appartenant à P_{po}, si le concept $c(R)$ est relié par une relation d'hyperonymie à un concept c' appartenant à $C2(S_{ko})$, on crée une nouvelle région dont la localisation coincide avec R et dont l'annotation est c'.
- Pour toute paire R et R' appartenant à P_{ko} et étant annotées par le même concept, si R et R' sont adjacentes ou si leur intersection est non vide, ces régions sont fusionnées en une seule.
- Pour toute région R appartenant à l'ensemble $P = P_{ko} \cup P_{po}$, on ajoute à l'ensemble des annotations de la région R, qui consiste pour l'instant en un singleton $c(R)$ tous les concepts synonymes de $c(R)$.

Ainsi, on rajoute aux partitions annotées avec les concepts du réseau sémantique S_Ω^{po} un autre ensemble de régions qui ne définit pas nécessairement une partition et qui contient les concepts du réseau sémantique S_Ω^{ko}. En effet, la relation sémantique "A "kind-of" B" correspond à une implication : si la zone est annotée par le concept A, alors est elle aussi annotée par le concept B. Ainsi, toute région annotée par A l'est aussi par B. Ensuite, les régions annotées par le même concept sont fusionnées pour obtenir des régions 4-connexes.

Pour compléter cette représentation de l'interprétation de l'image, on rajoute des relations spatiales entre toutes les paires de régions de l'ensemble P. Les relations que l'on retient entre deux régions sont : "adjacentes", "disjointes", "entourée par", "entoure", "se chevauchent", "envahit", "est envahi par".

Étant donné deux régions R_i et R_j, afin d'assigner la relation spatiale de R_i par rapport à R_j, on calcule les valeurs suivantes :

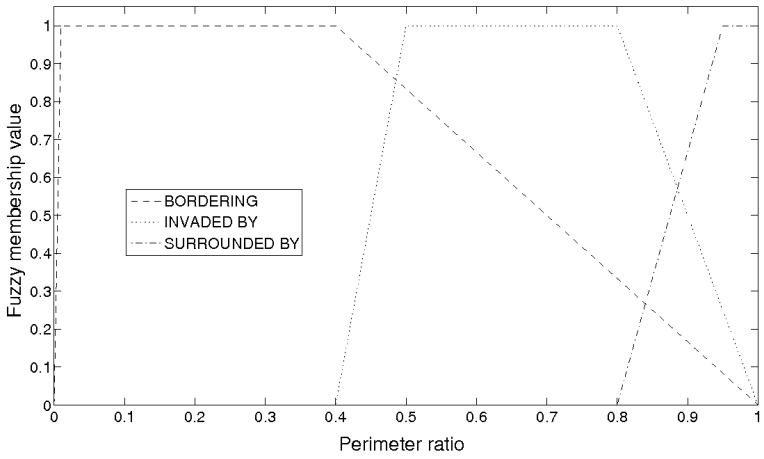
- Périmètre de R_i : π_i.
- Périmètre de R_j : π_j.
- Périmètre commun entre les régions : π_{ij}
- Rapport du périmètre commun au périmètre de la première région : $r_{ij}^1 = \pi_{ij}/\pi_i$
- Surface de R_i : σ_i.
- Surface de R_j : σ_j.
- Surface commune entre les régions σ_{ij}
- Rapport de la surface commune et de la surface de R_i : $r_{ij}^2 = \sigma_{ij}/\sigma_i$

Les relations spatiales entre R_i et R_j se définissent alors par une représentation floue selon les fonctions d'appartenance définies en 5.2. De plus, chaque région est caractérisée par les coordonnées de son centre de gravité, ses moments d'inertie, et l'orientation principale de l'ellipsoïde d'inertie.

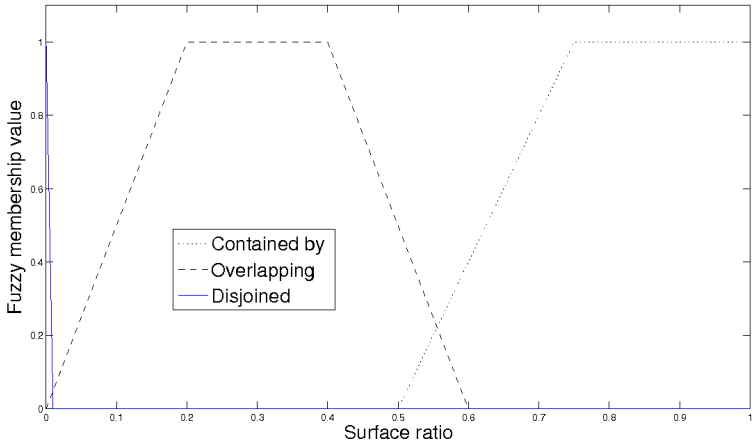
On obtient un ensemble des régions $P^{tot} = P_{po} \cup P_{ko}$ et une matrice de relations entre ces régions. Ainsi, après ces traitements, P^{tot} peut être transformé en un graphe G^{tot} dont les nœuds correspondent aux éléments de P^{tot} et dont les arcs sont étiquetés par les relations spatiales énoncées précédemment.

5.1.4 Test d'auto-cohérence du système d'annotation

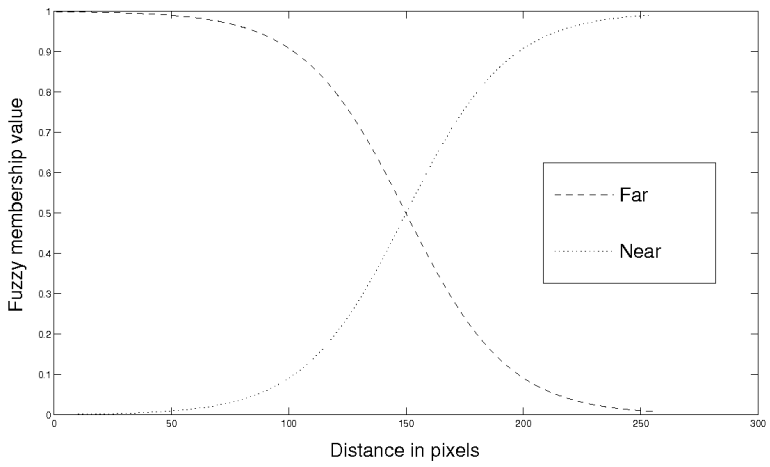
Nous testons ici la capacité du système, supposée élémentaire, à retrouver dans une image de test des imagettes d'apprentissage qui sont extraites dans cette même image. Nous avons ainsi extrait 7 imagettes dans une image de Marseille de taille 3000×3000, 3 images de carrière, 1 imagette de montagne, 1 imagette de zone résidentielle, 1 imagette de zone rurale, 1 imagette de mer. Ces imagettes sont utilisées à elles seules comme apprentissage pour estimer les paramètres du modèle bayésien naïf de chacun de ces 5 concepts. Le résultat est présenté figure 5.3. On observe que les différentes régions sont retrouvées tout à fait correctement dans l'image.



(a)



(b)



(c)

Fig. 5.2 – Relations spatiales floues (a) Relations liées au périmétre (b) Relations liées à la surface (c) Relations liées à la distance.

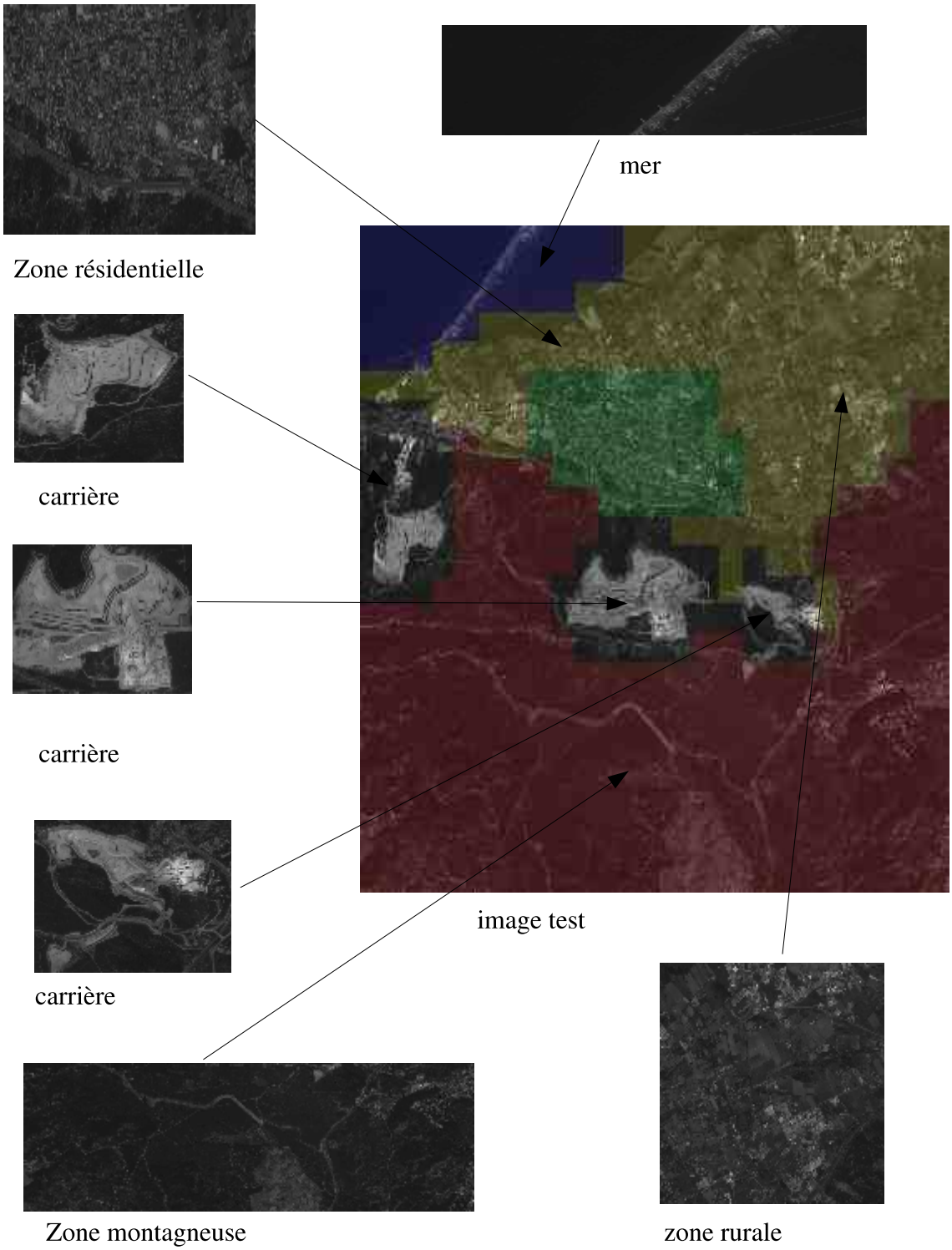


Fig. 5.3 – Segmentation de type "autocohérence" : 7 imagettes sont extraites d'une image 3000 × 3000. Les zones correspondantes sont correctement retrouvées dans l'image d'origine.

5.2 Évaluation quantitative des performances d'annotations

Le nombre important de travaux d'indexation et d'annotation d'images qui ont été réalisés au cours de la dernière décennie atteste de l'importance que représente ce domaine de recherche. Face à la multitude de méthodes qui ont été proposées, la communauté scientifique du traitement d'images a pris conscience de la nécessité d'évaluations quantitatives rigoureuses des résultats de ces méthodes. Trop souvent, les auteurs se contentaient d'illustrer leurs publications à partir de quelques résultats produits sur certaines images fréquemment utilisées ("Lena"), ou alors en prenant comme exemple des images synthétiques qui mettent en évidence les points forts de leurs méthodes. La comparaison avec d'autres algorithmes était souvent limitée à un nombre réduit d'exemples sur lesquels les paramètres du modèle ont souvent été soigneusement ajustés de façon à ce que le résultat apparaisse satisfaisant. On constate, dans les publications récentes, une attention plus grande portée à des évaluations quantitatives effectuées sur des bases de données annotées communes consacrées à des applications précises. Cette évolution est semblable à celle qui s'est produite dans le domaine du traitement de la parole. Émanant d'une prise de conscience de l'importance du processus d'évaluation, des campagnes ont en effet été lancées pour évaluer les algorithmes de traitement automatique du langage naturel en mettant à disposition des acteurs des corpus de grande taille et des métriques d'évaluation fiables. Un tel exemple de campagne en France est la campagne ESTER (Evaluation des systèmes de Transcription des Émission Radiophoniques) [49]. Toutefois, mettre en œuvre une évaluation quantitative rigoureuse d'un algorithme d'annotation d'images est une tâche qui s'avère délicate et qui est très coûteuse, et il est de manière générale très difficile d'évaluer de façon relative des algorithmes d'annotation sémantique selon un protocole expérimental commun. Cependant, des progrès significatifs ont été accomplis dans ce sens. Ainsi, pour la tâche d'annotation d'images par apprentissage supervisé, un certain nombre de groupes de recherches ont adopté le protocole Corel5K [37], [40], [79]. Ce protocole est basé sur la base d'image Corel qui comprend 5000 images et qui est séparé en un lot d'apprentissage de 4000 images, un lot de validation de 500 images, et un lot de test de 500 images. Les paramètres initiaux du modèle sont estimés à partir du lot d'apprentissage, les paramètres nécessitant une validation croisée sont ensuite optimisés sur le lot de validation, après quoi ce lot est fusionné avec le lot d'apprentissage pour construire un nouveau lot d'apprentissage. Chaque image comporte une légende d'une à cinq annotations parmi un vocabulaire de 371 mots. Les performances de l'annotation effectuée par le système sur les images du lot de test sont ensuite comparées avec l'annotation humaine qui constitue la "vérité terrain".

5.2.1 Métrique considérée

Pour évaluer la qualité de l'annotation, les annotations par les concepts de S_{ko} ne sont pas pris en compte car ils découlent naturellement de la première annotation avec les concepts de S_{po}. L'annotation d'une image test consiste à produire pour chaque couche j du modèle une partition annotée de cette image, à savoir un ensemble $\{R_1^j, R_2^j, ..., R_{m_j}^j\}$ où chaque région R_i^j est annotée par le concept $c(R_i^j)$, où $c(R_i^j)$ appartient à l'ensemble des concepts situés dans la couche j du modèle. Pour évaluer la qualité de cette annotation multi-couche, on souhaite la comparer à une "vérité terrain" qui consiste, pour chaque couche j, en un ensemble de régions $\{R_1^{jr}, R_2^{jr}, ..., R_{m_j^r}^{jr}\}$ où chaque région R_i^{jr} est annotée par le concept $c(R_i^{jr})$. On suppose que, pour chaque couche, l'ensemble des concepts d'annotation employés pour la vérité terrain est le même que celui employé par le système. Ainsi, on suppose que la hiérarchie des concepts apprise par le système correspond à celle utilisée pour effectuer la vérité terrain. Pour évaluer quantitativement l'annotation produite par le système, nous comparons individuellement les segmentations produites pour chaque couche.
Cependant, cette distance n'apparie pas toutes les régions, et ne prend donc pas en compte toute l'information. En effet, les régions seront appariées si elles ont un ensemble commun de textons important, ce qui a tendance a donner beaucoup d'importance aux grandes régions.
Pour évaluer les résultats de segmentation d'une couche donnée, deux approches peuvent être utilisées. La première consiste à voir le problème de l'annotation comme un problème de pure segmentation de l'image et des outils de comparaisons de segmentation avec une vérité terrain peuvent être employés comme la distance de Vinet (voir [23]). La deuxième approche consiste à utiliser des outils de traitement du langage naturel. En effet, l'objectif principal est de mettre en relation les images avec un vocabulaire qui est celui du langage naturel. Dans l'exemple d'une application de transcription de la parole, la sortie du système est une séquence de mots qui est reliée par programmation dynamique à la séquence de mots qui constitue la *vérité terrain*. La métrique qui est utilisée est la métrique dite "Word Error Rate" et qui s'exprime selon la formule :

$$WER = \frac{Elisions + Ajouts + Substitutions}{NombreTotalDeMots} \tag{5.5}$$

.
Dans notre cas, cette métrique peut être aisément adaptée en posant qu'une région du système R_1 et une région de la vérité terrain R_2 peuvent être appariées si elles se recouvrent de manière significative. Plus précisement, nous fixons la condition d'appariement de la façon suivante :

$$\frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} > 0.8 \tag{5.6}$$

.
où $|R|$ correspond au nombre de textons de la région.

En référence à la métrique WER, on définit alors une métrique que l'on appellera IAWER (Image Adapted World Error Rate) :
- Une élision est une région de la vérité terrain qui n'est pas appariée à une région de la partition produite par le système.
- Un ajout est une région de la partition produite par le système qui n'est pas appariée à une région de la vérité terrain.
- Une substitution est une région de la partition produite par le système qui est appariée à une région de la vérité terrain mais qui n'est pas annotée par le même concept.

La métrique IAWER est alors calculée selon la formule 5.5 avec la définition de l'élision, de l'addition, et de la substitution définies précédemment et selon les conditions d'appariement définies ci-dessus.
Pour chaque couche, l'appariement des région est réalisé de manière à minimiser la quantité IAWER. On effectue l'appariement de manière gloutonne en appariant les régions annotées par un même concept et correspondant à un taux de recouvrement optimal.
Ainsi, pour une couche de niveau j, l'appariement est effectué de la manière suivante : Pour i variant de 1 à m_j : pour chaque région de la vérité terrain annotée avec le même concept que R_i^j, on calcule leur taux de recouvrement avec R_i^j. Si aucune région ne vérifie la condition d'appariement (voir 5.6), R_i^j n'est pas appariée. Dans le cas contraire, soit $R_{i,opt}^{jr}$ la région de la vérité terrain correspondant au taux de recouvrement maximal. Si $R_{i,opt}^{jr}$ n'est pas encore appariée à une autre région, R_i^j est appariée avec $R_{i,opt}^{jr}$. Dans le cas contraire, à savoir si $R_{i,opt}^{jr}$ est déjà appariée avec une région R_k^j, on apparie $R_{i,opt}^{jr}$ avec celle des deux régions qui correspond au taux de recouvrement maximal. La région correspondant au taux de recouvrement minimal ne sera alors plus appariée à aucune autre région.

5.2.2 Expériences

Des expériences ont été menées pour mesurer les performances d'annotation de la méthode ASP. L'objectif principal est d'évaluer l'apport d'une modélisation multi-couches et d'un franchissement en plusieurs étapes du fossé sémantique pour des labels correspondant à des zones complexes. Aucune base de données d'images satellitaires disposant d'annotations adaptées (plusieurs niveaux d'annotations par complexité croissante) n'étant disponible, des annotations ont été faites manuellement à partir de bases d'images SPOT5 (2,5m de résolution) et d'images Quickbird (0,6m de résolution). Pour chaque base, un réseau sémantique à deux couches est donc fixé et les images sont segmentées deux fois, les régions de chaque segmentation étant annotées avec les labels de chaque couche respective. Les performances de la méthode ASP sont comparés à une méthode markovienne avec estimateur MPM où chacune des deux annotations est effectuée séparément (voir annexe B). Cette méthode concurrente a été choisie pour sa simplicité de mise en oeuvre.

