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 sémantique correspondant 'a 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 tout 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'images par le contenu ont utilisé directement ces caractéristiques symboliques qui ont donné quelques 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 pavillonnaire, 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" [9], 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" [68]. Ce "fossé" est une difficulté récurrente en vision par ordinateur, il n'est ni plus ni moins que le problème de liaison entre une description de bas-niveau et une description de haut-niveau d'une image, et c'est ce problème qui

a été traité dans ce travail de thèse.

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 [76]. 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'informations, 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 disciplines : 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. Ce domaine ne pouvant être défini plus précisément que comme l'étude du sens, l'extraction de sémantique dans les bases d'images est généralement défini simplement comme la mise en relation des images de la base avec des concepts sémantiques.

L'originalité principale du travail effectué ici consiste justement à poser le problème de l'extraction de sémantiques 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'images 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 lexicale 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".

L'approche considérée, illustrée figure 1.1, consiste à mettre en relation les modèles statistiques du système et les modèles sémantiques de l'utilisateur. Pour représenter des modèles sémantiques, nous utilisons le formalisme des réseaux sémantiques. à un modèle sémantique peut ainsi être attaché un modèle stochastique dont la structure lui correspond. 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 est conçue pour pouvoir s'appliquer à tout type d'images. Cependant, le choix de travailler avec des images satellitaires s'explique par des raisons de simplicité. En effet, contrairement aux images multimédia, les images satellitaires comportent l'avantage de pouvoir connaître précisément le 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 à 0,6 mètres de résolution de Pékin.

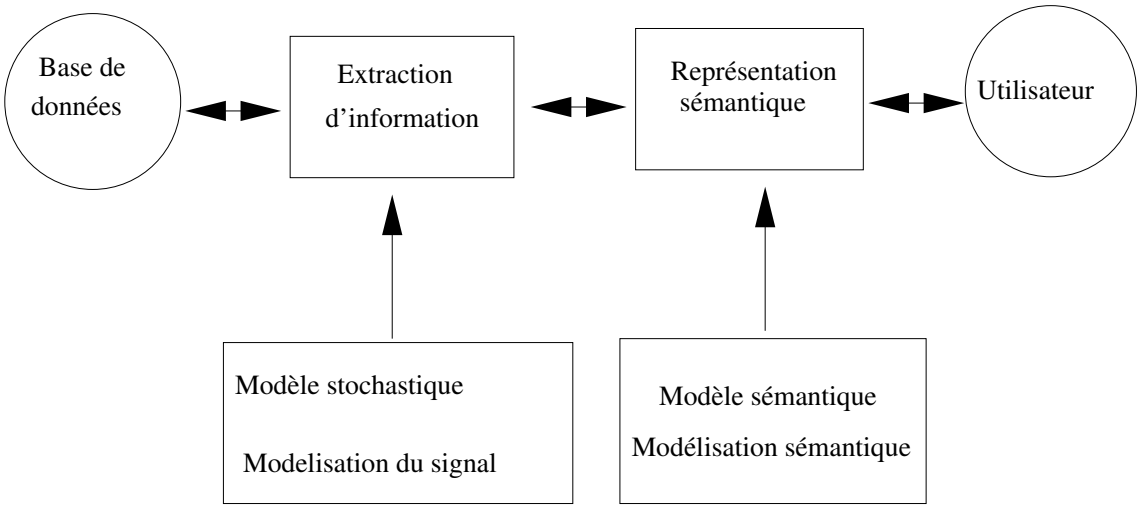


Figure 1.1 – Dualité modèle stochastique/modèle sémantique, proposée par M. Datcu

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 processus de fouille de données. Le deuxième chapitre exposera ensuite un état de l'art des techniques d'inférence traditionnellement utilisées en extraction de sémantiques 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.

Chapitre 2

Sémantique et fouille de données

Le sens est indiscutablement un élément incontournable dans notre utilisation des langues. Cependant, il existe une différence très importante entre appréhender le sens communément perçu d'un mot et l'établir en un objet d'étude linguistique. C'est sans doute la raison pour laquelle la sémantique est apparue assez tardivement en tant que discipline à part entière au sein des sciences du langage. Sa naissance est canoniquement fixée à 1887, année où Michel Bréal inaugure une science des significations dans son *Essai de sémantique*[10] : "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". Le cœur de cible de cette nouvelle *science*, telle que définie par Bréal, est de s'appuyer sur la linguistique pour déterminer des évolutions des significations des mots. Depuis, les objectifs de la linguistique ont considérablement évolué et se sont fortement diversifié.

2.1 Les différentes formes de la sémantique

Toutes les définitions de la sémantique s'accordent sur un même objet d'étude qui est le sens au sein des langues. Cependant, les techniques employées et les domaines d'étude varient considérablement à travers un vaste ensemble de courants théoriques. On distingue en effet quatre périodes principales [70]

— La période évolutionniste (à partir de 1897) : On observe une sémantique inspirée par Darwan qui chercher des lois décrivant l'évolution du sens des mots au cours du temps.

— La période structurale (à partir de 1931) : La sémantique s'inspire de l'étude allemande des champs lexicaux, et tente de compartimenter et structurer un ensemble d'éléments sémiques.

— La période des grammaires formelles (à partir de 1963) : La sémantique analyse les rapports entre les structures syntaxiques et les phrases.
— La période cognitive (à partir de 1978) : La sémantique cognitive vise à rattacher le sens des mots au fonctionnement général du cerveau.