5.2.2.1 Base de données SPOT5

Une base d'images SPOT5 de Paris, Marseille, Nice, et Angers à 2,5m de résolution annotées manuellement ont été utilisées pour effectuer une évaluation quantitative des performances d'annotation sémantique. Cette base de données est constituée de 42 images de taille 3000×3000 à 6000×6000 correspondant à différents types de paysages. Les labels utilisés pour l'annotation sont ceux listés en 4.20 et sont placés dans un réseau sémantique à deux couches qui est celui obtenu en 4.7.2.4. À chaque image de la base sont donc associées deux partitions annotées, correspondant à chaque couche de labels, et codées chacune comme un masque de l'image.

Apprentissage 15 images ont été sélectionnées pour l'apprentissage des modèles de manière à constituer un échantillon représentatif des différents types de zones présentes dans la base de données pour chaque label. La structure du réseau sémantique étant fixée à priori, seuls les paramètres du modèle sont à estimer en utilisant les formules 4.29 et 4.30. Les paramètres du modèle markovien sont appris sur cette base d'apprentissage en utilisant la méthode du gradient stochastique (voir annexe B section B.2.1).

Résultats Les images sont annotées en utilisant la méthode décrite en 5.1. Pour la méthode markovienne, les images sont annotées en utilisant la méthode MPM. Les performances d'annotation sont évaluées en utilisant deux critères : le critère de Vinet et la métrique adaptée du *Word Error Rate* présentée en section 5.2.1 avec un coefficient de regroupement fixé à 0,8 (voir 5.6) pour permettre un appariement entre régions. Le critère de Vinet évalue la sortie du système comme un résultat de segmentation, tandis que la métrique IAWER évalue la sortie du système comme un résultat d'annotation. Notons que le critère de Vinet évalue une segmentation comme étant bonne lorsqu'il est élevé (une segmentation étant parfaite lorsque le critère de Vinet est de 100%). Au contraire, la métrique IAWER correspond à un taux d'erreur, et une bonne segmentation doit donc correspondre à une valeur aussi faible que possible de cette métrique.
Comme on peut voir, les deux algorithmes fournissent des performances relativement similaires dans le cas de la première couche dans le cas de la mesure de Vinet (voir tableau 5.4 et l'image 5.7). En effet, l'algorithme MPM est basé sur une maximisation sur le nombre de sites correctement annotés. Par contre, les performances d'annotation sont nettement moins bonnes pour la modélisation markovienne selon le critère IAWER (voir tableau 5.5), ce qui peut s'interpréter par le fait que l'estimateur MPM effectue une optimisation au niveau des sites individuels qui tend à entrainer la création d'un certain nombre de petites régions ne correspondant pas à des régions sémantiques identifiables. Au contraire, pour la méthode ASP, la méthode bayésienne naïve utilisée par les modèles de la couche de plus bas-niveau comporte une force de rappel assez forte pour créer les régions à travers la loi de Poisson.
Les résultats sont nettement plus contrastés en ce qui concerne les résultats du

	couche 1	couche 2
Modèle ASP	83,27 %	86,14%
Modélisation markovienne	84,26%	68,27%

Fig. 5.4 – Résultats d'annotation estimés avec la mesure de Vinet sur la base SPOT5

	Élisions	Additions	Substitutions	IAWER
Modèle ASP	15	63	5	9,79 %
Modélisation hiérarchique	31	29	16	8,91 %

Fig. 5.5 – Résultats d'annotation estimés avec la métrique IAWER pour la première couche d'annotation sur la base SPOT5

processus d'annotation avec les concepts appartenant à la deuxième couche (voir tableau 5.6), l'annotation markovienne fournissant des résultats nettement moins bons. On constate également une nette dégradation des résultats de l'annotation markovienne par rapport à l'étape précédente. Cela peut s'interpréter par le fait qu'il existe un fossé sémantique trop important entre les caractéristiques de bas-niveau et ces concepts pour pouvoir être franchis en une seule étape d'inférence. Les résultats restent cependant stables en ce qui concerne notre méthode, ce qui tend à prouver la pertinence d'effectuer l'inférence de ce type de concepts à partir de régions contenant déja un certain niveau de sémantique. L'annotation ainsi produite sur les deux niveaux est montrée et comparée à la vérité terrain en 5.8. Ainsi, ces expériences tendent à prouver que l'inférence en plusieurs étapes effectuée par la méthode ASP permet de décrire des images en utilisant des labels complexes avec une qualité supérieure à celle que l'on peut obtenir par un mécanisme d'inférence directe. Ces résultats doivent cependant être confirmés sur des bases de données de taille plus importante.

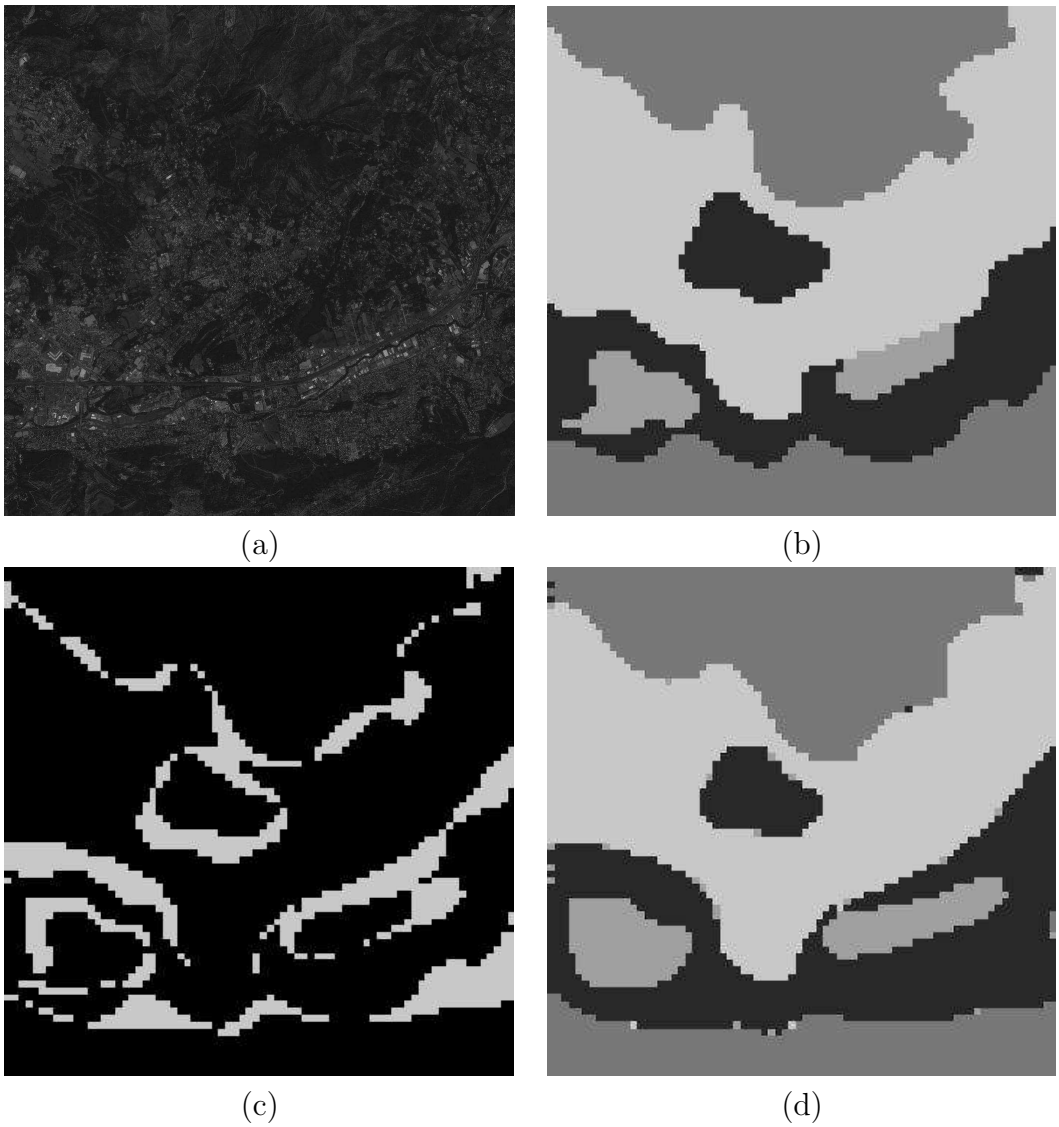
5.2.2.2 Base de données Quickbird

Des images Quickbird de Beijing à 0,6m de résolution ont été annotées manuellement grace à l'aide du BISM de Pékin (Beijing Institute of Survey and Mapping). Les concepts utilisés pour l'annotation sont ceux listés en 4.23 et sont placés dans un réseau sémantique de type "part-of" (voir Section 4.2.1).

Apprentissage Le protocole d'apprentissage est identique à celui employé sur la base de données d'images SPOT5 (voir section 5.2.2.1).

	Élisions	Additions	Substitutions	IAWER
Modèle ASP	11	24	3	23.14 %
Modélisation hiérarchique	9	8	3	12,34 %

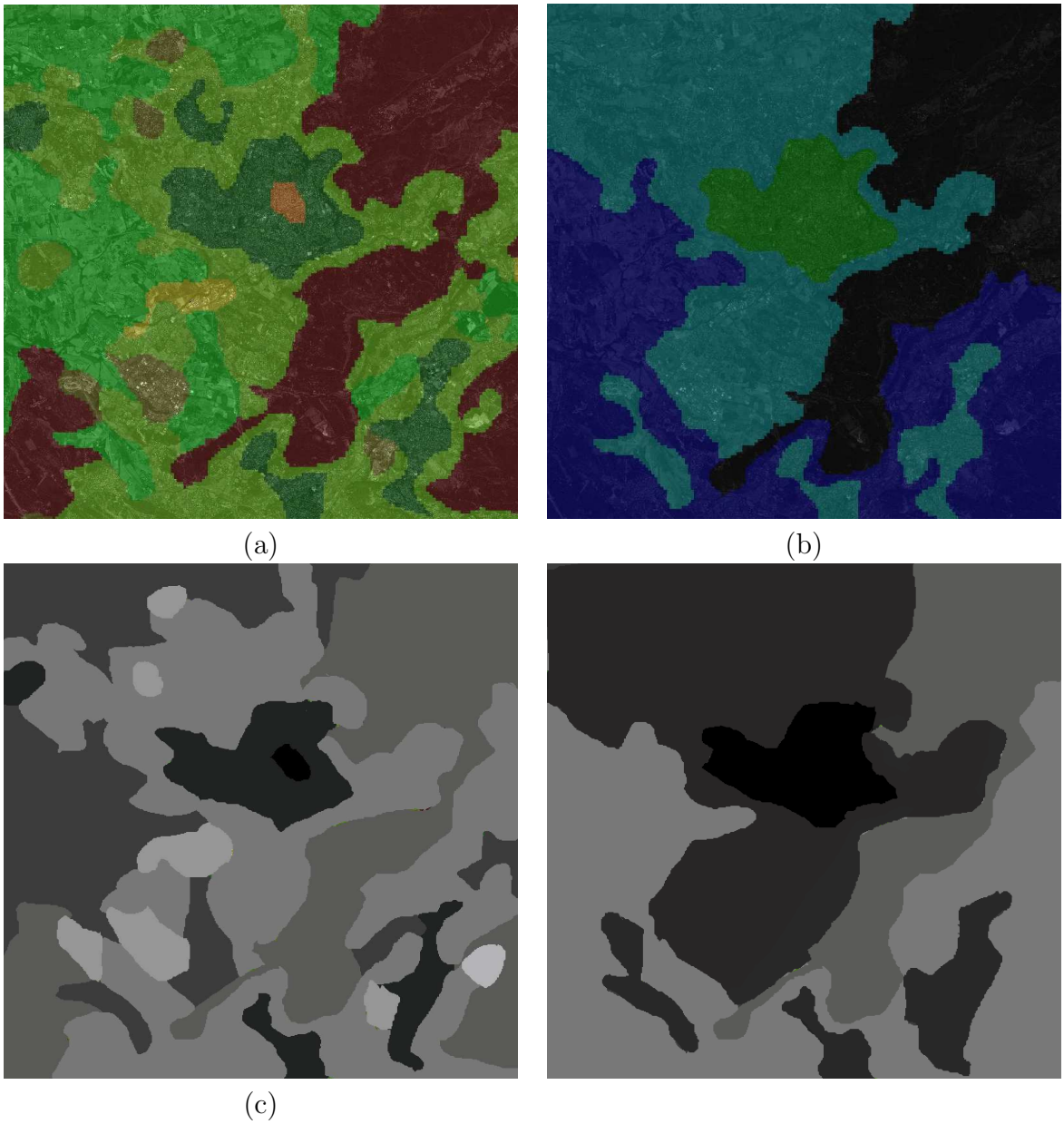
Fig. 5.6 – Résultats d'annotation estimés avec la métrique IAWER pour la deuxième couche d'annotation sur la base SPOT5



(a) (b)

(c) (d)

Fig. 5.7 – (a) Image d'origine SPOT5 de Marseille ©CNES (b) Masque de la vérité terrain (c) Pixels mal annotés (d) Masque de l'annotation du système



(a) (b)

(c)

Fig. 5.8 – Image SPOT5 à 2,5m de résolution 6000 × 6000 de la région d'Aix en Provence. (a) Première couche d'annotation (b) Deuxième couche d'annotation (c) Vérité terrain correspondant à la couche 1 du réseau. (d) Vérité terrain correspondant à la couche 2 du réseau

	couche 1	couche 2
Méthode ASP	72,78%	76,34%
Modélisation markovienne	74,68%	61,59%

Fig. 5.9 – Résultats d'annotation estimés avec la mesure de Vinet sur la base Quickbird

	Élisions	Additions	Substitutions	IAWER
Méthode ASP	20	25	6	21,39 %
Modélisation markovienne	14	40	7	25,02 %

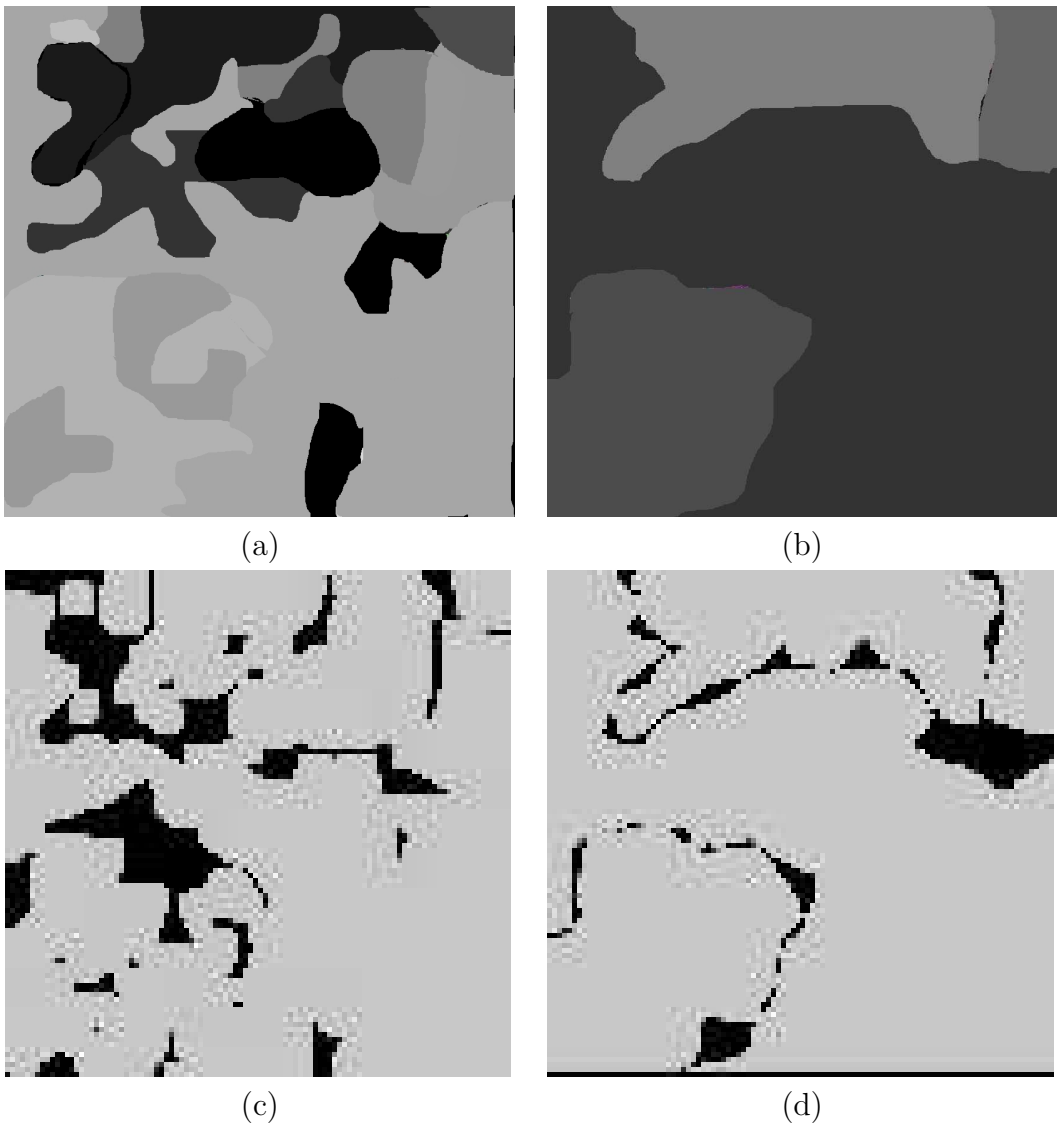
Fig. 5.10 – Résultats d'annotation estimés avec la métrique IAWER pour la première couche d'annotation sur la base Quickbird

Résultats Comme pour la base de données d'images SPOT5, les images sont annotées en utilisant la méthode décrite en 5.1. Pour la méthode markovienne, les images sont annotées en utilisant la méthode MPM. Les performances d'annotation sont évaluées en utilisant deux critères : la critère de Vinet et la métrique adaptée du *Word Error Rate* présentée en section 5.2.1. Les résultats quantitatifs sont présentés tableaux 5.9, 5.10 et 5.11. Des résultats visuels sont montrés figures 5.12, 5.13.

Des conclusions relativement analogues à celles tirées dans la section 5.2.2.1 peuvent être faites. Cependant, on constate une diminution globale des performances. Celle-ci provient de la plus grande complexité des données, du plus grand nombre de concepts d'annotation et probablement du moindre pouvoir de description des caractéristiques de bas-niveau qui sont employées. Une différence notable par rapport aux résultats obtenus sur la base SPOT5 est que les résultats sont meilleurs pour la couche 2 que pour la couche 1, tant pour la mesure de Vinet que pour la métrique IAWER. Les performances d'annotation de la deuxième couche avec le modèle hiérarchique sont presque deux fois meilleures que celles obtenues avec la modélisation markovienne. Comme pour la base d'images SPOT5, ces résultats nous confortent dans l'idée d'utiliser une inférence basée sur des labels sémantiques pour franchir le fossé sémantique. Cependant, ces résultats doivent être confirmés sur des bases de données de taille plus importante.