La sémantique cognitive étant assez éloignée du sujet traité ici, nous nous limiterons à détailler le propos des trois première sémantiques :

2.1.1 La période évolutionniste

La sémantique évolutionniste semble fortement influencé par l'objectif du darwinisme : découvrir les lois d'évolution des sens des mots. Elle emprunte aussi sa méthode scientifique d'observation des sens, et l'applique à un grand nombre de dialectes et sur de vastes périodes historiques. Meillet [45] met en lumière trois sortes d'évolutions dans les langues :

— Les changements d'ordre linguistique qui proviennent entre autres de lois phonétiques de modification des formes.
— Des changements d'ordre historique, qu'entraînent des contacts interculturels
— Des changements d'ordre social, tributaires de l'évolution des institutions humaines.

"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" affirme Meillet. Aussi estime-t-il indispensable de définir un système complet de causes qui "s'unissent, agissent et réagissent pour transformer le sens des mots".

2.1.2 La sémantique structurale

La sémantique structurale s'inscrit dans le courant théorique du structuralisme et, à la suite de Saussure, et se fonde sur les deux postulats suivants :

— Autonomie du langage : Le signe linguistique a pour fonction de relier un signifiant (image acoustique) à un signifié (concept).
— Indépendance de la forme : chaque langue constitue un système (une *structure*) et ses unités (sons, mots, sens) tirent à la fois leur essence et leur existence de leurs relations avec les autres unités du même système. L'organisation du système, ou de la structure, dépend des principes complémentaires de sélection (paradigmatique) et de combinaison (syntagmatique).

La notion d'*information sémantique* ([35]) provient de la possibilité de choisir une unité plutôt qu'une autre (information paradigmatique), et de la combiner avec d'autres unités douées de sens (information syntagmatique).

La sélection d'une unité plutôt qu'une autre relève de l'information paradigmatique. La sémantique structurale 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.

L'antonymie Le terme d'antonymie est utilisé 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, l'antonymie diffère de la synonymie par une forte binarité et par une plus grande difficulté à être caractérisée de manière précise. On ne peut pas en effet parler d'un seul type d'antonymie, mais plutôt de quatre types d'opposition :

— 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 visent à d'identifier des structures hiérarchiques inhérentes au lexique, en les dissociant des classifications conceptuelles en genres et espèces. C'est J.Lyons qui forge le néologisme d'hyponymie [35] 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 observe 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*. Cette classification par *superordination* permet un emboîtement successif de classes de plus en plus générales.

méronymie/holonymie : Les termes de méronymie et d'holonymie ont été introduits par A. Cruse [16]. Ils définissent des structurations lexicales hiérarchiques qui sont très différentes de celles induites par les relations d'hyponymie et d'hyperonymie. 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".

Il faut insister sur le fait 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

SÈME / LEXÈME	POUR S'ASSEOIR S1	SUR PIED S2	POUR UNE PERSONNE S3	AVEC DOSSIER S4	AVEC BRAS S5	MATÉRIAUX RIGIDE S6
siège	+	∅	∅	∅	∅	∅
chaise	+	+	+	+	-	+
fauteuil	+	+	+	+	+	+
tabouret	+	+	+	-	-	+
canapé	+	+	-	+	+	+
pouf	+	-	+	-	-	-

Figure 2.1 – Décomposition des sièges en éléments sémiques [55].

même l'impossibilité lexicale, et pas seulement syntaxique, de trouver des termes qui se correspondent termes à termes. En effet, il arrive fréquemment qu'une langue concentre dans un seul item lexical (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.

La thèse de l'autonomie du langage a permis d'envisager la description du sens à partir d'une analyse componentielle ou sémique différentielle, c'est-à-dire une méthode de description du sens des mots qui repose sur la thèse qu'on peut analyser le sens de chaque lexème à partir de composants de sens plus généraux (ou traits sémantiques) dont certains sont partagés par plusieurs lexèmes de la langue. Ainsi, les traits sémantiques sont des traits distinctifs minimaux (c'est-à-dire indécomposables) de sens qui opèrent dans un seul champ lexical et qui servent à structurer ce champs en termes de différents types d'opposition. Leur inventaire doit être fini. Le sème commun à tous les lexèmes est appelé "noyau sémique". L'exemple classique est celui du champ lexical des sièges ([55]) dont le noyau sémique est "pour s'asseoir" et dont les différents sèmes sont "sur pied", "pour une personne", "avec dossier", "avec bras", "matériaux rigides" etc. (voir tableau 2.1)

2.1.3 La sémantique des grammaires formelles

La sémantique des grammaires formelles analyse un langage en s'appuyant sur un ensemble d'états finis, en caractérisant une phrase sur la façon dont sont regroupés ses constituants.

Une grammaire formelle est constituée d'un 4-uplet : (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 $\alpha -> \beta$ 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 terminaux correspondent aux mots constitutifs de la phrase lorsque le processus de génération se termine. Le processus de génération consiste à appliquer à chaque étape 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.

2.2 Réseaux sémantiques

2.2.1 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. 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. La représentation de réseaux sémantiques dépasse le projet de définition d'un simple dictionnaire : les réseaux sémantiques reflètent la façon complexe dont est structurée un champs de 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 nœuds sont des concepts individuels et des arcs étiquetés reliant ces nœuds entre eux.