	Élisions	Additions	Substitutions	IAWER
Méthode ASP	9	8	0	20,99 %
Modélisation markovienne	13	15	2	37,04 %

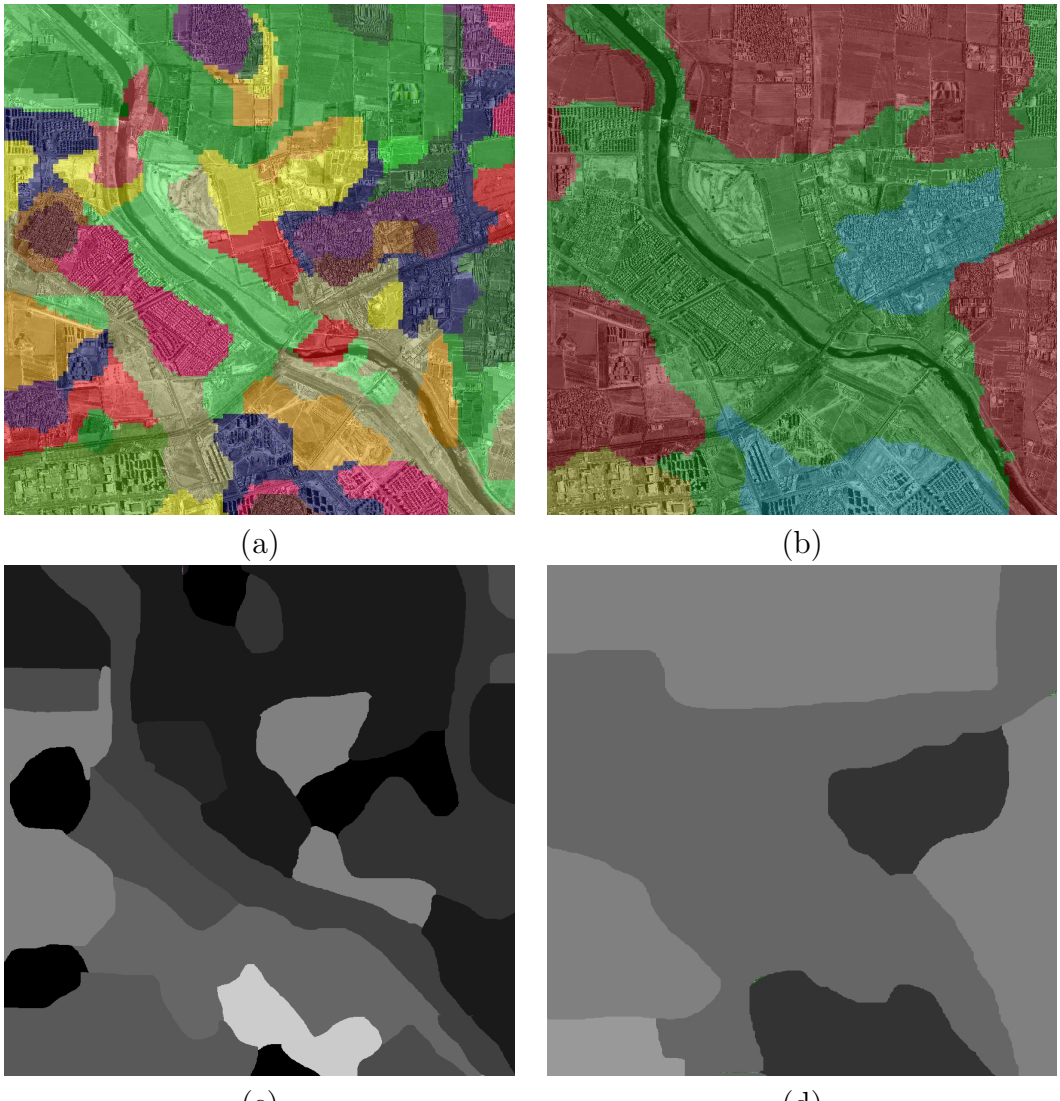
Fig. 5.11 – Résultats d'annotation estimés avec la métrique IAWER pour la deuxième couche d'annotation sur la base Quickbird



(a) (b)

(c) (d)

Fig. 5.12 – Résultats d'annotation sur une image Quickbird de Pékin (a) Masque de la érité terrain, couche 1 (b) Masque de la vérité terrain, couche 2 (c) Pixels mal annotés de la couche 1 (d) Pixels mal annotés de la couche 2. ©LIAMA



(a) (b)

(c) (d)

Fig. 5.13 – (a) Annotation produite par le système pour la couche 1 (b) Annotation produite par le système pour la couche 2 (c) Masque de la vérité terrain, couche 1 (b) Masque de la vérité terrain, couche 2. Pour l'image (b), Le rouge correspond aux zones rurales, le jaune aux complexes industrielles, le vert aux zones pavillonaires, le bleu foncé aux jardins publics, le bleu clair aux zones urbaines éparses. ©LIAMA

5.3 Utilisation des annotations pour la recherche d'images par le contenu

Nous traitons ici une méthode permettant d'exploiter les annotations fournies par le système afin de formuler des requêtes sémantiques dans une base d'images. Les requêtes définissent un ensemble de régions annotées par certains labels et dans une configuration spatiale particulière (exemple : zone industrielle en bordure d'une zone rurale et à proximité d'un lac). On définit alors une fonction de cohérence qui permet de séléctionner les meilleures hypothèses parmi un ensemble de groupes de régions annotées en réponse à cette requête.

5.3.1 Fonction de cohérence

Étant donné un graphe de N régions correspondant à l'annotation d'une zone d'une image de la base de données, on mesure l'adéquation entre une partition annotée P_s et une requête Req formulée par un utilisateur en utilisant la fonction de cohérence définie dans [89] :

$$C_{Req}(P_s) = \sum_{i,j \in \{1,\ldots,N\}} \sum_{\mathbf{R}} (C)_{\mathbf{R}}(R_i, R_j)$$

.

où :

$$(C)_{\mathbf{R}}(R_i, R_j) = P(R_i)P(R_j)F(R_i\mathbf{R}R_j)Eval_{Req}(R_i, \mathbf{R}, R_j)$$

.

- $P(R_i)$ est un indice de confiance de l'annotation de la région R_i par le concept $c(R_i)$ avec le modèle qui a été utilisé pour évaluer la partition. Ce modèle étant exprimé ici avec un modèle bayésien naïf, la probabilité d'annotation décroit avec le nombre de textons selon une suite géométrique et il n'est donc pas possible d'utiliser directement cette probabilité comme indice de confiance, sous peine de fortement pénaliser les grandes régions. Nous procédons ainsi à une normalisation par la quantité $p_c^{|R_i|}$ où p_c est définie par la formule suivante : $p_c = (P(X_{c(R_i)}))^{\frac{1}{|X_{c(R_i)}|}}$, où $|X_c|$ correspond au nombre de textons dans la base d'apprentissage du concept c.
- $\mathbf{R}$ est une relation spatiale entre deux objets parmi les relations qui ont été listées en 5.1.2.
- $F(R_i\mathbf{R}R_j)$ est la fonction d'appartenance de cette relation et est employée pour mesurer le degré avec lequel la relation la relation spatiale $\mathbf{R}$ entre R_i et R_j est vérifiée.
- $Eval(R_i, \mathbf{R}, R_j)$ est une fonction évaluant si la relation entre les deux régions est conforme à la requête. Elle vaut 0 si la relation $(R_i, \mathbf{R}, R_j)$ n'est pas mentionnée dans la requête et 1 si elle l'est.

Ainsi, si une relation $\mathbf{R}$ entre deux régions R_1 et R_2 est mentionnée dans la requête de l'utilisateur ("Zone industrielle à proximité de la montagne"), $Eval(R_1, \mathbf{R}, R_2) = 1$.

5.3.2 Recherche d'images par le contenu

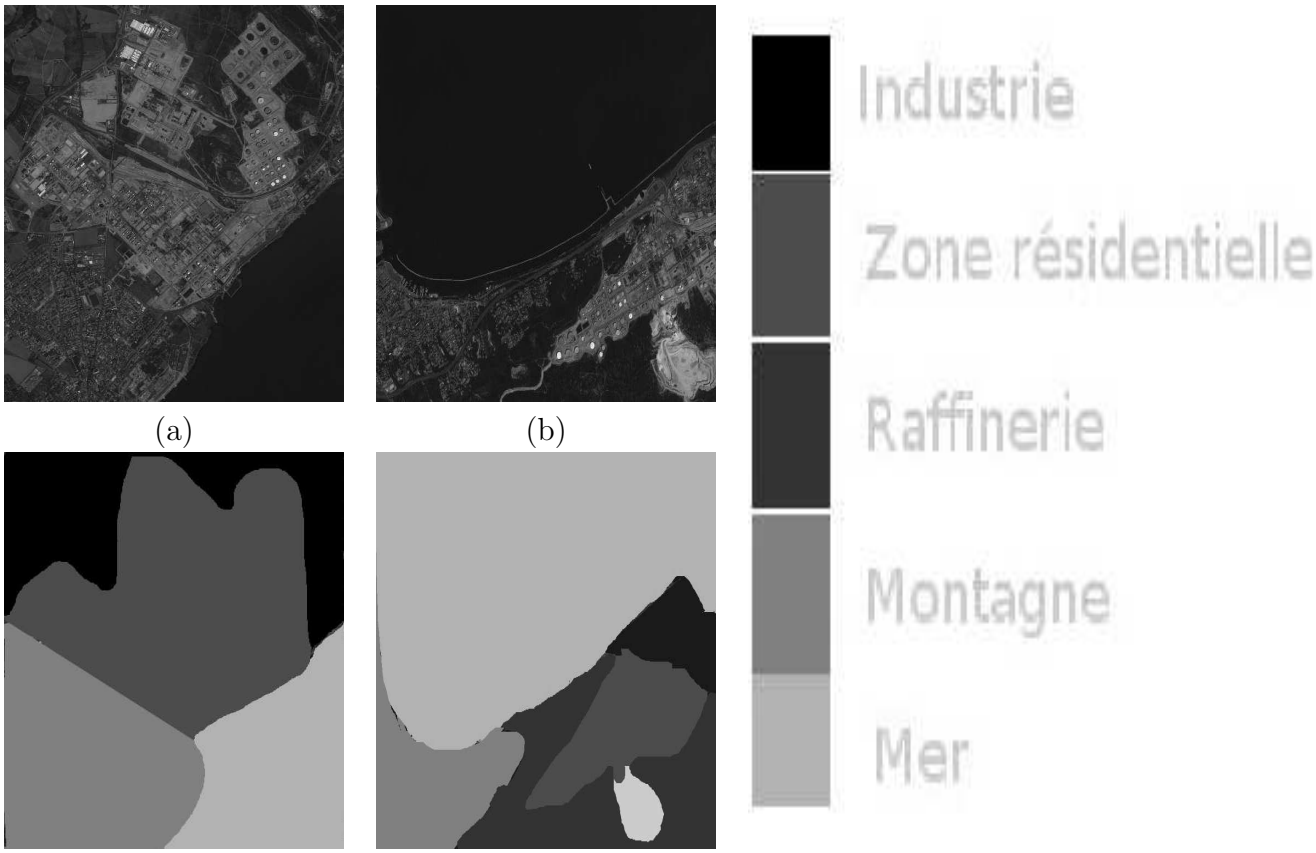
Etant donné une requête, la recherche des groupes de régions maximisant la fonction de cohérence est effectuée de façon combinatoire. Les images étant découpées en sous-images 6000×6000, le nombre de régions qui y sont présentes est de l'ordre d'une cinquantaine et le nombre de régions impliquées par les requêtes que l'on considère varie de 2 à 4 au maximum. Ainsi, la recherche exhaustive ne pose pas de difficultés dans notre cas, mais pour des requêtes impliquant davantage de régions et où le nombre de configurations à considérer serait trop important, il peut devenir nécessaire d'employer des algorithmes d'optimisation plus sophistiqués (comme le recuit simulé). Les résultats sont visuellement satisfaisants, cependant des résultats quantitatifs seraient nécessaires pour une évaluation rigoureuse des performances (voir 5.14 et 5.15).

5.4 Couverture sémantique d'une base d'images

Nous souhaitons donner une caractérisation de la *couverture sémantique* d'un vocabulaire d'annotations par analogie avec la couverture d'un corpus par un vocabulaire dans le domaine du traitement du langage naturel. Nous exprimons cette mesure d'un ensemble d'images D par un vocabulaire Ω de taille n en utilisant l'inverse de la log-vraisemblance de la base de données $-\log P(D|M_\Omega)$, où M_Ω correspond au modèle stochastique estimé sur une base d'apprentissage X_Ω. Afin d'évaluer l'évolution de la couverture sémantique en fonction de la taille du vocabulaire, on estime pour $i \in \{1, \ldots, n\}$, la couverture sémantique SC_i de D en utilisant un sous-ensemble ω_i de Ω de taille i.

- Initialisation : $\omega_0 = \emptyset$
- For $i \in \{1, ..., n-1\}$
  - $\forall c_j \in \Omega - \omega_i$, on définit $\omega'_{i+1,j} = \omega_i \cup c_j$. Le modèle $M_{\omega'_{i+1}}$ est appris avec ce sous-ensemble de vocabulaire sur $X_{\omega'_{i+1}} = \cup_{c \in \omega'_{i+1,j}} X_i$, l'information de Shannon est alors calculée : $-logP(D|M_{\omega'_{i+1,j}})$
  - On définit $SC_i = -logP(D|M_{\omega_{i+1}}) = \min_{\omega'_{i+1,j}}(-logP(D|M_{\omega'_{i+1,j}})$

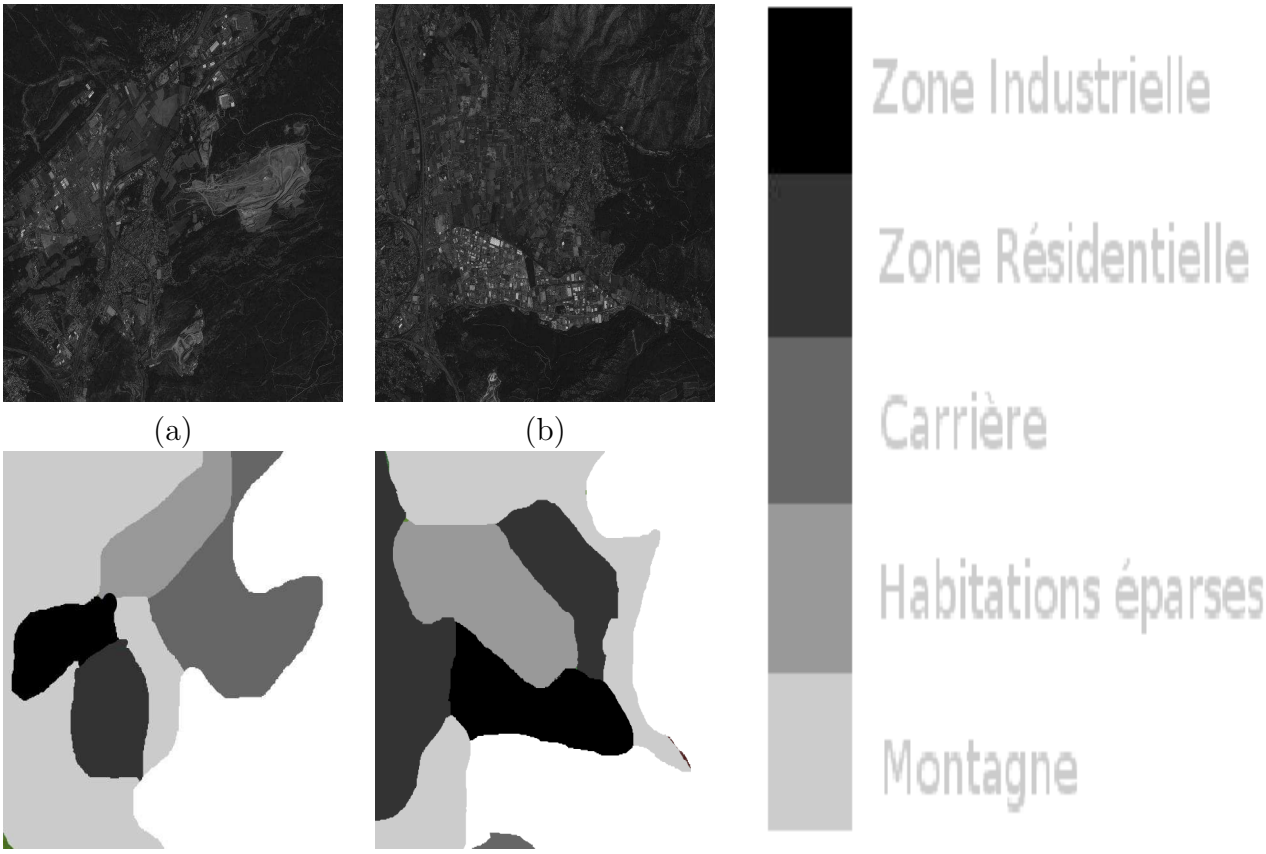
On applique cet algorithme sur la base d'images SPOT5 annotée utilisée pour l'évaluation quantitative. Les concepts introduits dans l'ordre sont : Champs, Montagne, Mer, Habitations éparses , Montagne, Bois, Zone résidentielle, Carrière, Zone d'activité, Aéroport, Centre ville, Zone rurale, Zone montagneuse, Banlieue industrielle, Raffinerie, Zone maritime, Agglomération, Cimetière, Marais salant On voit que l'ordre relative d'*importance* des concepts pour la couverture sémantique de la base de données dépend directement de la surface relative des zones couvertes par les différents concepts qui les annotent pour la base de données que l'on considère 5.4. Le lien avec la couverture linguistique d'un corpus apparait



(a) (b)

(c) (d)

Fig. 5.14 – Requête effectuée : Raffinerie à proximité de la mer et d'une zone résidentielle : (a) et (b) Images ayant les meilleurs scores de la fonction de cohérence ©CNES (c) et (d) : masque de la vérité terrain correspondant aux zones renvoyées



(a) (b)

(c) (d)

Fig. 5.15 – Requête effectuée : Zone industrielle bordant une zone résidentielle et à proximité de la montagne : (a) et (b) Images ayant les meilleurs scores de la fonction de cohérence ©CNES (c) et (d) : masque de la vérité terrain correspondant aux zones renvoyées

Concept	Pourcentage de surface couverte
Champs	22,29 %
Montagne	22,19 %
Mer	17,83 %
Habitations éparses	15,76 %
Bois	9,98 %
Zone résidentielle	7,87 %
Carrière	1,35 %
Zone d'activité	1,21 %
Aéroport	1,04 %
Centre ville	0,32 %
Raffinerie	0,08 %
Marais salants	0,07 %
Cimetière	0,01 %

Fig. 5.16 – Pourcentage de couverture des zones couvertes par les régions annotées par les différents concepts dans la base de données considérée.

de ce point de vue comme pertinente. Cet ordre dépend bien évidemment très fortement des paysages rencontrés dans la base de données. Les images d'Angers contiennent beaucoup de paysages champêtres, et les images de Marseille contiennent beaucoup de montagnes. D'où la prédominance des champs et des montagnes pour cette base de données.
Un bruit a été rajouté sur les caractéristiques de la base de données afin d'évaluer l'impact de la discriminance des caractéristiques de bas-niveau. L'intensité du bruit σ est ajustée, et le bruit est rajouté sur les images de test ainsi que sur les images d'apprentissage. L'apprentissage est alors effectué, où seuls les paramètres sont estimés, la structure du modèle étant fixée comme étant celle trouvée en 4.7.2.4. La fonction $P(D|M_\omega)$ est présentée figure 5.17. On voit que la fonction converge vers une limite dépendant de la discriminance des caractéristiques de bas-niveau. On peut conjecturer que cette limite dépend également de la complexité intrinsèque de la base de données.