2.2.2 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 six types de réseaux sémantiques que l'on peut considérer comme les plus couramment utilisés :

— Les réseaux sémantiques de définition (Definitional network) utilise 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 utilisent l'implication comme relation de base entre nœuds connectés. Ils peuvent être utilisés pour représenter des causes ou des inférences.

— Les réseaux d'apprentissage construisent ou étendent leur représentions en acquérant de la connaissance. La nouvelle connaissance peut changer l'ancien réseau en ajoutant ou supprimant des nœuds, ou en modifiant des valeurs numériques, appelés poids, associés avec les nœuds et les arcs.
— Les réseaux exécutables contiennent certains mécanismes, comme des procédures associées, qui peuvent mettre en œuvre des inférences ou chercher des associations.
— Les réseaux hybrides combinent deux ou plusieurs des techniques précédentes, soit dans un simple réseau, soit des réseaux interagissant les uns avec les autres.

2.2.3 Ontologies

La définition la plus générale que l'on puisse faire de l'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 dans un but de clarification.

Il existe plusieurs types d'ontologies et ses applications sont diverses : elles sont notamment exploitées pour élaborer la structure d'une base de données, dans le développement de logiciels, ainsi que dans le Web sémantique. Nous dressons ici les principaux types d'ontologie :

Ontologie d'un domaine L'ontologie du domaine est fonctionnelle et est utilisée pour représenter un domaine (de la génétique, l'aéronautique, etc.) sous forme de base de connaissances. Elle présente les concepts-clés, les attributs, les instances relatifs au domaine. Elle permet à une communauté de se mettre d'accord sur un vocabulaire et une structuration communs. Certains éditeurs existent pour construire une ontologie d'un domaine, donc le plus utilisé est "Protégé".

Ontologie informatique Les ontologies informatiques sont des représentations graphiques qui permettent de structurer un corpus de connaissances sous une forme utilisable par une machine. Elles représentent un ensemble organisé de concepts, dont les relations peuvent être des relations sémantiques et/ou des relations de composition et d'héritage (au sens objet). Des outils existent pour édifier cette structuration, dont le plus connu est le langage UML (Unified Modeling Language), qui est un formalisme permettant de construire des ontologies informatiques.

2.3 Sémantique et fouille d'images

2.3.1 Fouille d'images

Les méthodes de fouille d'images (data-mining) ont pour but de *faire émerger* la sémantique d'un ensemble de données en dégageant du sens de cette base de

données. Les données sont tout d'abord représentées à l'aide de différents outils : graphe, arbre, vecteur de nombres, tableau etc. Ensuite, il est impératif de faire intervenir un expert humain du domaine pour situer la sémantique extraite et lui donner de la valeur.

2.3.2 Définition des concepts d'annotation

La définition des concepts nécessite :
— la définition de termes sémantiques qui seront utilisés pour décrire une image
— la définition des précédents termes en caractéristiques que nous pouvons extraire de l'image.

Les deux éléments sont importants car la définition seule d'un vocabulaire ne permettra pas au système de retrouver ces termes dans l'analyse automatique de l'image. Dans la plupart des problématiques d'extraction de sémantiques dans le texte et dans l'image. Il s'agit d'attacher un certain nombre de concepts à un texte, à une image, ou à une partie d'une image de façon à permettre à l'utilisateur de naviguer efficacement dans la base de données à travers des requêtes sur ces concepts sémantiques. La définition de ces concepts pour l'annotation d'une image est donc un élément particulièrement important du système, et dépend à priori du corpus d'images, ainsi que de la perspective avec laquelle on souhaite le traiter : comment pourrait on associer les labels "intérieur" ou "extérieur" à une échographie d'un fœtus ?

Nous nous plaçons ici dans le cadre d'étude d'une base d'images satellitaires, ce qui définit précisément la nature du corpus étudié. Cependant, la description en termes conceptuels de ce corpus dépend également de la résolution de ces images, et de l'objectif poursuivi. Celui-ci dépend de la perspective dans laquelle se place l'utilisateur : agronomique, cartographique, prévention de catastrophes etc. Il semble difficilement envisageable de définir un vocabulaire qui permette de satisfaire à n'importe quelle application. Notons également que les concepts avec lesquels on est susceptible d'annoter l'image dépendent de l'échelle de l'image. Les concepts à employer peuvent être vue comme une pyramide en fonction de la résolution. Plus la résolution est grande, plus les concepts se spécifient, plus les objets urbains s'individualisent [57].

2.3.3 Interaction avec l'utilisateur

La construction de systèmes interactifs, ou semi-interactifs pour résoudre des problèmes de fouille d'images se répand de plus en plus. En effet, de tels systèmes permettent de mieux cerner les attentes de l'utilisateur à travers une série d'interactions. Cette interaction est complémentaire, entre la machine qui est efficace pour traiter un nombre important de données, et l'utilisateur qui définit un contexte d'utilisation. Prenons l'exemple des moteurs de rechercher existant sur Internet. à partir d'une première requête, l'utilisateur sélectionne des réponses qui le satisfont, puis reformule généralement une deuxième requête en ajoutant des mots clés qui lui permettent d'avoir des résultats qu'il considère plus approprié. Ainsi, quelques travaux

ont cherché à transposer les techniques d'enrichissement de la requête, entre autres par le relevance feedback (traduit en franéais par "retour de pertinence") [61] [28]. A partir d'une première étape d'annotation automatique de l'image, l'utilisateur va donner des exemples positifs et négatifs d'images qui vont permettre d'améliorer ces annotations.

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, domaine antérieur à celui de la fouille d'image, a fait l'objet de davantage de recherches que ce dernier. C'est pourquoi il est intéressant d'étudier l'extraction de sémantique dans le texte 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 "dose" 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 sémantique. Ainsi, on peut supposer que le *fossé sémantique* existant entre la sémantique et l'image est bien plus large que celui existant entre le texte et la sémantique.

La fouille textuelle s'est développée principalement au début des années 60 à travers l'accès à l'information à partir de requêtes utilisateurs dans des bibliothèques numériques. 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 data-mining sur ces objets formalisés. Les applications sont notamment :

— Le résumé automatique
— L'indexation automatique
— La génération d'index de livre
— La classification automatique de documents textuels
— L'extraction de concepts
— Le rapprochement entre textes

Nous commencerons 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 de documents

3.1.1.1 Représentations vectorielles de document :

Nous discutons ici des représentations d'un document en sac de mots (bag-of-words) qui ne prennent pas en compte l'ordre des mots dans un document.

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,

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

Vecteur fréquentiel : La représentation binaire d'un document reste limitée 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 $occ_D(i)$ la fonction donnant le nombre d'occurrences du mot-clé i dans le document D, nous définissons d de la façon suivante :

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

Une des limites de la représentation fréquentielle est qu'un document long aura un vecteur de normes plus élevées qu'un document plus court, ce qui peut engendrer des problèmes dans certaines méthodes de clustering ou de classification, ou même dans certains modèles de documents comme le modèle bayésien naïf que nous verrons ci-dessous, dans lesquels 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.

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'informations sur un document même s'ils sont

peu souvent présents, alors que certains mots sont très présents mais constitutif qui décrit la loi de répartitions des mots dans un corpus de documents.

Loi de Zipf : La loi de Zipf 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.

Représentation TF-IDF : Plusieurs formules ont été proposées pour prendre en compte la loi de Zipf dans le codage TF-IDF, elles reposent toutes sur l'hypothèse que la composante correspondant à un terme du vecteur représentant un document est calculée par le produit entre un facteur qui concerne le poids du terme dans le document et un autre qui concerne le poids du terme dans le corpus. 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 *domain frequency*). On peut l'écrire formellement de la façon suivante :

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

3.1.1.2 Représentation séquentielle

La représentation "sac-de-mots" des documents textuels étant insuffisante pour certaines applications (transcription automatique de la parole), il est nécessaire de faire appel à des modélisations plus réaliste du langage gardant une information sur sa séquentialité. Ainsi, les modèles "n-grammes" estiment les probabilités de toute séquence de n mots. Ensuite, 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 :

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

L'inconvénient d'une telle modélisation est le nombre énorme de paramètres à estimer. En effet, si le vocabulaire comporte N mots (comportant en général plusieurs dizaines de milliers de mots), il est nécessaire d'estimer N^n paramètres, ce qui même pour un modèle bigramme ou trigramme constitue un ensemble de paramètres très volumineux et difficile à estimer. Ainsi, seuls les modèles bigramme et trigramme sont couramment utilisés en pratique.

3.1.2 Modélisations non probabilistes du texte

3.1.2.1 Analyse sémantique latente (LSI/LSA)

La technique la plus simple qui consiste à répondre à une requête d'utilisateur en retournant les documents qui contiennent le plus d'occurrences des mots contenus dans la requête, 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) [19, 6]. 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 contienne 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_i^j est l'occurrence 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 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_k, S_k et V_k :

$$A_k = U_k S_k V_k^t$$

La matrice U_k, de dimension $M * k$ donne les coordonnées de chacun des M

termes dans le sous-espace linéaire dans lequel on va projeter les documents. 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_k 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. 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. Elle analyse le contexte d'utilisation d'un terme par rapport aux autres. Dans [19], Deerwester affirme ainsi 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.

3.1.2.2 Méthodes à noyaux

L'idée générale des méthodes à noyaux est que certains algorithmes n'ont besoin d'accéder aux données que par le biais des produits scalaires entre exemples. Ainsi, la matrice symétrique des produits scalaires entre chaque paire d'exemples (dite *matrice de Gram*) regroupe l'ensemble des informations à propos du corpus d'apprentissage dont ces méthodes ont besoin. Par conséquent, les méthodes à noyaux vont "déformer" la structure de l'espace en proposant de nouveaux produits scalaires entre exemples (c'est à dire un nouveau noyau), plus pertinents que le produit scalaire classique.

3.1.3 Modélisations probabilistes du texte

Dans ce chapitre, nous faisons le lien entre le domaine de la recherche d'information (RI) et celui de l'apprentissage. En effet, les modèles les plus performants de la RI ont été des modèles d'apprentissage dont les paramètres ont été adaptés au domaine du traitement de texte. Nous détaillerons ici particulièrement les modèles génératifs, plutôt que les modèles discriminants, plus difficiles à utiliser dans notre cas. Un modèles génératif est défini comme étant un modèle associé à un processus stochastique qui modélise comment certaines informations présentes dans un corpus de documents ont été générées par ce processus. Nous détaillerons ici successivement le modèle bayésien naïf, le modèle PLSA, puis LDA.