5.5 Compression sémantique

Le problème d'annotation sémantique peut également être vu sous la forme d'une compression avec perte d'une base d'images. Les archives d'images satellitaires étant énormes, il convient de réduire leur taille tout en gardant l'information essentielle pour les utilisateurs. Une compression de type JPEG ou avec ondelettes permet d'atteindre des taux de compression de l'ordre de 10 à 20 sans que la perte de qualité de l'image ne soit visible. Mais pour obtenir plusieurs ordres de grandeur du taux de compression de l'image, on peut conserver l'interprétation sémantique de l'image sous la forme d'un graphe de régions annotées. Même si cette représentation ne permet pas de reconstruire l'image, elle conserve

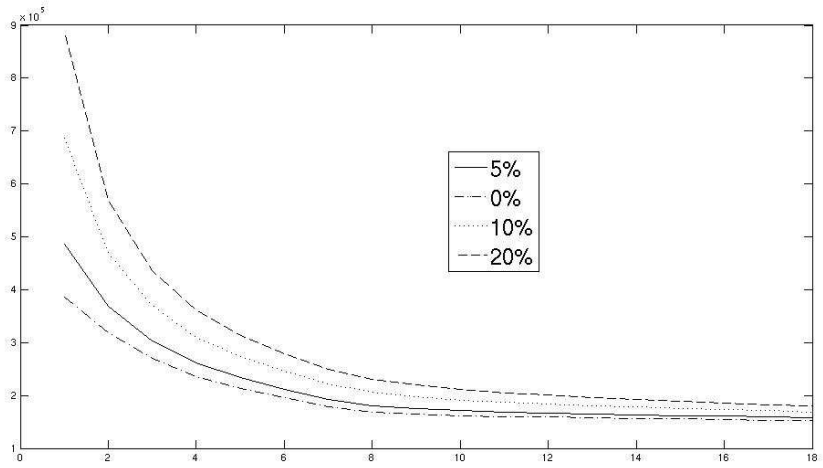


Fig. 5.17 – Log vraisemblance de la base de données en fonction du nombre de concepts inclus dans le modèle statistique pour différentes valeurs du bruit ajouté sur les caractéristiques.

l'information présente dans l'image qui est essentielle pour l'utilisateur, à savoir l'information sémantique. La représentation présentée en 5.1.3 donne ainsi de l'image une représentation sous forme d'un graphe dont les nœuds sont des concepts et dont les arcs sont des relations entre les régions annotées par ces concepts. Il convient ainsi de définir un équilibre entre la finesse de la description de la région et la compression souhaitée. En effet, les nœuds peuvent contenir des informations suppémentaires sur les régions telles que l'extension de la région ou des informations sur les frontières, tandis que les arcs peuvent contenir des informations plus fines sur les relations entre les régions telles que des relations floues.
Prenons une image SPOT5 panchromatique en niveau de gris codés sur 8 bits de taille 6000×6000 est codée sans perte sur 36Mo. Avec une seule couche de concepts, en considérant les 11 concepts que nous avons retenus en première couche pour annoter les images SPOT5, on considère une annotation de l'ordre de 12 régions avec une longueur de code de $\frac{-log((1/11)^1 2)}{8}$ octets, à savoir 4 octets. On code ensuite les relations entre régions avec 8 matrices

$$12 \times 12$$

en type float, ce qui fait un total de 4608 octets. Viennent ensuite les centre de gravité et les moments d'inertie des régions, qui nécéssitent 192 octets. Ainsi, une représentation en une couche nécéssite donc 4,8ko. Ce qui fait une compression d'un facteur 7500 conservant l'essentiel de l'information sémantique présente dans l'image.

Chapitre 6

Conclusion

6.1 Résumé et part d'innovation dans le travail effectué

Dans ce rapport, nous avons présenté une modélisation statistique d'une base d'images annotées. L'idée directrice est la mise en correspondance d'un réseau sémantique définissant des relations entre les différents labels servant à l'annotation, et d'un modèle statistique permettant de coder les images de la base d'apprentissage. A chaque type de relation sémantique liant différents concepts ("méronymie", "synonymie", "hyponymie") correspond une modélisation stochastique particulière. Un réseau sémantique unique est associé à un modèle stochastique global. Il est donc possible, partant d'une liste de labels et d'une base de données d'images exemple, d'effectuer une sélection du *meilleur* modèle (selon le critère MDL) et donc d'obtenir la structure du réseau sémantique contenant le vocabulaire annotant la base de données. Un nouveau label peut être incorporé de façon très simple au réseau existant, ce qui permet au modèle d'être très souple.
La structure du modèle étant définie et les paramètres étant estimés, il est possible d'annoter des images tests avec les labels contenus dans le réseau sémantique qui a été. L'image est décomposée en régions selon un nombre de partitions égal au nombre de couches du réseau sémantique. Cela permet ainsi une annotation selon plusieurs niveaux de complexité et de généralité. Des relations spatiales sont ensuite introduites et viennent enrichir les différentes partitions de l'image inférées par le système. Cela permet ainsi à un utilisateur d'effectuer des requêtes du type : "zone de végétation bordant des installations sportives et proche d'une zone industrielle".
Ainsi, la méthode proposée unifie deux types de méthodes :

– Les méthodes procédant par apprentissage statistique apprenant des classes à partir d'une base d'images (annotées faiblement ou non). Le problème d'annotation d'une nouvelle image est alors vu comme un problème de classification avec les classes dont les paramètres ont été estimées à l'apprentissage. Ces méthodes ont une certaine flexibilité car elles utilisent une phase d'apprentissage, mais ne prennent pas en compte les relations sémantiques

entre les différents labels utilisés pour l'annotation.
– Les méthodes utilisant un réseau sémantique. Dans ces méthodes, les relations sémantiques entre les concepts d'annotations sont exploitées mais ces méthodes sont généralement peu flexibles car le réseau sémantique contient une information à priori qui est fournie par un expert.

La méthode ASP permet donc de combiner la flexibilité des méthodes d'inférence statistiques avec la richesse descriptive provenant de la prise en compte des relations sémantiques. Les outils utilisés par la méthode ASP sont des outils relativement classiques dans le domaine du traitement d'images, de la reconnaissance de formes et de l'intelligence artificielle : modélisation par variables latentes, séléction de modèles, réseaux sémantiques etc. Cependant, l'innovation apportée dans ce travail se situe au niveau de la modélisation globale, à savoir la traduction sous forme de modèles mathématiques des relations sémantiques qui sont introduites (modèle de mélange d'unigrammes pour la relation d'hyponymie), et la définition globale permettant d'exprimer la vraisemblance d'une base d'images annotées. L'apport se situe également sur la modélisation de l'inférence des annotations attachées à une nouvelle image à partir du modèle stochastique et du réseau sémantique qui ont été appris lors de la phase d'apprentissage.

6.2 Perspectives d'amélioration dans le domaine de l'annotation sémantique

Des travaux de recherche supplémentaires seraient nécéssaires pour améliorer les résultats ou étendre les possibilités de la méthode présentée dans ce rapport. Cette section suggère quelques orientations qui pourraient être explorées.

6.2.1 Prise en compte d'un plus grand nombre de structures

Afin d'éviter des modélisations statistiques trop complexes et trop délicates à estimer, l'ensemble des structures possibles pour le réseau sémantique et le modèle statistique a été limité à un modèle par couches très restrictif. Il est possible d'envisager un réseau de relations entre les concepts qui serait plus complexe, en s'affranchissant par exemple de la modélisation en couches successives et qui pourrait ainsi décrire la réalité de manière plus fine (voir figure 6.1). La modélisation qui en résulterait s'en trouverait cependant considérablement compliquée. Or, dans le domaine très délicat de la modélisation, il est nécessaire de faire un compromis permanent entre la richesse de la description du réel, à savoir des données, et les capacités à estimer les paramètres qui sont à disposition. En effet, estimer les paramètres d'un modèle est un processus très délicat qu'il s'agit de prendre en compte avec réalisme et humilité. Ainsi, mieux vaut un modèle très simple, voire simpliste, donc on peut estimer correctement les paramètres, qu'un modèle très sophistiqué dont les paramètres ne peuvent pas être estimés de manière fiable et satisfaisante. Or, même dans le modèle très simple que nous

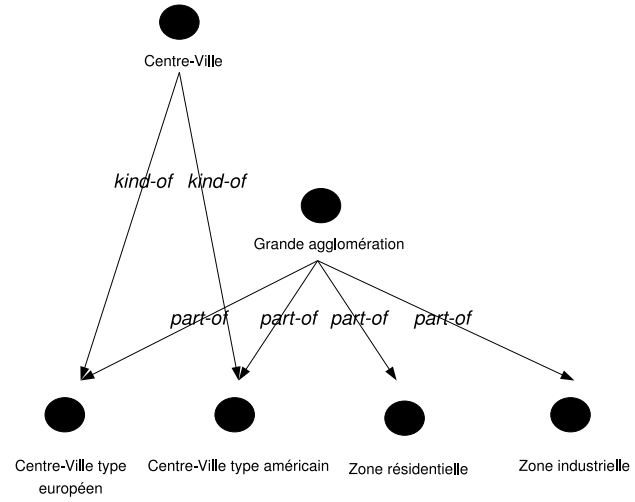
avons présenté, l'estimation des paramètres et les algorithmes d'optimisation reposent sur de fortes approximations. Une plus grande souplesse dans la structure du réseau sémantique doit donc aller de pair avec le développement d'algorithmes d'apprentissage, d'estimation et d'inférence efficaces.

6.2.2 Introduction d'information spatiale

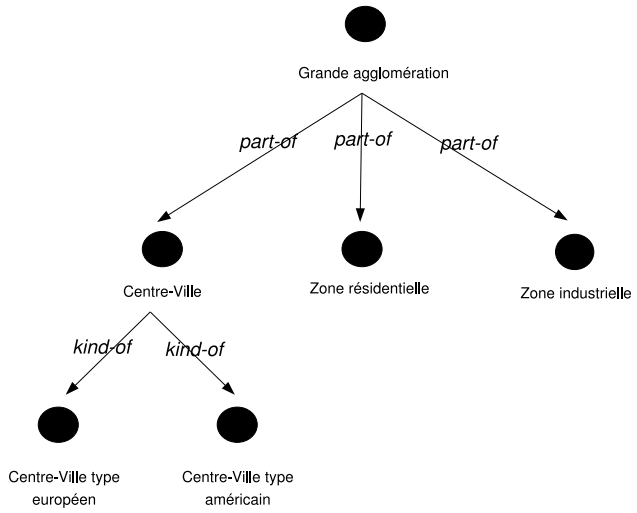
Dans le modèle qui est proposé, les relations spatiales ne sont prises en compte que tout à la fin du processus d'inférence, afin d'enrichir les annotations de l'image. Mais elles n'interviennent pas dans le modèle : ni au niveau du texton, ni au niveau des régions. Dans notre modélisation, le modèle le plus "bas" dans la hiérarchie est un modèle bayésien naïf, qui par définition ne prend pas en compte le contexte de chaque texton. Or, même si, en ce qui concerne les caractéristiques de bas-niveau utilisées dans nos travaux, les textons contiennent une information sur les relations spatiales entre les textons, une information sans doute très importante est contenue dans les relations spatiales entre les sites d'où sont extraites les caractéristiques de bas-niveau. De même, les relations spatiales entre les régions ne sont prises en compte que de façon très faible. En effet, le système pourra apprendre qu'une ville est susceptible de comprendre une gare, un centre-ville, une zone pavillonaire etc. Mais il ne pourra pas différencier une ville française selon les types "ville ancienne" ou "ville moderne" selon la disposition spatiale de ces différents éléments (gare présente au centre ou en périphérie). Il serait ainsi intéressant de pouvoir améliorer notre modélisation en prenant en compte ces informations spatiales.
Cependant, cette prise en compte de l'information spatiale nécéssiterait de faire évoluer le modèle ASP. Une difficulté importante de cette prise en compte est l'application délicate de la Minimisation de la Complexité Stochastique qui en résulte. En effet, dans notre approche, les différents modèles codent une même information : la séquence des valeurs des textons dans les régions. Si l'on souhaite introduire, des relations spatiales, les relations spatiales existantes entre régions situées en deuxième couche seraient différentes des relations de voisinages définies entre textons, et une information différente serait alors codée selon le placement dans la hiérarchie du concept d'annotation.

6.2.3 Amélioration des algorithmes d'optimisation

Les algorithmes employés dans ce travail, tant pour l'annotation que pour l'apprentissage du modèle, sont des algorithmes gloutons. Simples à mettre en oeuvre, ils permettent également d'obtenir des résultats dans tes temps raisonnables. Cependant, des algorithmes d'optimisation plus élaborés sont souhaitables afin que l'apprentissage et l'inférence des annotations soient plus fiables. En particulier, la simplification choisie dans le processus d'annotation, qui consiste à effectuer la maximisation de la vraisemblance de l'image sachant les p partitions en p étapes d'optimisation successives, a un inconvénient majeur. En effet, une partition de haut-niveau étant inférée directement à partir de la couche immédiatement inférieure, elle ne peut pas contribuer à donner un score à



(a)



(b)

Fig. 6.1 – (a) Méthode ASP : Les deux labels, "Centre-Ville de type européen" et "Centre-Ville de type américain", sont hyponymes du label "Centre-Ville", ce dernier label est donc mis automatiquement dans la couche C_2^{ko}. Les concepts "Centre-Ville de type européen" et "Centre-Ville de type américain" seront quant à eux reliés directement par une relation d'holonymie au label "Grande agglomération". Pourtant, il semble plus pertinent, comme dans le schéma (b), de représenter une agglomération comme étant constitué entre autres d'un centre-ville, celui-ci pouvant être spécialisé en type américain ou européen.

plusieurs configurations concurrentes. Cela ne permet donc pas, le cas échéant, de pénaliser des configurations de bas-niveau qui seraient incohérentes (cimetière isolé au milieu de la mer).

6.2.4 Création automatique de labels par le système

Les définitions des liens "kind-of" et "part-of" qui sont utilisées par la méthode ASP sont très contraignantes. En effet, un label général ne peut être déplacé dans une couche supérieure qu'à condition que toutes ses spécifications soient présentes dans le système. Ainsi, si un label "zone industrielle" peut se décomposer en sous-classes "raffinerie", "technopole" et "zone d'exploitation agricole" mais que seules les deux premières ont été introduites dans le système, il semble judicieux de permettre au système de créer un label supplémentaire afin que le label "zone industrielle" puisse être déplacé dans une couche plus générale. Ainsi, un label "zone industrielle autre que raffinerie et technopole" peut être ajouté avec des paramètres estimés sur les images de raffinerie qui ne sont pas correctement décrites par les modèles correspondant aux images de raffinerie et de technopole. Cela crée alors un modèle n'ayant pas de label introduit par un utilisateur. Cependant, il est envisageable d'informer l'utilisateur de la possible création de ce nouveau label et de lui demander de nommer les images qui y correspondent.

Annexe A

Classification non-supervisée de patchs dans des images Quickbird

Depuis son introduction, le descripteur SIFT (Scale Invariance Feature Transform) a suscité beaucoup d'enthousiasme dans le communauté de vision par ordinateur et est à présent considéré comme un descripteur compétitif relativement à d'autres descripteurs [83] [88]. Etant donné un point où est calculé le descripteur SIFT, quatre fenêtres 4×4 sont considérées : chacune est pondérée par l'amplitude du gradient et par une fenêtre circulaire gaussienne avec une écart type d'une valeur de l'ordre de 4 ou 6. Par la suite, les histogrammes locaux d'orientation sont calculés pour chacune de ces 4 fenêtres : dans notre cas, des histogrammes à 4 valeurs sont utilisés, couvrant l'ensemble $[0, \pi]$ (des descriptions opposées sont supposées décrire le même type d'objets). Les histogrammes sont ensuite normalisés. Chacune des 4 fenêtres est ainsi décrite par un histogramme à 4 valeurs : la concaténation de ces 4 descripteurs produit un descripteur local de taille 16 : le SIFT. Afin de garder l'invariance par rotation du descripteur, le voisinage des caractéristiques calculées subit une rotation de façon à ce que le gradient local ait une direction horizontale. Ici, contrairement à ce qui est fait habituellement, le descripteur SIFT n'est pas extrait en des points de Harris, mais selon une grille régulière ayant un pas de 8 textons. En effet, notre objectif n'est pas de faire de matching d'objets, mais simplement d'avoir une caractéristation des images ayant une structure de grille régulière. En effet, nous supposons que le descripteur SIFT extrait des informations géométriques pertinentes pour caractériser des images à haute résolution.

A.1 Quantification des descripteurs SIFT

Les descripteurs SIFT qui sont ainsi extraits dans le corpus d'images sont ensuite quantifiés. Pour cela, une partie des vecteurs de caractéristiques qui sont extraits sont utilisés pour faire l'apprentissage d'un "codebook". De même que pour la section 3.1.1, les codewords ayant été calculés, toutes les caractéristiques du corpus d'images sont quantifiées et leur localisation est prise également en compte. Nous

avons donc un nouveau lot d'images dont les valeurs de textons sont les indices des codewords dans le codebook, ces valeurs sont donc comprises dans l'ensemble $\{1, ..., N\}$. Ainsi, nous voyons ces vecteurs quantifiés comme des textons d'une nouvelle image dont les niveaux de gris sont les indices des codewords dans le codebook. Nous appelerons dans la suite ces vecteurs quantifiés "textons".