3.1.3.1 Le modèle "bayésien naïf"

Le modèle bayésien naïf est un modèle génératif classique qui fut 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 (appelé filtre bayésien dans les applications grand public) pour le filtrage parental, le filtrage de spam,... Plusieurs versions du modèle bayésien

naïf existent, elles reposent sur des hypothèses statistiques légèrement différentes (notamment en ce qui concerne la longueur des documents). Nous expliquons ici simplement le principe général du modèle, puis son application à des documents textuels.

Soit une séquence d'objets discrets $x = (x_1, x_2, ..., x_n)$, où n représente la longueur de la séquence et où les x_i sont à valeurs dans un ensemble discret quelconque $V = (V_1, ..V_m)$ appelé vocabulaire. Soit θ l'ensemble des paramètres du modèle, on peut écrire :

$$P(x|\theta) = P(x_1, x_2, ., x_n|\theta)$$

$$P(x|\theta) = \prod_{i=1}^{n} P(x_i|x_{i-1}, ..|\theta)$$

Le modèle bayésien naïf repose sur l'indépendance conditionnelle des éléments de la séquence entre eux, ce qui est une hypothèse très forte. On tire ainsi de l'équation précédente :

$$P(x|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

L'expression du modèle bayésien naïf est donc simple. Ce modèle possède une inférence de complexité linéaire en fonction de la longueur de la séquence.

Si un document est représenté par un vecteur fréquentiel : $x = (x_1, x_2, ..., x_n)$, où x_i représente donc le nombre d'occurrence du terme i dans le document, et si on note p_i la probabilité d'apprition du terme i dans le document, la probabilité de génération de ce document sachant le modèle θ s'écrit :

$$P(x|\theta) = \prod_{i=1}^{n} p_i^{x_i};$$

3.1.3.2 Modèles de mélange

Un théorème classique formulé par de Finetti affirme que la loi de toute collection de variables aléatoires interchangeables peut s'écrire comme un mélange. Ainsi, si l'on utilise une modélisation où l'on considère que les mots ou les documents sont interchangeables, il est nécessaire de considérer des modèles de mélange. Les deux modèles ci-dessous ont comme point commun de faire cette hypothèse d'interchangeabilité sur les mots. Mais la modélisation LDA est plus complète au sens où une modélisation probabiliste au niveau des documents du corpus est également effectuée. Généralement, un modèle de mélange, par rapport à un modèle unigramme, permet simplement une meilleure description des données, moyennant une complication du modèle. Mais dans les modélisation qui suivent, il est important de noter que les poids des différents modèle dans le mélange véhiculent une information sémantique, soit le poids d'un "topic" dans un texte.

Le modèle pLSA Le modèle pLSA (Probabilistic Latent Semantic Analysis) [6] est un autre modèle génératif de document largement utilisé. Cette méthode suppose qu'un document d et un mot w_n sont conditionnellement indépendants, étant donné une variable latente z, appelée "topic". Étant donné un document, elle suppose le processus génératif suivant :

Pour chaque mot :

— Choisir un topic z_n selon la loi $p(z|d)$

— Choisir un mot w_n suivant la loi $p(w_n|z_n)$, loi de probabilité conditionnée par le topic z_n.

Ainsi, la probabilité jointe d'un document et d'un mot w_n est donnée par :

$$P(d, w_n) = p(d) \sum_z (p(w_n|z)p(z|d))$$

La pLSA essaie d'assouplir l'hypothèse faite par la méthode bayésien naïf selon laquelle chaque document est généré par un seul "topic", à savoir la condition d'homogénéité sémantique. Elle capture la possibilité qu'un document contienne plusieurs topics, et $p(z|d)$ représente ainsi les poids de ces mélanges pour le document. d est un indice représentant un document dans l'ensemble d'apprentissage. Étant donné un nouveau document, ces probabilités doivent être recalculées. Ainsi, la taille du modèle croit linéairement avec le nombre de textes présent 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.

le modèle "LDA' : Latent Dirichlet Analysis [8] La LDA 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 topics latents, où chaque topic est caractérisé par une distribution sur l'ensemble des mots.

La 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 mot w_n :

  — Choisir un topic z_n suivant une loi multinomiale de paramètre θ.

  — Choisir un mot w_n suivant la loi $p(w_n|z_n, \beta)$, loi multinomiale conditionnée par le topic z_n.

Dans ce modèle, la variable θ est un vecteur qui prend ses valeurs dans le $(k-1)$ simplex, on suppose que sa dimension est fixée et que la dimension du vecteur z, c'est-à-dire le nombre de topics possibles, est fixée également. La matrice β contient les probabilités des mots sachant le topic : $\beta_{ij} = P(w_j = 1|z_i = 1)$, cette quantité est à estimer. 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 bayésien naïf 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 des méthodes variationnelles dont ils affirment qu'elles permettent d'estimer correctement les divers paramètres.

3.2 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 d'ordre sémantique. Les systèmes sont généralement annotés automatiquement par des mots-clés sémantiques, 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 :
a. L'annotation automatique d'images nouvelles.
b. 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 sémantiques 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 dans l'image.

3.2.1 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 dans l'image.

3.2.1.1 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. (voir figure 2.4). Le but de l'annotation sémantique est, étant donné une image I, d'extraire un ensemble d'étiquettes w, qui décrit I de façon optimale. 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. [74].

3.2.1.2 Étiquetage supervisé

Vu comme un étiquetage supervisé, l'étiquetage est formulé comme étant composé de T problèmes de détection déterminant la présence oul'absence des concepts dans L. Considérons ainsi le i-ème problème de détection et la variable Y_i définie par :

$Y_i = 1$ si I contient w_i 0, sinon

Etant donné un vecteur de q caractéristiques $X = \{x_1, ..., x_q\}$ extrait de I, le but est d'inférer l'état de Y_i en minimisant la probabilité d'erreur, pour tout $i \in \{1, ..., T\}$. En utilisant des résultats bien connus de théorie de la décision [3], ce problème peut être résolu en posant que le concept est présent est :

$$P_{X|Y_i}(X|1)P_{X|Y_i}(1) = P_{X|Y_i}(X|0)P_{Y_i}(0)$$

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 i, l'ensemble D_1 des images annotées par l'étiquette w_i et l'ensemble D_1 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 abondement 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 [69], des peintures et des photographies [17], des êtres humains et des animaux [24], des villes et des paysages [73].

3.2.1.3 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.

Dans [2], l'annotation sémantique porte sur des images satellitaires, une classification bayésienne des caractéristiques de bas niveau (spectral, radiométrie, texture) est tout d'abord opérée dans le but d'obtenir des "régions prototypes". Des caractéristiques globales sont extraites sur chaque région prototype et 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”. Une grammaire visuelle est ensuite calculée pour les différents concepts sémantiques à l'apprentissage à partir d'exemples fournit par l'utilisateur par comptage des différents types de configuration entre les régions prototypes. Le système est alors capable d'apprendre des notions sémantiques qui auraient été plus difficiles à retrouver sans la prise en compte de la spatialité entre les régions prototypes 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 [50] 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 inconsistantes (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.

3.2.2 Application de techniques textuelles à l'image

Un autre grand axe de l'annotation sémantique d'images consiste à extraire un certain nombre de *mots visuels* dans l'image dont on souhaite qu'ils se situent à un niveau intermédiaire entre les pixels et la sémantique. L'introduction de ce vocabulaire discret permet alors la mise en œuvre de techniques d'inférence statistique qui ont prouvé leur efficacité dans la recherche de documents textuels.

3.2.2.1 Primitives de l'image

Dans les années 1960-1970, Julesz supposa les ”textons” comme les éléments atomiques d'une perception visuelle des structures locales [39]. 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” ([44]). 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 images afin de traiter de grandes bases de données d'images ([1]).

3.2.2.2 Groupements géométriques basiques

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 ([78],[40]). 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'œil humain.

3.2.2.3 Modèles textuels

Dans [8], les auteurs adaptent un modèle génératif hiérarchique proposé pour le texte par Hofman [32] [33]. Ce modèle regroupe les documents dans des clusters 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.

[7] est un autre article particulièrement important traitant de l'annotation d'images par modèles textuels. Les auteurs utilisent trois méthodes hiérarchiques de génération de données annotées, adaptées de modèles traditionnellement utilisés pour générer des 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". Étant 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 :