A.2 Regroupement en patchs de descripteurs SIFT

Les descripteurs SIFT apportant à eux seuls une information beaucoup trop locale, ils sont regroupés en ce que nous appelerons ici des patchs, à savoir des fenêtres carrées de taille fixée (typiquement 40*40 ou 50*50) permettant d'agréger ces informations locales. La surface des patchs doit être suffisante pour permettre de faire des statistiques fiables sur les occurrences de chaque type de texton, tout en étant également suffisamment faible pour ne pas contenir des zones trop hétérogènes. La répartition spatiale des textons à l'intérieur des patchs est négligée et seul l'histogramme des types de textons présents à l'intérieur des patchs est pris en compte.
Les notations employées ici sont les suivantes : l'ensemble des textons présents dans les images est noté O, l'ensemble des textons présents dans le patch i est noté O_i, N est le nombre de patchs, n est le nombre de types de textons. La modélisation probabiliste suivante est alors effectuée : nous supposons qu'à chaque patch correspond une réalisation d'une variable aléatoire tirée indépendamment pour chaque patch, les textons sont ensuite engendrés selon une loi de probabilité multinomiale dont les paramètres dépendent de la valeur prise par la variable latente. Cette variable latente prend une valeur parmi un vocabulaire de taille K : $\{V_1, ..., V_K\}$ avec probabilité $\pi = \{\pi_1, ...\pi_K\}$. Si un patch est associé à la réalisation V_i de la variable aléatoire, les textons présents dans le patch sont supposés être tirés par la loi multinomiale de paramètres $\theta_i = \{p_{i1}, p_{i2}, ...p_{in}\}$ (voir figure A.1). On appelle θ la matrice dont les colonnes sont les θ_i.

A.3 Apprentissage.

On souhaite déterminer les vecteurs θ et π qui maximisent la vraisemblance de la réalisation des observations O conditionnellement aux modèles θ et π. Pour cela, un algorithme d'expectation-maximisation est mis en œuvre. En effet, l'algorithme EM est basé sur l'introduction de variables cachées dont la connaissance permet d'optimiser plus simplement la vraisemblance. Le modèle étudié ici étant un modèle de mixture, les variables cachées ici peuvent être choisies ici très naturellement comme les variables latentes associées aux patchs. L'algorithme EM permet ici de trouver les paramètres des lois multinomiales correspondant à un maximum local de la vraisemblance de la réalisation de O. Les calculs qui en découlent sont les suivants :

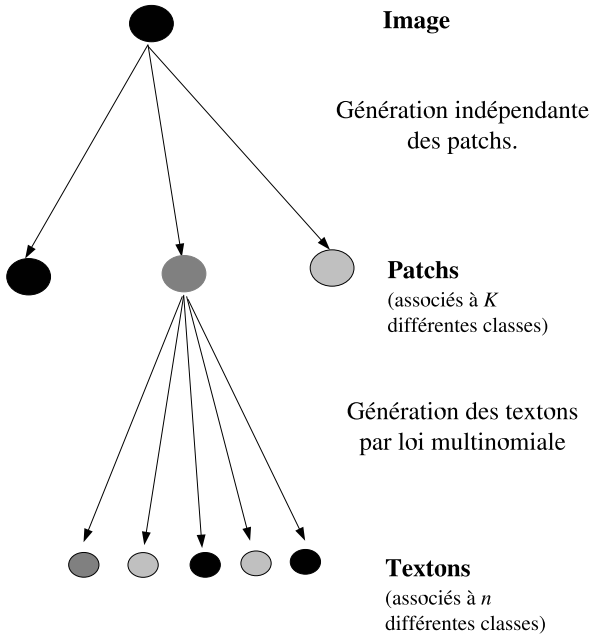


Fig. A.1 – Schéma de classification des patchs utilisé

Étant donné l'indépendance de la génération des patchs, on écrit :

$$P(O|\theta,\pi) = \prod_{i=1}^{N} P(O_i|\theta,\pi)$$

.
En prenant le logarithme de cette expression, et en conditionnant par rapport aux valeurs possibles de la variable latente, on obtient, N_i étant le nombre de textons dans le patch i :

$$logP(O|\theta,\pi) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} log(\pi_k Mult_{N_i,\theta_k}(O_i))$$

.
Où z_{ik} est une variable qui vaut 1 si la variable latente $Z_i = k$, c'est-à-dire la variable latente Z vaut 1 pour le patch i : En notant N_{ij} le nombre de textons de type j présents dans le patch i, $Mult$ est la loi multinomiale définie par :

$$Mult_{N_i}(O_i|\theta_z) = \frac{N_i!}{N_{i1}!N_{i2}...N_{iK}!} p_1^{N_{i1}}...p_K^{N_{iK}} \quad (A.1)$$

On prend l'espérance par rapport à O, θ et π, pour obtenir :

$$E_{Z|O,\theta,\pi}(logP(O|\theta,\pi)) = \sum_{i=1}^{n}\sum_{k=1}^{K} E(z_{ik}|O_i,\theta) log(\pi_k Mult_{N_i,\theta_z}(O_i))$$

.

En posant $\gamma_k(O_i) = E(z_{ik}|O_i,\theta)$, on obtient la formule :

$$E_{Z|O,\theta,\pi}(logP(O|\theta,\pi)) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k(O_i) log(\pi_k Mult_{N_i,\theta_z}(O_i))$$

.

En introduisant dans cette équation la formule A.1, on peut écrire :

$$E_{Z|O,\theta,\pi}(log(P(O,z|\theta,\pi) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k(O_i)(log(\pi_k) + \sum_{j=1}^{N_i} j - \sum_{l=1}^{K}\sum_{j=1}^{N_{il}} log(j) + \sum_{j=1}^{K} N_{ij} log(p_{kj}))$$

.

L'algorithme EM permet de trouver un maximum local de cette vraisemblance.

Etape E : calcul de $\gamma_k(O_i)$, pour tout k et i par la règle d'inversion de Bayes.

$$\gamma_k(O_i) = \frac{\pi_k Mult_{n_i,\theta_k}(O_i)}{\sum_{j=1}^{K} \pi_j Mult_{n_i,\theta_j}(O_i)}$$

.

Etape M : maximisation de $E_{Z|O,\theta,\pi}(logP(O,z|\theta,\pi))$, pour cela la méthode des multiplicateurs de Lagrange est utilisée pour optimiser $E_{Z|O,\theta,\pi}(log(P(O,z|\theta,\pi))$, on obtient :

$$p_i^k = \frac{\sum_{i=1}^{n} \gamma_k(O_i) n_i}{\sum_{i=1}^{n} \gamma_k(O_i) N_i}$$

.

$$\pi_k = \frac{\sum_{i=1}^{n} \gamma_k(O_i)}{n}$$

.

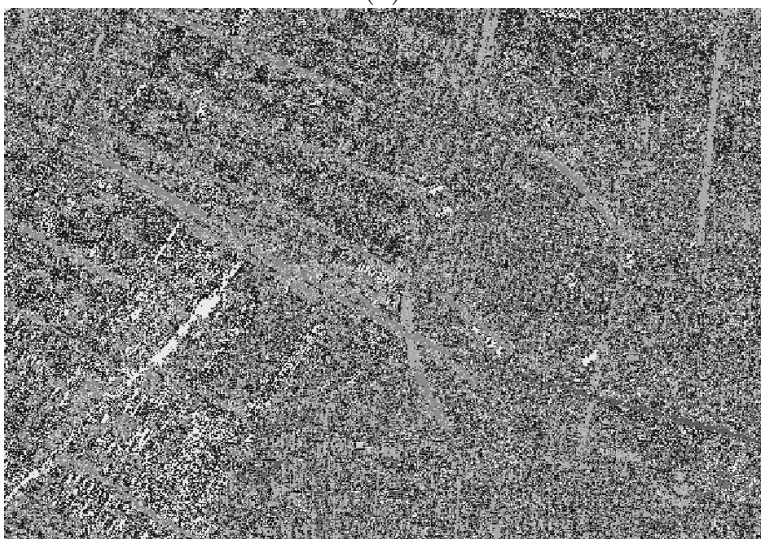
A.4 Classification des patchs

Les paramètres du modèle ayant été calculés, on peut alors classifier les patchs d'une nouvelle image en choisissant les valeurs des variables latentes qui maximisent la probabilité *a posteriori* $P(Z|O)$. En gardant les mêmes notations que dans la section précédente, on peut écrire, par indépendance des patchs :

$$P(Z|O) = \prod_{i=1}^{N} P(Z_i|O_i)$$

Ainsi, trouver les réalisations des variables latentes correspondantes à chaque patch revient très simplement à maximiser $P(Z_i|O_i)$ indépendemment pour chaque patch i. Ainsi, pour chaque patch, on prend comme réalisation de la variable aléatoire : $argmax_k(\pi_k Mult_{\theta_k}(O_i))$.
Nous avons appliqué cette méthode pour deux valeurs possibles de la variable latente sur une base de 30 images 512*512. Des textons sont extraits sur une grille régulière tous les 5 textons. Un patch a une taille de 40 textons de côté, il contient ainsi 64 textons, un texton pouvant avoir 25 valeurs possibles.

(a)



(b)

Fig. A.2 – Résultat de quantification des descripteurs SIFT : (a) l'image d'origine est de taille 5500×4000 ©(LIAMA) (b) : la nouvelle image obtenue est de taille 687×512, elle est obtenue après que les descripteurs SIFT aient été quantifiées. Les valeurs de niveau de gris de cette image correspondent aux indices des codewords dans le codebook.

Annexe B

Modélisation markovienne

Dans ce chapitre, nous décrivons un bref aperçu des notions de base de la modélisation markovienne. Les champs de Markov cachés ont connu un essor considérable en traitement d'images à partir des années quatre-vingts. En effet, ils emploient un formalisme mathématique rigoureux pour prendre en compte l'information contextuelle de l'image. Chaque texton (indice de codeword) est vu comme la réalisation d'un champ $Y = (Y_s)_{s \in S}$ correspondant à la sémantique sous-jacente, formalisée comme un champ aléatoire caché $X = (X_s)_{s \in S}$, où S est l'ensemble des textons. Le champ $X = (X_s)_{s \in S}$ prend ses valeurs dans l'ensemble des concepts $\Omega = \{s_1, ..., s_n\}$, tandis que le champ aléatoire $Y = (Y_s)_{s \in S}$ prend ses valeurs dans l'ensemble des valeurs des textons $\{1, ..., N\}$, où N est la taille du codebook construit lors du clustering des caractéristiques. Ainsi, le processus d'extraction de sémantique consistera à estimer $X = (X_s)_{s \in S}$ sachant $Y = (Y_s)_{s \in S}$. Ce problème d'inférence statistique très complexe peut être traité efficacement, dans le cadre des modèles markoviens, par des méthodes d'estimations bayésiennes de X à partir de Y. Ces techniques d'estimation sont réalisées grâce à des techniques générales de simulation dites "méthodes de Monte Carlo par chaines de Markov"(MCMC de l'anglais "Monte Carlo Markov Chains") qui sont applicables à condition que la loi de X conditionnelle à $Y = y$ (sa loi "*a posteriori*") soit de Markov.

B.1 Modélisation markovienne

B.1.1 Généralités

B.1.1.1 Système de voisinage et de cliques

Un sous-ensemble c d'une image I est appelée clique relative au système de voisinage V, si c est un singleton ou si tous les sites distincts de c sont voisins. Une clique est dite d'ordre P si elle contient P éléments, en d'autres termes, l'ordre d'une clique est par définition son cardinal. On notera C l'ensemble des cliques associées à un système de voisinage.

B.1.1.2 Champ de Markov et distribution de Gibbs par rapport à un système de cliques

Champ de Markov Soit S un ensemble de textons et $X = (X_s)_{s \in S}$ une famille de variables aléatoires, définies sur S, où chaque X_s est à valeurs dans un ensemble fini de classes $\Omega = \{s_1, ..., s_m\}$. Soit un système de voisinage V, on note par la même lettre p les diverses lois de probabilité liées à X.
On dit qu'un champ X est un champ de Markov relativement à V si et seulement si deux les propriétés suivantes sont vérifiées :
– Positivité :

$$\forall x \in \Omega^{Card(S)}, p(x) \geqslant 0$$

– Markovianité :

$$p(x_s|x_t, t \neq s) = p(x_s|x_t, t \in V_s) \tag{B.1}$$

La propriété B.1 signifie que la probabilité en un site s conditionnellement à tous les autres sites de l'image n'est fonction que de la configuration du voisinage du site considéré.

Champ de Gibbs Soit S un ensemble de textons muni d'un système de voisinage V, et soit C l'ensemble des cliques c associées à V. Le champ X est un champ de Gibbs, si et seulement si sa loi est définie par :

$$p(x) = \frac{1}{Z} exp(-U(x))$$

avec : $U(x) = \sum_{c \in C} \phi_c(x_c)$ qui est dite "fonction d'énergie", ainsi que la fonction Z définie par :

$$Z = \sum_{x \in \Omega^{Card(S)}} exp(-U(x)) \tag{B.2}$$

Une telle loi est dite "distribution de Gibbs". Notons que ϕ_c sont des applications de $\Omega^{Card(S)}$ dans R appelées "fonction potentiel" avec x_c al restriction de x à c et Z est une constante de normalisation, également "fonction de partition" qui fait de $p(x)$ une probabilité.Etant donné que cette constante est une somme sur toutes les configurations possibles du champ X, son calcul est dans la pratique généralement impossible.

Equivalence entre un champ de Gibbs et un champ de Markov Comme nous l'avons évoqué précedemment, le champ de Markov est caractérisé par sa propriété locale tandis que le champ de Gibbs est caractérisé par sa propriété globale. Le théorème de Hammersley-Clifford (1971) établit l'équivalence entre ces deux propriétés.

Théorème Soit S un ensemble de textons muni d'un système de voisinage V. Un champ X sur S est un champ de Markov relativement à V, si et seulement si X est un champ de Gibbs de potentiel associé à V.
L'intérêt pratique de ce théorème est qu'il permet d'accéder d'une manière simple à une forme exploitable ds probabilités jointes et ce, en spécifiant les fonctions potentiel ϕ_c définies sur les restrictions x_c de x aux cliques c. Afin de calculer la probabilité jointe du champ de Markov qui est une distribution de Gibbs, il est nécessaire d'évaluer la fonction de partition B.2. Comme la somme englobe l'ensemble des configurations, ce calcul est en général impossible. Cependant, il est possible de simuler des réalisations de ce champ à partir des caractéristiques locales, grâce à des méthodes de relaxation que nous allons voir dans la section suivante.
Ainsi, l'échantilloneur de Gibbs permet cette simulation lorsque les probabilités conditionnelles :

$$p(x_s|x_v) = \frac{exp(-\sum_{c\in C}\phi_c(x_s))}{\sum_{x_s\in\Omega} exp - \sum_{c\in C}\phi_c(x_s)}$$

sont connues. Sachant que la taille des cliques ne dépasse généralement pas quatre éléments, les probablités conditionnelles sont calculables.

B.1.2 Estimation bayesienne

Soit $X = (X_s)_{s\in S}$ et $Y = (Y_s)_{s\in S}$ deux champs aléatoires comme décrits ci-dessus. le problème de la segmentation bayesienne consiste en l'estimation de la réalisation invisible X à partir des données observées $Y = y$. Ainsi, le problème est de déterminer une estimation $\hat{x} \in \Omega$ de x à partir de y, obtenue en optimisant un certain critère.
L'estimation bayesienne nécessite l'estimation d'une fonction de coût, L, définie dans

$$\Omega^{card(S)} \times \Omega^{card(S)} \rightarrow R^+$$

. Celle-ci possède les propriétés suivantes :

$$\forall x, \hat{x} \in \Omega^{Card(S)} \times \Omega^{Card(S)} :$$

$$L(x, \hat{x}) \geq 0;$$

$$L(x, \hat{x}) = 0 \Leftrightarrow x = \hat{x}$$

Le risque bayesien associé à la stratégie $\hat{s} : R^{Card(S)} \rightarrow \Omega^{Card(S)}$ est donné par le coût moyen $R = E[L(\hat{s}(Y), X])]$. La stratégie bayesienne $\hat{s}_B$ est une stratégie dont le risque bayesien est minimum :

$$E[L(\hat{s}(Y), X])] = min_{\hat{s}} E[L(\hat{s}(Y), X])]$$

L'estimateur baysien $\hat{s}_B$ est alors obtenu en minimisant l'espérance de cette fonction conditionnellement aux observations :

$$\hat{s}_B(y) = argmin_{\hat{x}\in\Omega} E[L(\hat{x}(Y), X)|Y = y]$$

A chaque fonction de coût est associé un estimateur bayesien. Nous présentons dans la suite deux fonctions de coût très utilisées, qui correspondent à deux fonctions de coût, très utilisées, qui définissent les estimateurs les plus répendus dans la littérature, à savoir les estimateurs du "*maximum a posteriori*" (MAP), et l'estimateur du "*mode des marginales a posteriori*"(MPM).
Le choix d'un estimateur est laissé le plus souvent à l'appréciation de l'utilisateur. Cependant, dans la suite de ce chapitre, notre choix s'est porté sur l'estimateur MPM.

Estimateur Maximum a posteriori MAP l'estimateur MAP est associé à la fonction de coût suivante :

$$L(\hat{x}, x) = 1 - \delta(\hat{x}, x)$$

où la fonction δ est définie par $\delta(x_s, x_t) = 0$ pour $x_s \neq x_t$ et $\delta(x_s, x_t) = 1$ pour $x_s = x_t$.
Remarquons que cette fonction pénalise de la même manière une erreur sur un site et une erreur sur plusieurs sites. On peut alors écrire :

$$E[L(\hat{x}, x)|Y = y] = \sum_{x\in\Omega} L(\hat{x}, x)p(x|y)$$

$$E[L(\hat{x}, x)|Y = y] = 1 - p(\hat{x}|y)$$

Par conséquent, l'estimation bayesienne est :

$$\hat{x} = argmin_{\hat{x}\in\Omega}(1 - p(x|y)) = argmax_{\hat{x}\in\Omega}(p(x|y))$$

L'estimateur au sens du MAP revient donc à maximiser la probabilité *a posteriori*.

Estimateur du Mode des Marginales a Posteriori (MMP) Pour éviter la sévérité de la fonction de coût de l'estimateur MAP, une autre fonction de coût moins restrictive, est associées à l'estimateur MPM :

$$L(\hat{x_s}, x_s) = \sum_{s\in S} 1 - \delta(\hat{x_s}, x_s)$$

Cette fonction consiste à pénaliser l'erreur commise en fonction du nombre de sites mal estimés.
On montre que l'estimation bayesienne $\hat{x_s}$ est ici obtenue suivant :

$$\hat{x_s} = argmax_{\hat{x_s}} p(x_s|y)$$

Cette estimation ressemble à l'estimateur MAP, mais effectuée de façon locale. Autrement dit, on passe de la probabilité conditionnelle globale d'une configuration , à la probabilité conditionnelle en un site. Ainsi, la configuration

optimale est obtenue lorsque toute les lois marginales en chaque site sont maximisées. Cependant, le calcul direct des probabilités *a posteriori* $p(x_s|y)$ est impossible, compte tenu du gigantisme en pratique de l'espace des configurations., mais le fait de pouvoir simuler des réalisations de X par des approximations de type Monte Carlo permet leur approximation. On pose ainsi :

$$\hat{p}(x_s = \omega|y) = \frac{1_{[x_s^1=\omega]} + ... + 1_{[x_s^N=\omega]}}{N}$$

Finalement, pour la segmentation MPM, on procède de la manière suivante :
- Simuler N réalisations $x^1, x^2, ..., x^N$ de X selon la probabilité $Y = y$ en utilisant l'échantilloneur de Gibbs ;
- Estimer à partir des réalisations $x^1, x^2, ..., x^N$ la loi de chaque $X_s$par les fréquences ;
- Retenir la classe maximisant la loi ainsi obtenue.