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

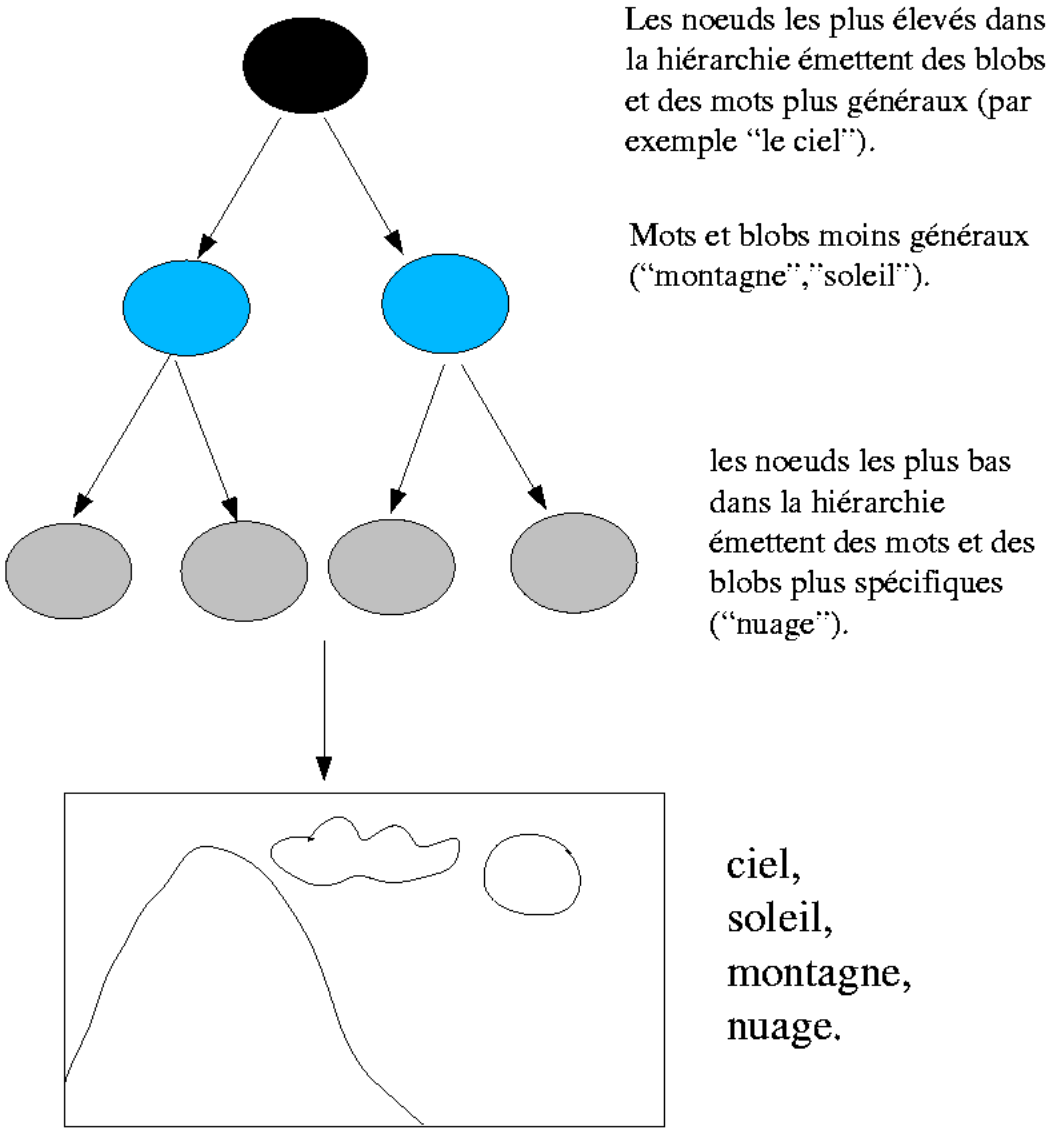


FIGURE 3.1 – Illustration du processus génératif implicite au modèle statistique.

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.(voir figure 3.3) Ainsi, étant donné un document, une variable aléatoire θ est tout d'abord générée avec une distribution de Dirichlet, ses composantes définissent la distribution de probabilité des variables latentes. Pour chacune des N régions et chacun des M mots, une variable aléatoire est tirée avec la distribution $Mult_\lambda(\theta)$ et le descripteur de région r_n ou le mot v_m est généré conditionnellement à cette variable latente.

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

Si ce modèle capture la possibilité d'avoir des régions ou des mots issus de différents modèles, il ne fait pas de lien direct entre la description d'une région et son annotation, ce que fait par contre le troisième modèle : "Correspondance LDA" :

Le modèle Corr-LDA est décrit figure 3.4, il combine la flexibilité de GMM-LDA et l'associativité de GM-mixture. Une variable aléatoire θ ayant une distribution de Dirichlet est générée pour tout le document, et donne les coefficients de la loi multinomiale de la variable aléatoire latente z qui est générée pour chaque région. Les N régions r_n sont tout d'abord générées avec un modèle LDA. Ensuite, pour chacune des M annotations, une des régions est sélectionnée dans l'image et le mot w_m est généré en tant qu'annotation de l'image, conditionnellement à la variable latente qui a engendré la région en question.

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

Ce modèle est un compromis entre l'extrême correspondance du modèle GM-mixture, où dans l'image entière, les descripteurs de région et les mots sont conditionnés au même facteur, et le manque de correspondance du modèle GM-LDA, où les descripteurs de région et les mots d'annotation sont conditionnels à deux lots séparés de variables latentes.

3.2.2.4 Traitement de l'image comme une collection discrète

Dans la thèse de Pecenovic [54], 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 sémantique de plus petite dimension. Les caractéristiques réelles des vecteurs de caractéristiques calculés dans l'image sont gardés pour créer la matrice de