B.1.3 Simulation d'un champ de Markov

La simulation des réalisations d'un champ de Markov est un outil très utile pour résoudre un problème dont on ne peut pas proposer une solution analytique. En effet, dans le cas de la segmentation d'images non supervisée par exemple, où les paramètres du modèle sont à estimer, certaines méthodes d'etimation des paramètres demandent la réalisation du champ X. Nous présentons ici l'échantilloneur de Gibbs.
- Initialiser une première carte x^0 de façon arbitraire
- A chaque itération n :
  - Balayer l'ensemble des sites $s\epsilon S$, et en chaque site s :
    - Calculer les probabilités $p(x_s^n|x_{V_s})$
    - Effectuer un tirage aléatoire d'une variable dans Ω, selon ces probabilités conditionnelles et poser $x_s^{n+1} =$.

On obtient ainsi une suite $x^0, x^1, ..., x^N$ de réalisations aléatoires du champ X, dont la loi converge vers $p(X)$.

B.2 Apprentissage des paramètres

Les paramètres de la loi du couple (X, Y) est constitué d'un ensemble α qu'on appelle "*ensemble de paramètres d'interaction* qui définit la loi du champ caché X, et d'un ensemble η définissant la loi de Y conditionnelle à X (ou encore le *bruit*). On notera $\theta = (\alpha, \eta)$ l'ensemble des paramètres nécéssaires pour mettre en oeuvre une méthode de segmentation choisie. Cependant, dans les applications réelles, ces paramètres sont la plupart du temps inconnus, d'où la nécessité de développer des méthodes pour procéder à leur estimation.
Deux situations de complexité différentes sont alors possibles. La première concerne le cas où les données sont dites "complètes" : dans ce cas, la réalisation de X ainsi que celle de Y sont connues, c'est-à-dire que l'on dispose des observations ains que de la "vérité terrain". Le deuxième cas de figure, beaucoup plus difficile à

traiter, est le cas du problème à "*données incomplètes*". Dans ce cas, l'estimation se fait uniquement à partir des observations, c'est-à-dire à partir de la réalisation du champ Y. Pour ces problèmes, une série de méthodes sont à disposition, la plupart d'entre elles utilisant le critère du maximum de vraisemblance.

B.2.1 Apprentissage à données complètes : le gradient stochastique

Proposé par Younes [144], la méthode du gradient stochastique est une méthode itérative qui permet de chercher un optimum local du maximum de vraisemblance. Le principe de l'algorithme est le suivant :
- Initialiser le vecteur de paramètre α_0
- A chaque itération n, calculer α_{n_1} à partir de α_n et de $X = x$ par la formule de mise à jour suivante :

$$\alpha_{n+1} = \alpha_n \frac{K}{n+1}(\vec{\nabla}_{\alpha_n U(x_{n+1}) - \vec{\nabla}_{\alpha_n} U(x)})$$

où $\vec{\nabla}_{\alpha_n} U(x)$ est le gradient de $U(x)$ par rapport à α pris en α_n, et K une constante. Notons qu'en pratique, on prend la constante K inversement proportionnelle au nombre de textons.

B.2.2 Apprentissage à données incomplètes

Dans cette section, nous décrivons différentes manières de traiter le problème de l'estimation des paramètres dans le cas d'un apprentissage à données incomplètes. C'est un cas de figure où l'on ne dispose que de la réalisation du champ Y, et pas du champ des classes X. De nombreux travaux ont été réalisés à ce sujet et ont débouchés sur différents algorithmes. Nous présenterons ici simplement l'algorithme "*Stochastic Expectation Maximization*" et l'algorithme du "*Gradient stochastique*"

B.2.2.1 Stochastic Expectation-Maximisation

Nous rappelons que l'algorithme "*Expectation-Maximisation*" introduit par Dempster [33] se décompose de la façon suivante :
- Initialisation du vecteur de paramètres θ_0
- A chaque itération n :
  - Etape E : Calcul de l'epérance

$$E_{\theta_n}(log(p_{\theta_n}(X, y)|Y = y)$$

  - Etape M : Maximisation de

$$\theta_{n+1} = argmax E_{\theta_n}(log(p_\theta(X, y)|Y = y)) \quad \text{(B.3)}$$

Dans le cadre des champs Markoviens, l'estimation décrite par la dernière formule n'est pas réalisable pratiquement. C'est pourquoi il est nécessaire de mettre en oeuvre, pour conserver le principe de l'algorithme EM, de le modifier en un algrithme stochastique de complexité comparable. Ainsi, l'algorithme Stochastic Expectation Maximisation est un algorithme itératif qui est une variante stochastique de l'algorithme EM. L'idée principale est d'incorporer une étape stochastique (étape S) précédant l'étape d'estimation (étape E) de l'algorithme EM. Cette étape S repose sur un principe d'affectation aléatoire selon un tirage de X suivant les lois *a posteriori*.
Plus précisement, les étapes de l'algorithme sont les suivantes :
- Initialiser le vecteur de paramètre θ_0
- A chaque étape n d'itération,
  - Simulation d'une réalisation x^n de X selon la probabilité *a posteriori* $p(x|y, \theta_n)$
  - Estimation de θ à partir de (x^n, y) par l'estimateur du maximum de vraisemblance θ_{MV} : $\theta_{n+1} = \theta_{MV}(x^n, y)$.

B.2.2.2 Gradient stochastique

L'algorithme de gradient stochastique de Younes, dont le principe est de rechercher est un estimateur du maximum de vraisemblance, a d'abord été proposé dans le cas des données complètes [144] avant d'être généralisé au cas des données incomplètes [145]. Ainsi, étant donné la réalisation du champ $Y = y$, il s'agit de maximiser $p_\theta(y)$
- Initialiser le vecteur de paramètre α_0
- A chaque itération n, calculer α_{n_1} à partir de α_n et de $X = x$ par la formule de mise à jour suivante :

$$\alpha_{n+1} = \alpha_n \frac{K}{n+1} (\vec{\nabla}_{\alpha_n U(x_{n+1}) - \vec{\nabla}_{\alpha_n} U(x)})$$

  où $\vec{\nabla}_{\alpha_n} U(x)$ est le gradient de $U(x)$ par rapport à α pris en α_n, et K une constante. Notons qu'en pratique, on prend la constante K inversement proportionnelle au nombre de textons.

où (x_n, y_n) est une réalisation du champ (X, Y), simulée par l'échantilloneur de Gibbs, en utilisant le paramètre θ_n, et x_{n+1} une réalisation du champ X, simulée en utilisant l'échantilloneur de Gibbs, suivant la loi *a posteriori* utilisant θ_n, K étant une constante, prise généralement égale à l'inverse du nombre de textons.

Annexe C

Inférence probabiliste de concepts sémantiques dans des images satellitaires

Dans l'approche que nous présentons dans ce chapitre, des vecteurs de caractéristiques de bas-niveau sont tout d'abord extraits dans l'image selon une grille régulière. Ces vecteurs sont ensuite quantifiés. Ainsi, chaque vecteur de caractéristiques est associé à un indice, ce qui permet de travailler à partir d'un vocabulaire discret. La localisation spatiale de chaque vecteur de caractéristiques étant conservée, on verra ainsi ces caractéristiques discrêtisées comme une nouvelle image dont les niveaux de gris de chaque texton sont les indices auxquels est attaché chaque vecteur de caractéristiques. Notons qu'aucune comparaison n'est possible entre les "textons" ainsi obtenus, car une similarité entre deux indices ne correspond en rien à une similarité dans l'espace des caractéristiques.
Nous décrivons ici une modélisation probabiliste de la génération de ces textons par un loi de mélange. Etant donné une région, une hypothèse d'indépendance des textons sachant le concept sémantique est supposée. Seul l'histogramme des valeurs des textons est ainsi pris en compte. Si la spatialité entre les textons n'est pas prise en compte, notons cependant que chacun de ces textons est un vecteur de caractéristiques quantifié. A ce titre, le texton contient en lui même une information sur la répartition spatiale des textons de l'image d'origine sur laquelle a été calculé le vecteur de caractéristiques qui lui correspond.
L'apprentissage est fait à partir d'une base de données d'apprentissage fournie par l'utilisateur. Une annotation complète d'un lot d'images étant une téche très coûteuse, nous supposons que cette base de donnée consiste en des images découpées correspondant chacune à un concept sémantique. Une méthode Expectation-Maximisation est mise en œuvre pour l'apprentissage des paramètres de la loi de mélange. Le critère de minimisation de la complexité stochastique est utilisé pour déterminer la complexité optimale du modèle.
Etant donné une image nouvelle à annoter par des concepts sémantiques, un algorithme "glouton" est mis en œuvre pour fournir, en un temps raisonnable, une annotation de l'image correspondant à un maximum local de la vraisemblance sur

l'ensemble des annotations possibles de l'image.
Des évaluations ont été faites sur des images de Pékin en utilisant des descripteurs SIFT comme caractéristiques de bas niveau. Un ensemble de onze concepts sémantiques a été défini : zone urbaine dense, zone résidentielle pavillonaire, chantier, zone résidentielle grand ensemble, champs, serres, terrain vague, étang, zone commerciale, zone industrielle, nœud autoroutier, installation sportive. Nous avons pour but de proposer une méthode qui soit aussi générale que possible tant au niveau des caractéristiques de bas-niveau utilisées que des concepts choisis pour annoter l'iamge. Nous évaluerons ainsi cette approche sur des images panchromatiques SPOT5 en utilisant des caractéristiques de Haralick.

C.1 Principe de la méthode

L'approche que nous détaillons ici se prêtant à une utilisation abondante de notations, elles sont rassembées dans le tableau C.1.

C.1.1 Modélisation bayésienne

Soient n le nombre de concepts avec lesquels on souhaite annoter le corpus d'images que l'on considère. Soit $G_I = \{S_1, S_2, ..., S_m\}$ l'ensemble des régions sémantiques qui ont été trouvées dans une image donnée I appartenant à ce corpus, m étant le nombre de régions trouvées dans l'image. Comme dit en introduction de ce chapitre, les caractéristiques de bas-niveau quantifiées extraites selon une grille régulière seront considérées comme des textons que nous allons relier aux concepts sémantiques. L'ensemble des textons de l'image sera ici noté O_I. On définit une régions sémantique S_l par un concept qui lui est attaché ("zone urbaine dense", "banlieue résidentielle" etc.) et un ensemble 4-connexe de textons qui sont supposés être générés par cette région et que nous notons ici $S_l(O_I)$. Nous supposons pour le moment que chaque texton doit être relié à une et une seule région, ce qui implique que $\{S_l(O_I)\}_{l\epsilon\{1,...,m\}}$ forme une partition de O_I.
Afin de définir l'ensemble de régions sémantiques G_I qui correspond le mieux à I avec les concepts à disposition, on souhaite maximiser le maximum a posteriori $P(G_I|O_I)$ des textons sachant l'annotation de l'image.

$$max_{G_I} P(G_I|O_I) = max_{G_I} \frac{P(G_I)P(G_I|O_I)}{P(O_I)} \tag{C.1}$$

$P(O_I)$ ne dépendant pas de G_I, maximiser la probabilité a posteriori revient à maximiser le produit $P(G_I)P(G_I|O_I)$. De plus, les textons sont supposés ne dépendre que de la région qui les a générés, par conséquent, on peut écrire :

$$max_{G_I} P(G_I|O_I) = P(G_I) \prod_{l=1}^{m} P(S_l(O_I)|S_l) \tag{C.2}$$

Les modélisations proposées pour exprimer $P(S_l(O_I)|S_l)$ et $P(G_I)$ seront présentées respectivement dans les deux sections suivantes.

Symbole	**Sens**
N	Nombre de codewords dans le codebook
m	Nombre de régions sémantiques trouvées dans une image
n	Nombre de concepts (définis par l'utilisateur)
s	Numéro d'indice d'un concept
K_s	Nombre de modèles dans le mélange associé au concept s
N_i	Nombre de textons dans l'image d'indice i du lot d'apprentissage associé à un concept
N_{ij}	Nombre de textons dont la valeur est j dans l'image d'indice i du lot d'apprentissage d'un concept spécifique
p_j^{ks}	probabilité de génération d'un texton de valeur j dans le modèle d'indice k du mélange associé à un concept s
N_s	Nombre de textons de la région sémantique S
j	indice de la valeur d'un texton
O	Ensemble de textons d'une image dont les valeurs sont les indices des codewords après quantification des descripteurs SIFT
i	Indice d'une image dans un lot d'apprentissage associé à un concept spécifique
k	Indice d'un modèle dans un mélange associé à un concept spécifique
S_l	Région sémantique d'indice l trouvée dans une image
Z	Variable latente associée à un mélange de modèles
n_s	Nombre d'images donné par l'utilisateur pour l'apprentissage d'un mélange associé au concept s
M_s	Modèle de mélange associé à la sémantique s
X_s	Lot d'images donné par l'utilisateur pour le concept s
X_{si}	i-ème image du lot d'apprentissage X_s

Fig. C.1 – Notations utilisées dans ce chapitre

C.1.2 Mélange de modèles associé à un concept sémantique.

Un mélange de modèles génératifs est associé à chaque concept s. Nous détaillons dans cette section, étant donné une image d'indice i, comment est généré l'ensemble O_i des textons qui la composent. Tout d'abord, une variable latente Z est tirée dont la valeur est comprise dans l'ensemble discret $\{1, ..., K_s\}$ correspondant à l'indice du modèle qui est choisi pour générer les données. K_s peut ainsi être vu comme la complexité du mélange pour le concept s. Les paramètres pour ce modèle sont les probabilités p_j^{ks} de génération du texton de valeur j par le modèle k du mélange, et les probabilités à priori de chaque modèle $\pi_{ks} = P(Z = k)$, les paramètres de taille de la région pour chaque modèle λ_{ks}. Le nombre total de paramètres est ainsi $N(k + 1)$.
Plus précisement, on suppose ainsi qu'une région sémantique d'indice i et associée au concept s génère l'ensemble de textons O_i de la façon suivante :
- Le modèle k est choisi avec probabilité π_{ks}.
- Le nombre N_i de textons de la région est généré avec une loi de Poisson de paramètre λ_s.
- Chaque texton de la région est choisi indépendemment des autres avec probabilité p_j^{ks}, où j correspond à la valeur du texton.

Soit $\{N_{i1}, ..., N_{iN}\}$ l'histogramme des valeurs des textons au sein de la région i. La probabilité de génération de O_i est donnée par :

$$P(O = O_i|s, Z = k) = Poiss_{\lambda_{ks}}(N_k)\pi_{ks}\prod_{j=1}^{N}(p_j^{ks})^{N_{ij}} \tag{C.3}$$

En conditionnant sur les valeurs possibles de la variable latente Z :

$$P(O = O_i|s) = \sum_{k=1}^{K_s} P(Z = k)P(O = O_i|s, Z = k)$$

Par définition, $P(Z = k) = \pi_{ks}$. De plus, $P(O = O_i|s, Z = k)$ est la probabilité de génération des textons étant donné le concept s et la variable latente.
En remplaéant cette probabilité par son expression dans l'équation C.3, on peut écrire :

$$P(O = O_i|s) = \sum_{k=1}^{K_s} \pi_{ks} Poiss_{\lambda_{ks}}(N_i)\prod_{j=1}^{N}(p_j^{ks})^{N_{ij}} \tag{C.4}$$

La vraisemblance du lot d'observation O est ainsi exprimée comme un mélange de modèles.

C.2 Apprentissage du modèle

Nous supposons ici qu'une segmentation annotée de la base d'apprentissage est fournie.

C.2.1 Expectation-Maximization

Dans cette section, nous supposons fixé le nombre K_s de lots de paramètres correspondant au concept s, l'algorithme présenté ici est utilisé pour estimer la valeurs des K_s lots de paramètres maximisant la vraisemblance $P(X_s|M_s)$. Ces paramètres sont estimés en utilisant l'algorithme Expectation-Maximization (EM). On suppose l'indépendance entre les n_s images X_{si} du lot d'apprentissage, par conséquent :

$$P(X_s|M_s) = \prod_{i=1}^{n_s} P(X_{si}|M_s) \tag{C.5}$$

Soit z_{ik} la variable dont la valeur vaut 1 si $Z = k$ pour l'image i et la quantité $\gamma_k(X_{si}) = E_{Z|X_{si},M_s}(z_{ik})$, où k est l'indice du modèle, et i l'indice de l'image dans le lot d'apprentissage. Cette quantité peut être interprétée comme la correspondance du modèle k pour l'image i relativement aux autres modèles. En prenant le logarithme de l'expression C.5, et en conditionnant par rapport aux valeurs possibles de la variable latente, on obtient, N_i étant le nombre de textons dans le patch i :

$$logP(X_s|M_s) = \sum_{i=1}^{n_s}\sum_{k=1}^{K} z_{ik} log(\pi_k \lambda_{ks}(N_i) \prod_{j=1}^{N} (p_j^{ks})^{N_{ij}}) \tag{C.6}$$

Oo'u z_{ik} est une variable qui vaut 1 si la variable latente $Z_i = k$, c'est-à-dire la variable latente Z vaut 1 pour le patch i, et où N_{ij} est le nombre de textons de type j présents dans l'image i. En prenant l'espérance par rapport à X_s et aux paramètres du modèle M_s de l'équation C.6, on obtient :

$$E_{Z|X_s,M_s}(logP(X_s|M_s)) = \sum_{i=1}^{n_s}\sum_{k=1}^{K} E(z_{ik}|X_s, M_s) log(\pi_k \lambda_{ks}(N_i) \prod_{j=1}^{N} (p_j^{ks})^{N_{ij}})$$

En posant $\gamma_k(X_s) = E(z_{ik}|O_i, M_s)$, on obtient la formule :

$$E_{Z|X_s,M_s}(logP(X_s|M_s)) = \sum_{i=1}^{n_s}\sum_{k=1}^{K} \gamma_k(X_s) log(\pi_k \lambda_{ks}(N_i) prod_{j=1}^{N} (p_j^{ks})^{N_{ij}})$$

L'algorithme EM utilise les deux étapes suivantes pour trouver un maximum local de la vraisemblance :

– étape E : Calcul de $\gamma_k(X_{si})$, pour tout modèle k et toute image i, en utilisant la loi d'inversion de Bayes :

$$\gamma_k(X_{si}) = \frac{\pi_k \prod_{j=1}^{N} (p_j^{ks})^{N_{ij}}}{\sum_{m=1}^{K_s} \pi_m \prod_{j=1}^{N} (p_j^{ms})^{N_{ij}}}$$

Cette expression est écrite comme la probabilité de génération conditionnellement au modèle k sur la vraisemblance des observations dans l'image i. Cela semble logique, étant donné que l'interprétation de la quantité

$\gamma_k(X_{si})$, comprise entre 0 et 1, est une mesure de l'adéquation du modèle k à l'histogramme des textons de l'image i.
– étape M : maximisation de $E_{Z|X_s,M_s}(logP(X_s, Z|M_s))$. La méthode des multiplicateurs de Lagrange est utilisée pour maximiser cette quantité. La formule de mise-à-jour des paramètres est donnée comme suit :

$$p_j^{ks} = \frac{\sum_{i=1}^{n_s} \gamma_k(X_{si}) N_{ij}}{\sum_{i=1}^{n_s} \gamma_k(X_{si}) N_i} \pi_{ks} = \frac{\sum_{i=1}^{n_s} \gamma_k(X_{si})}{n_s} \lambda_{ks} = \sum_{i=1}^{n_s} \gamma_k(X_{si}) N_i \tag{C.7}$$

Notons que l'estimation de p_j^{ks} correspond, comme on pouvait s'y attendre, au rapport des occurrences de textons de valeur j présents dans le lot d'apprentissage sur le nombre total de textons, pondéré par les quantités $\gamma_k(X_{si})$. Les paramètres de probabilité a priori π_{ks} ont également une interprétation très intuitive comme le rapport des quantités $\gamma_k(X_{si})$ sur le nombre total d'images de la base.

C.2.2 Apprentissage non supervisé des paramètres

C.2.2.1 Méthode employée

Nous supposons que, pour chaque concept d'indice $s \in \{1, ..., n\}$, l'utilisateur fournit un lot X_s de n_s imagettes découpées manuellement pour l'apprentissage des paramètres du mélange qui lui est associé. La procédure d'apprentissage est alors la suivante :
– Pour K variant de 1 à n_s :
  – Les paramètres du modèle sont estimés à partir des formules C.7, le modèle obtenu est noté $M_{s,K}$. La vraisemblance de la base de donnée, notée $P_K(X_s|M_{s,K})$ est alors calculée.
  – La complexité stochastique $CS(X_s, M_{s,K})$ est calculée.
– La complexité K choisie est celle minimisant la complexité stochastique :

$$K_s = argmin_{K \in \{1,...,n_s\}} CS(X_s, M_{s,K})$$

C.3 Annotation d'images

C.3.1 Méthode d'annotation

I étant une image à annoter, et les paramètres des mélanges de modèles pour chacun des concepts ayant été estimés, trouver l'ensemble optimal de régions sémantiques $G_I = \{S_1, ..., S_{m_I}\}$ parmi l'ensemble $\mathbf{G}$ de toutes les configurations possibles est un problème très complexe. En effet, le cardinal gigantesque de $\mathbf{G}$ rend impossible toute recherche exhaustive. C'est pourquoi nous détaillons ici un algorithme qui explore en temps raisonnable un chemin dans l'ensemble $\mathbf{G}$. Le principe de cet algorithme est de partir d'une configuration initiale qui est complexe et de la simplifier en fusionnant itérativement des régions voisines en choisissant à chaque étape la fusion qui optimise la vraisemblance. L'algorithme s'arrête lorsqu'il ne reste plus qu'une seule région pour toute l'image. Cet

algorithme est dit glouton car il choisit à chaque étape la meilleure fusion au sens du maximum de vraisemblance. On n'autorise pas de "retour" dans le chemin exploré parmi toutes les configurations de régions sémantiques. Ainsi, cet algorithme peut très bien fournir un simple optimum local de la vraisemblance. Nous détaillons à présent plus en détail les trois étapes de l'algorithme :

- Initialisation de l'algorithme : Chaque texton l de l'image I est lié à un concept en prenant en compte sa valeur, ainsi que la valeur des textons de son voisinage $NE(l)$ en choisissant le concept s qui minimise la quantité (cf équation C.4) :

$$P(NE(l)|S=s) = \sum_{j=1}^{k_s} \pi_{js} \prod_{l \in NE(l)} p_{v(l)}^{js}$$

  où $v(l)$ est la valeur du texton l. Le voisinage $NE(l)$ est défini comme l'ensemble des textons contenu dans un carré centré en l et dont le cété est de taille t. Ensuite, les régions sémantiques sont créees en définissant les régions 4-connexes de textons reliés au même concept. On obtient ainsi un lot initial de régions sémantiques G_0. La vraisemblance $P(X_I|G_0)$ est alors calculée (cf Equation C.2). Notons que plus la valeur de t est grande, moins il y a de régions dans G_0, et plus le chemin exploré dans $\mathbf{G}$ est réduit.
- Soit i le nombre d'itérations ayant déja été effectuées dans cette boucle. Tant que le nombre de régions est supérieur à 1 :
  - On considère toutes le fusions possibles entre régions sémantiques adjacentes
    - Pour chaque fusion possible, il est nécessaire de relier un concept a la région qui a été créée. On calcule ainsi les n vraisemblances possibles pour chacun des n concepts qu'on peut lui assigner. Pour chacun de ces cas, si des régions sémantiques sont adjacentes et ont le même concept, elles sont fusionnées.
  - La configuration maximisant la vraisemblance est gardée et notée G_i.
- On garde la configuration maximisant la vraisemblance sur tous les ensembles G_i trouvés à chaque itération

Le nombre d'itérations possibles est inférieur à $card(G_0)$, à savoir le nombre de régions sémantiques trouvées lors de l'initialisation. En effet, à chaque itération, au moins deux régions sont fusionnées, on a ainsi : $card(G_i) \leq card(G_i) - 1$, ce qui nous assure que l'algorithme termine en un nombre fini d'itérations.

C.3.2 Evaluation visuelle

Les images exemples étant fournies pour chaque concept, nous calculons à partir de ces imagettes les paramètres de génération et les paramètres de Poisson par la méthode d'apprentissage à données complètes exposée dans la section précédente. Les paramètres d'interaction entre régions voisines sont initialisés comme étant équiprobables. Une partie de la base de données d'image est ensuite utilisée pour faire un apprentissage à données incomplètes des paramètres, l'algorithme 3.3.4 est alors mis en œuvre pour estimer plus précisement les paramètres. Nous avons fait des évaluations visuelles pour les deux bases d'images à disposition.

C.3.2.1 Images Quickbird

Présentation de la base de données La base d'image Quickbird à disposition consiste en 16 images 16000*16000 et couvre un ainsi une aire de 11km de cété centrée sur la municipalité de Pékin. Les concepts suivants sont utilisés pour faire une évaluation visuelle : zone urbain dense, zone résidentielle pavillonaire, zone industrielle, nœud autoroutier, zone résidentielle grand-ensemble, zone commerciale, chantier, terrain vague, champs, serres, lac.
La base d'images exemple contient environ 140 imagettes de taille variant de 400×400 à 1000×1000. Une validation croisée a été effectuée en prenant 80% de la base pour l'apprentissage et 20% pour le test. Le nombre d'imagettes n'étant pas identique pour chaque classe, nous avons pris à chaque fois 80% d'imagettes de chaque classe pour l'apprentissage et 20% pour le test. Ceci nous permet d'éviter un problème combinatoire. Nous avons un résultat de $96,4\%$ de bonne classification.

Résultats Nous avons fait des évaluations visuelles en utilisant la base d'imagettes pour l'apprentissage des modèles, et une image de test à annoter avec les concepts introduits. Les résultats semblent satisfaisants. Les zones mal annotées correspondent généralement à des zones ne correspondant à aucun des concepts introduits. Pour améliorer les résultats, il serait nécessaire de laisser la possibilité au système de ne pas annoter certaines zones avec une certaine pénalité. On pourrait ainsi permettre à l'utilisateur de fixer un paramètre de pénalité, permettant ainsi de choisir entre une annotation très complète de l'image mais pouvant comporter des annotations peu fiables, et une annotation très fiable, mais pouvant comporter beaucoup de zones non annotées.
Remarquons que les classes choisies sont très texturées. On pourrait ainsi s'interroger sur la pertinence de caractéristiques de texture, qui pourrait vraisemblablement donner de bons résultats. Il serait ainsi intéressant de voir la différence entre la sortie du système proposée ici et la sortie d'une classification fenêtre par fenêtre (d'une taille par exemple 128) utilisant des caractéristiques de texture ou encore les caractéristiques utilisées ici. Une telle expérience permettrait de mettre en avant la "plus-value" du système qui aggrège des informations locales pour les relier au concept.

C.3.2.2 Images SPOT5

Présentation de la base de données La base d'images SPOT5 qui a été traitée ici comporte des villes diverses, permettant d'avoir à disposition des paysages variées. Cependant, les images utilisées jusqu'à présent sont toutes des villes françaises : Nimes, Paris, Marseilles, Angers, Nice. Cela nous permet d'avoir une "vérité-terrain" cartographique de ces villes en les superposant avec des cartes IGN. Une base d'environ 200 imagettes exemples ont ainsi été extraites sur ces 5 villes (voir C.3.2.2), de taille variant de 250×250 à 1000×1000.

	ZI	ZR	CV	ZC	ZB	C	Ch	ZM
ZI	29	2	0	0	0	0	0	0
ZR	1	25	0	0	0	0	0	0
CV	0	1	15	0	0	0	0	0
ZC	0	0	0	9	0	0	0	0
ZB	0	0	0	0	12	0	1	0
C	0	0	0	0	0	8	0	
Ch	0	0	0	0	0	0	35	0
ZM	0	0	0	0	0	0	0	17

Fig. C.2 – Matrice de confusion pour les tests de validation croisée sur la base d'image SPOT5. Les notations sont les suivantes : ZI : Zone industrielle. ZR : Zone résidentielle. CV : Centre ville. ZC : Zone commerciale. ZB : Zone boisée. C : Carrière. Ch : champs. ZM : Zone montagneuse.

Validation croisée Nous définissons ici le protocole que nous utilisons pour faire une validation croisée de la base d'imagettes extraites dans les images SPOT5. Nous ne prenons pour cela qu'un sous-ensemble de la base que nous avons à disposition, nous conservons seulement les 8 classes suivantes : zone industrielle, zone résidentielle, centre ville, zone commerciale, zone boisée, carrière, champs, zone montagneuse. Nous utilisons à chaque itération 20% de la base de données comme base de test et 80% comme base d'apprentissage. Les paramètres génératifs ayant été calculés. Chaque image d'indice i de l'ensemble de test est classifiée dans le concept s qui maximise la probabilité à priori :

$$P(O = O_i|s) = \sum_{k=1}^{K_s} \pi_{ks} \prod_{j=1}^{N} (p_j^{ks})^{N_{ij}}$$

La matrice de confusion est présentée en C.3.2.2. On obtient au total $96,7\%$ de bonne classification.

Protocole d'apprentissage Nous choisissons pour effectuer l'apprentissage un ensemble de 60 imagettes parmi la base de données extraites présentées précédemment et un ensemble de 20 images 3000×3000 choisies dans la base d'images SPOT5 de façon à contenir toutes les différentes sémantiques choisies. Notons que certaines des imagettes utilisées. Les 13 sémantiques considérées ici sont listées dans la figure C.3.2.2. Le protocole d'apprentissage non supervisé est celui décrit en 3.3.4 avec 5 itérations au cours desquelles les 20 images 3000×3000 sont segmentées puis les paramètres des modèles de mélange sont estimés.

Analyse des résultats

– Nous voyons tout d'abord sur la figure 3.14 la différence entre une classification effectuée comme maximisation du maximum de vraisemblance pour chaque et la classification obtenue à partir de la méthode proposée.

Concept	**nombre d'imagettes d'initialisation**	**Nombre d'imagettes à la dernière itération**	**nombre de modèles dans le mélange à la dernière itération**
Carrière	5	19	3
Bois	5	16	2
Champs	5	17	1
Montagne	5	24	1
Mer	5	10	1
Aeroport	1	5	3
Centre ville	5	9	2
Marais salant	1	1	1
Zone rurale	5	17	2
Raffinerie	3	5	2
Village	5	27	3
Zone industrielle	5	22	3
Zone résidentielle	10	25	3

– L'image (a) de la figure 3.13 illustre l'apport de la modélisation de l'interaction entre régions. En effet, les concepts "village" et "zone résidentielle" ont des distributions de probabilité très proches. L'interaction entre régions permet de prendre en compte ce problème. Un village a en effet une plus forte probabilité de se situer à proximité d'une zone rurale qu'une zone résidentielle, tandis qu'une zone résidentielle aura une plus forte probabilité de se trouver à proximité d'un centre ville ou d'une zone industrielle qu'un village.
– Les régions associées au concept "aéroport" (voir figure 3.15) correspondent à un taux d'erreur relativement fréquent en terme de "fausses alarmes", même si ce taux peut difficilement être quantifié à présent (voir images (b) et (c) de la figure 3.13). Ce fort taux de fausses alarmes provient probablement de la complexité intrinsèque de cette classe. En effet, la densité de probabilité associée au concept aéroport (voir figure 3.17) est beaucoup plus proche de la densité uniforme que d'autres densité. Par conséquent, ce concept a tendance à annoter par défaut des zones correspondant relativement à aucune autre sémantique. On peut en conclure que le concept "aéroport" correspond à des régions trop complexes pour être pris en charge par un modèle aussi simple, où qu'il serait nécessaire d'introduire un niveau intermédiaire de sémantique pour faire le lien entre un concept d'aussi haut niveau et les caractéristiques de bas-niveau, dépourvues de sémantique. Les régions aéroportuaires contiennent en effet des régions diverses associées chacunes à un concept sémantique : "pistes", "aérogoares","terminaux" .. et le "saut sémantique" que l'on en fait en passant à des caractéristiques de texture, caractéristiques purement symbolique, au

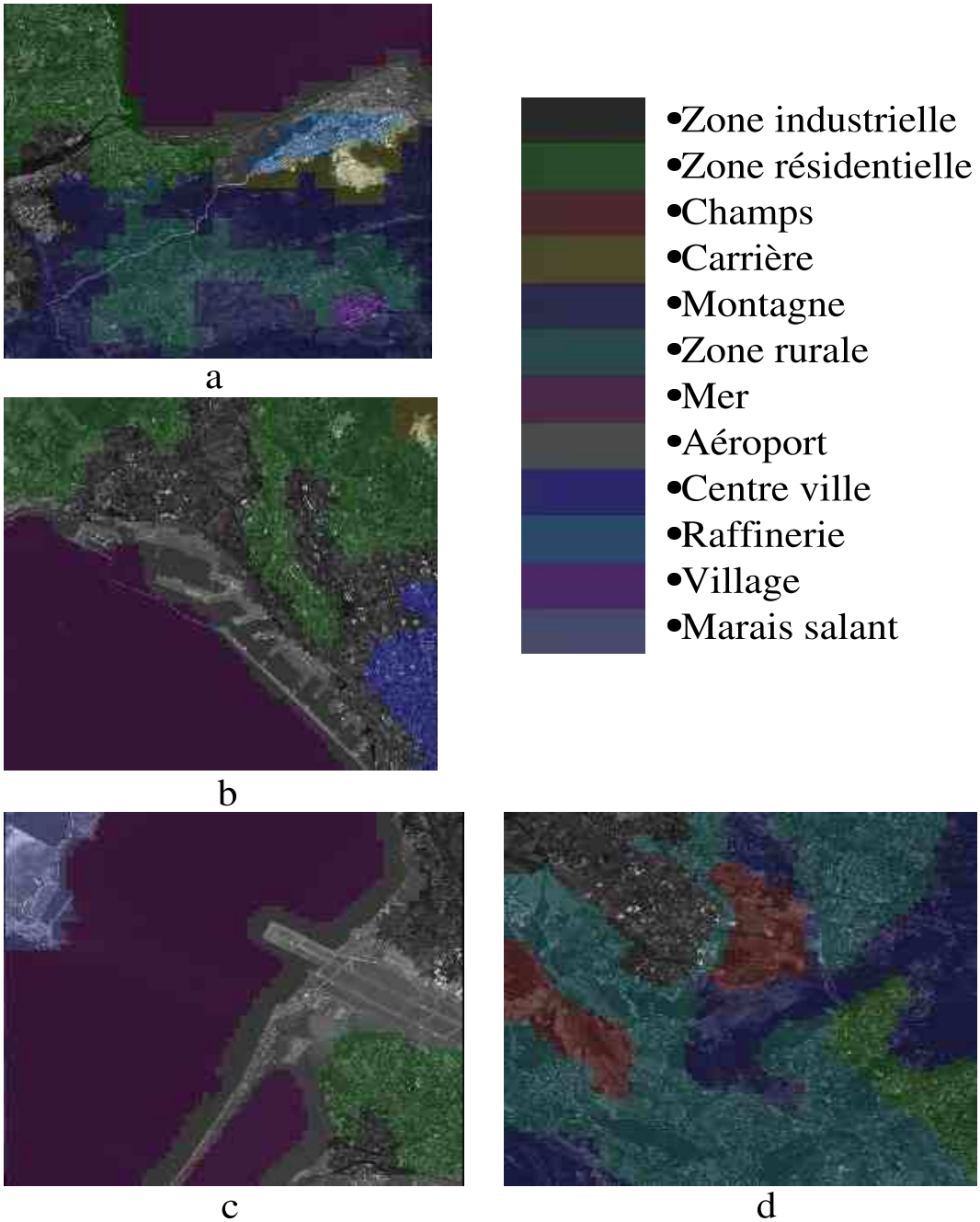
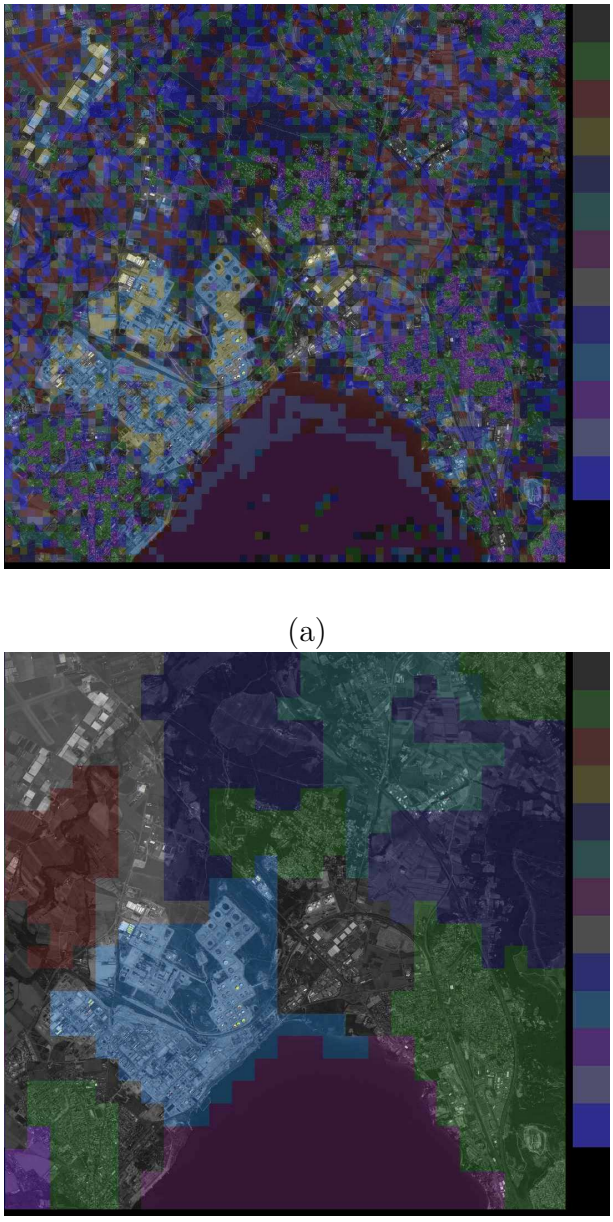


Fig. C.3 – Annotation sémantique d'images SPOT5 3000×3000 textons de Marseille.



(a)

(b)

Fig. C.4 – (a) : image annotée par classification de chaque vecteur de caractéristique. (b) : image annotée par la méthode proposée dans ce chapitre.

Catégorie	**concept**	**nombre d'imagettes**
Végétation	carrière	13
	bois	18
	champs	37
	montagne	17
	prairie	3
Eau	mer	3
	lac	2
	bassin	2
Urbain	aeroport	4
	centre ville	16
	cimetière	1
	marais salant	1
	port de plaisance	10
	port industriel	1
	raffinerie	4
	village	29
	zone commerciale	9
	zone industrielle	36
	zone résidentielle	53

Fig. C.5 – Base de données constituée à partir de la base d'images SPOT5

concept "aéroport" est probablement trop important dans la modélisation que nous employons dans ce chapitre.

Bibliographie

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. of the Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973.

[2] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. Tilton. Learning bayesian classifiers for scene classification with a visual grammar. *Geoscience and remote sensing*, 43(3) :993–1022, 2005.

[3] D. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilités de Saint-Flour XIII 1983*, pages 1–198. Springer, 1985.

[4] P. Auer. On learning from multi-instance examples : Empirical evaluation of a theoretical approach. In *Proc. of international Conf. Computer Vision*, volume 2, 1997.

[5] B. Bachimont. Bibliothèques numériques audiovisuelles. *Document numérique*, 2-3, 1998.

[6] R. Balian. *Cours de physique satistique de l'école polytechnique*, volume 1. Ellipse, 1982.

[7] K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3 :1107–1135, 2003.

[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures, 2002.

[9] J. Bellegarda. Latent semantic mapping. *Signal Processing*, 22(5) :70–80, 2005.

[10] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. 2001.

[11] M. Berry and al. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37 :573–595, 1995.

[12] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004.

[13] D. Blei and M. Jordan. Modeling annotated data. In *Proc. of the 26th annual intetational ACM SIGIR conference*, pages 127–134, 2003.

[14] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022, 2003.

[15] A. Boucher and T. Lee. Comment extraire la sémantique d'une image ? In *Proc. of the 3rd International Conference : Sciences of Electronic. Technologies of Information and Telecommunication*, 2005.

[16] J. Bower. *Human Associative memory*. Winston and Sons, 1973.

[17] M. Breal. *Essai de sémantique (science des significations)*. Hachette, 1897.

[18] J. Bresnan and R. Kaplan. *a formal system for grammatical representation*. MIT Press, Cambridge, Massachussetts, 1981.

[19] J. Buckner, M.Pahl, and O.Stahlhut. Geoaida-a knowledge based automatic image data analyser for remote sensing data. In *Second International ICSC Symposium AIDA*, Bangor, Wales, U.K., 2000. CIMA.

[20] K. Burnham and D. Anderson, editors. *Model Selection and Multimodel Inference*. Springer-Verlag, 2002.

[21] R. Burton. Semantic grammar. an engineering technique for constructing natural language understanding systems. *Technical Report 3353, BBN, Cambridge, Massachusetts*, 1976.

[22] M. Campedel, B. Luo, H. Maître, E. Moulines, M. Roux, and I. Kyrgyzov. Indexation des images satellitaires. Technical report, École Nationale Supérieure des Télécommunications, 2004.

[23] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent. Unsupervised performance evaluation of image segmentation. *EURASIP J. Appl. Signal Process.*, 1 :217–217, 2006.

[24] H. Chen, Z. Xu, Z. Liu, and S. Zhu. Composite templates for cloth modeling and sketching. *Proc of the IEEE Conference of Pattern Recognition on Computer Vision*, June 2006.

[25] H. Chen, Z. Xu, Z.Q. Liu, and S.C Zhu. A high resolution grammatical model for face representation and sketching. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

[26] Z. Chi and S. Geman. Estimation of probabilistic context free grammar. *Computational linguistics*, 24(2) :299–305, 1998.

[27] J. Courtès. *La sémiotique du langage*. Armand Colin, 2007.

[28] D. Cruse. *Meaning in languages : an introduction to Semantics and Pragmatics*. Oxford University Press, 2000.

[29] F. Cutzu, R. Hammoud, and A. Leykin. Distinguishing paintings from photographs. *Computer Vision and Image Understanding*, 100(3) :249–273, 2005.

[30] B. de Finetti. *Theory of probability*, volume 1-2. John Wiley Ltd, Chichester, 1975.

[31] F. de Saussure. *Écrits de linguistique générale. Texte établi et édité par Simon Bouquet et Rudolf Engler*. Gallimard, 2002.

[32] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic indexing. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.

[33] A. Dempster. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B39, 1977.

[34] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proc. of the 11th international conference on World Wide Web*, pages 662–673, New York, 2002. ACM.

[35] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29 :103–137, 1997.

[36] D. Donoho, M. Vetterli, and R. Devore. From volumes to view, an approach to 3d objects recognition. *IEEE Transactions Information Theory*, 6 :2435–2476, 1998.

[37] P. Dyugulu, K. Barnard, and D.F.N Freitas. Object recognition as machine translation : learning a lexicon for a fixed image vocabulary. *Proc. of the IEEE European Conference on Computer Vision*, 2002.

[38] M. Eisen and al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25) :14863–14868, 1998.

[39] C. Fellbaum, editor. *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998.

[40] S. Feng, R. Manmatha, and V. Lavrenko. Mutliple bernouilli relevance models for image and video annotation. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2004.

[41] C. Fillmore. *The case for case, Universals in Linguistic Theory*. Rinehart and Winston Inc., 1968.

[42] D. Forsyth and M. Fleck. Body plans. In *Proc. of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 678, Washington, DC, USA, 1997. IEEE Computer Society.

[43] C. Fraley and A. Raftery. How many clusters ? which clustering method ? answer via model-based cluster analysis. *The Computer Journal*, 41(8) :578–588, 1998.

[44] J. Friedman. Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5) :1189–1232, 2001.

[45] J. Friedman and T. Hastie. Additive logistic regression : a statistical view of boosting. *Annal of statistics*, 38(2) :337–374, 2002.

[46] N. Friedman. The bayesian structural em algorithm. In *Proc. of the 14th Conference on Uncertainty in AI*, 1998.

[47] K. Fu. *Syntactic Pattern recognition and applications*. Prentice Hall, 1982.

[48] G. Furnas, T. Landauer, L. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication : an analysis and a solution. *Communications of the ACM*, 30 :964–971, 1987.

[49] S. Galliano and al. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proc. Language Evaluation and Resources Conference*, 2006.

[50] A. Gelfand and A. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 685(410) :398–409, 1990.

[51] S. Geman and D. Potter. Composition system. *Quaterly of applied mathematics*, 60 :707–736, 2002.

[52] J. Gennari and al. The evolution of protégé : an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1) :89–123, 2003.

[53] S. Golder and B. A. Huberman. The structure of collaborative tagging systems, Aug 2005.

[54] U. Grenander. *General Pattern theory*. Oxford University Press, 1993.

[55] T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Proc. of the 24th Annual Conference of the Cognitive Science Society*, 2002.

[56] F. Han and S.Chun Zhu. Primal sketch : integrating texture and structure. *Proc. of the IEEE International Conference on Computer Vision*, 2005.

[57] R. Haralick. Statistical and structural approaches to texture. 67(5) :786–804, 1979.

[58] J. Hare, P. Lewis, Peter G. Enser, and C. Sandom. Mind the gap : another look at the problem of the semantic gap in image retrieval. *Management and retrieval*, 6073, 2006.

[59] L. Hjelmslev. *Prolégomènes à une théorie du langage*. Minuit, 1968.

[60] T. Hofmann. The cluster-abstraction model : Unsupervised learning of topic hierarchies from text data. In *In IJCAI*, pages 682–687, 1999.

[61] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the Special Interest Group in Information Retrieval*, pages 25–44, 1999.

[62] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning journal*, 42(1) :177–196, 2001.

[63] H. Maitre I. Kyrgyzov and M. Campedel. Kernel mdl to determine the number of clusters. 2007.

[64] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th international ACM SIGIR Conference*, pages 119–126, 2003.

[65] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2006.

[66] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine learning*, 37 :183–233, 1999.

[67] A. Joshi and L. Levy. Tree adjunct grammars. *Journal of Computer and System Sciences*, 1975.

[68] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290 :91–97, 1981.

[69] A. Lee K. Pedersen and D. Mumford. The non-linear stastics of high-contrast patchesin natural images. *IJCV*, 54 :83–103, 2003.

[70] G. Kanisza. *Organization in vision.* Praeger, 1974.

[71] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3), 1987.

[72] A. Kiryakov and al. Semantic annotation, indexing and retrieval. *Journal of Web Semantics*, 2(1) :49–79, 2004s.

[73] H. Kucera and N. Francis. *Computational Analysis of Present-Day American English.* Brown University, 1962.

[74] H. Kuck, P. Carbonetto, and O. De Freitas. A constrained semi-supervised learning approach to data association. In *Proc. of the European Conference for Computer Vision*, pages 1–12. Springer, 2004.

[75] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79–86, 1951.

[76] F. Kummert, H. Niemann, G. Sagerer, and S. Schroder. *Werkzeuge zur modellgesteuerten Bildanalyse und Wissensakqusition–Das System ERNEST*, pages 556–570. Springer-Verlag, 1987.

[77] D. Kunz, K. Schilling, and T. ogtle. A new approach for satellite image analysis by means of a semantic network, 1997.

[78] G. Lakoff. *Women, Fire and Dangerous things. What Categories Reveal about the Mind.* University of Chicago Press, 1987.

[79] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantic of pictures. *Proc. of the Conference on Advances in Neural Information Processing Systems*, 2003.

[80] E. Lebarbier and T. Mary-Huard. Le critère bic : fondements théoriques et interprétation, 2004.

[81] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :1075–1088, 2003.

[82] W. Li. Random texts exhibits zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6), 1992.

[83] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[84] J. Lyons. *Eléments de sémantique.* Larousse Université, Paris, 1978.

[85] B. Mandelbrot. *Information Theory and Psycholinguistic.* Basic Books, 1968.

[86] D. Marr. *Vision.* Freeman Publisher, 1983.

[87] A. Meillet. Comment les mots changent de sens. *L'Année sociologique*, 9, 1905.

[88] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR*, 2003.

[89] C. Millet. *Annotation automatique d'images : annotation cohérente et création automatique d'une base d'apprentissage*. PhD thesis, École nationale supérieure des télécommunications, 2008.

[90] F. Min, J. Suo, S. Zhu, and N. Sang. An automatic portrait system based on and-or graph representation. In Yuille et al., editor, *Energy Maximization Methods in Computer Vision and Pattern Recognition*, pages 184–197. Springer, August 2007.

[91] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM International Conference on Multimedia*, pages 271–274, November 2003.

[92] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proc. of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[93] G. Mounin. *La sémantique*. Seghers, 1972.

[94] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of computer vision*, 14 :5–24, 1995.

[95] S. Newsam, L. Wang, S. Bhagavathy, and B.S. Manjunath. Using texture to analyze and manage large collections of remote sensed image and video data. *Applied optics*, 43(2) :210–217, Jan 2004.

[96] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification, 1999.

[97] M. Nilsson. The semantic web : How rdf will change learning technology standards. Technical report, Center for User-Oriented IT-design, Royal Institute of Technology, Stockholm, 2001.

[98] M. Oder, H. Rehrauer, K. Seidel, and M. Datcu. Interactive learning and probabilistic retrieval in remote sensing image archives. *Geoscience and Remote Sensing*, 38(5) :2288–2298, September 2000.

[99] Y. Ohta, T. Kanade, and T. Sakai. An analysis for scenes containing objects with substructures. In *Proc. of the 4th International Joint Conference on Pattern Recognition*, pages 752–754, Kyoto, 1978.

[100] Z. Pecenovic. *image retrieval using latent semantic indexing*. PhD thesis, Ecole polytechnique fédérale de Lausanne, 1997.

[101] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda : A generative model for particular object discovery. In *Proc. of the British Machine Vision Conference*, 2008.

[102] V. Prince and Y. Kodratoff. Revue des nouvelles technologies de l'information. *Journal of Information Science*, 32(10) :1–14, February 2007.

[103] R. Quillian. *M.R. Semantic memory*. PhD thesis, Carnegie- Mellon U, February 1966.

[104] F. Rastier. *Sémantique et recherches cognitives*. PUF, Paris, 1991.

[105] F. Rastier. Histoire de la sémantique, 1890-1990. *Histoire, Epistémologie, Langage*, 15(1) :153–187, 1993.

[106] F. Rastier. Ontologie(s). *Revue des sciences et technologies de l'information*, 18(1) :15–40, 2004.

[107] J. Rekers and A. Schurr. A parsing algorithm for context sensitive graph grammars. *Leiden Univ*, 1995.

[108] I. Rish. An empirical study of the naive bayes classifier.

[109] J. Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978.

[110] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific, 1989.

[111] E. Rosch. Semantic representation of semantic categories. *Journal of experimental psychology*, 104(3) :192–233, 1975.

[112] B. Russell and al. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[113] R. Schapire. The boosting approach to machine learning : an overview. *MSRI Workshop on nonlinear Estimation and Classification*, 2002.

[114] R. Schwartz and al. Language understanding using hidden understanding models. In *Proc. of the International Conference on Spoken Langugage Processing*, pages 997–1000, 1996.

[115] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[116] E. Segal, D. Koller, and D. Ormoneit. Probabilistic abstraction hierarchies. In *Advances in Neuronal Information Processing Systems*, volume 14. MIT Press, 2001.

[117] C. Shannon. A mathematical theory of communication. *Bell Syst Technology*, 27 :379–423, 1948.

[118] K. Sheldonand and R. Simmons. Syntactic dependence and the computer generation of coherent discourse. *Mechanical Translation*, 1963.

[119] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of the 17th international conference on World Wide Web*, 2008.

[120] E. Simoncelli and W.T. Freeman. Shiftable multi-scale transforms. *IEEE Transactions Information Theory*, 38(2) :587–607, 1992.

[121] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. of the International Conference on Computer Vision*, 2005.

[122] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[123] J. Sivic and A. Zisserman. Video google : A text retrieval approach to object matching in videos. In *ICCV '03 : Proc. of the Ninth IEEE International Conference on Computer Vision*, page 1470, Washington, DC, USA, 2003. IEEE Computer Society.

[124] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349–1380, 2000.

[125] M. Smith, C. Welty, and D. McGuiness. Owl web ontology language guide. w3c recommandations 10 february 2004, 2004.

[126] J. Sowa. Semantic networks. revised and extended version of an article originally written for the encyclopedia of artificial intelligence, edited by stuart c. shapiro, wiley, 1987, second edition, 1992, 2006.

[127] S. Staab and al. An annotation framework for the semantic web. In *Proc. of the First Workshop on Multimedia Annotation*, pages 30–31, 2001.

[128] N. Sugiura. Further analysts of the data by akaike' s information criterion and the finite corrections – further analysts of the data by akaike' s. *Communications in Statistics - Theory and Methods*, 7(1) :13–26, 1978.

[129] M. Szummer and R. Picard. Indoor-outdoor image classification. In *Proc. of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, page 42, Washington, DC, USA, 1998. IEEE Computer Society.

[130] I. Tamba. *La sémantique*. Puf, 2005.

[131] S. Todorovic and N. Ahuja. Extracting subimages of unknown category from a set of images. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognitio*, 2006.

[132] Z. Tu, X. Xhen, A. Yuille, and S.C Zhu. Image parsing : unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2) :113–140, 2005.

[133] A. Vailaya, A. Jain, and H. J. Zhang. On image classification : City vs. landscape. In *Proc. of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, page 3, Washington, DC, USA, 1998. IEEE Computer Society.

[134] M. Vargas-Vera and al. Mnm : Ontology driven tool for semantic markup. In *Proc. of the Workshop Semantic Authoring, Annotation and Knowledge Markup*. ECAI, 2002.

[135] N. Vasconcelos and G. Carneiro. Formulating semantic image annotation as a supervised learning problem. *CVPR*, 5 :163–168, 2005.

[136] N. Vasconcelos, G. Carneiro, and P. Moreno. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions Pattern Intelligence and Machine Analysis*, 29(3) :394–410, 2007.

[137] K. Vezina. Survol du monde de l'indexation d'images. Technical Report 1, École de bibliothéconomie et des sciences de l'information de l'UniversitÃ©é de Montréal, 1997.

[138] P. Vossen. Eurowordnet : a multilingual database for information retrieval. In *Proc. of the DELOS workshop on Cross-language Information Retrieval*, 1997.

[139] W. Wang and I. Pollak. Hierarchical stochastic grammars for classification and segmentation. *IEEE transactions on image processing*, 15(10) :3033–3052, Oct 2006.

[140] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[141] C. Weber and A. Puissant. Une démarche orientée-objets pour extraire des objets urbains sur des images thr. *Société Française de Photogrammetrie et de Télédétection*, 2004.

[142] B. Yao, X. Yang, and S-C. Zhu. Introduction to a large-scale general purpose ground truth database : methodology, annotation tool and benchmarks. In Yuille et al., editor, *Energy Maximization Methods in Computer Vision and Pattern Recognition*, pages 169–183. Springer, August 2007.

[143] D. Yihong. Web semantic annotation using data-extraction ontologies, 2004.

[144] L Younes. *Estimation and annealing for Gibbsian Fields*, volume 2. Annales de l'institut Poincaré, 1988.

[145] L. Younes. *Parametric Inference for Imperfectly Observed Gibbsian Fields, Springer-Verlag Probability Theory and Related Fields*, volume 82. Springer Berlin, 1989.

[146] S.C. Zhu. Embedding gestalt laws in markov random fields. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 21, 1999.

[147] G. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Mass, 1932.