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

FIGURE 3.2 – résultats d'annotation d'images de la base Corel

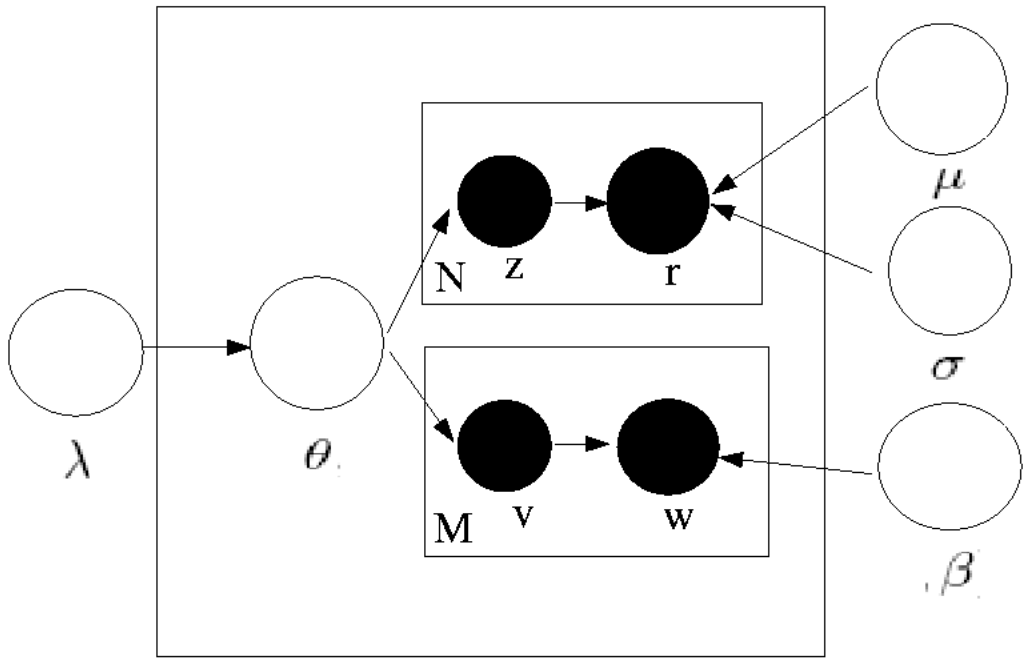


Figure 3.3 – Modèle de mélange d'images GMM-LDA. Contrairement au modèle GMM, chaque mot w et chaque région r peut provenir d'une variable latente différente.

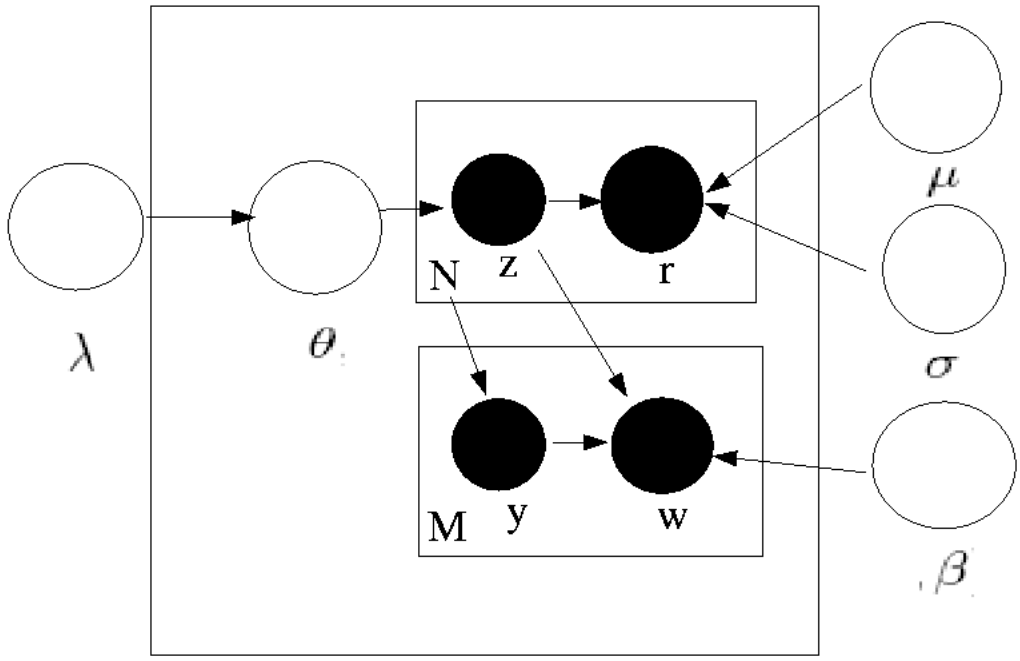


FIGURE 3.4 – Modèle de mélange d'images CORR-LDA. Notons que les variables y_m sont conditionnées par N, le nombre de régions de l'image.

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 [48], les auteurs étendent cette idée, qui consiste à appliquer des méthodes qui ont montré certains succès en texte pour la recherche d'images, pour l'annotation automatique d'images et vont plus loin dans le rapprochement avec le texte en travaillant à partir de caractéristiques quantifiées. L'image est segmentée grossièrement en 3 régions : le centre, la partie haute et la partie basse. Pour chaque région, des caractéristiques calculées à partir de l'histogramme sont quantifiées et le tout est mis dans un sac de mots. On rajoute dans ce sac de mots 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. La LSA est le modèle le plus simple car l'apprentissage nécessite simplement une décomposition en valeurs singulières (SVD : Singular Value Decomposition) de la matrice termes/documents, c'est à dire que les variables latentes sont calculées directement à partir du signal. Les matrices obtenues par SVD permettent de faire le passage d'un sac de mots représentant un document à un vecteur dans l'espace latent, et réciproquement. A partir d'une image non-annotée, le vecteur latent est calculé et comparé aux vecteurs latents des images annotées. Les annotations sont ainsi propagées au reste de la base.

Comme expliqué en 1.1.2, la PLSA modélise chaque terme du document comme étant émis par une variable latente qui a une distribution dépendant de chaque document, et qui représente les "topics" présents dans celui-ci. Contrairement à la LSA, la PLSA propose une modélisation probabiliste de génération du document. La probabilité jointe document/mots, si nous supposons qu'il y a K "topics" qui sont susceptibles d'être présents dans le document, est exprimée par la formule suivante :

$$P(w_j, d_i) = P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

Les probabilités $P(w_j|z_k)$ sont estimées à l'apprentissage, mais les probabilités $P(z_k|d)$ doivent être estimées pour chaque nouveau document non annoté d. Les annotations sont obtenues par un calcul de probabilité en maximisant la probabilité à posteriori des termes du vocabulaire sachant le nouveau document q. Les résultats obtenus montrent que les gains de performance de la PLSA par rapport à la LSA ne sont pas significatifs malgré la différence de complexité entre les deux modèles.

Dans [20], les auteurs définissent clairement une "approche d'adaptation" pour

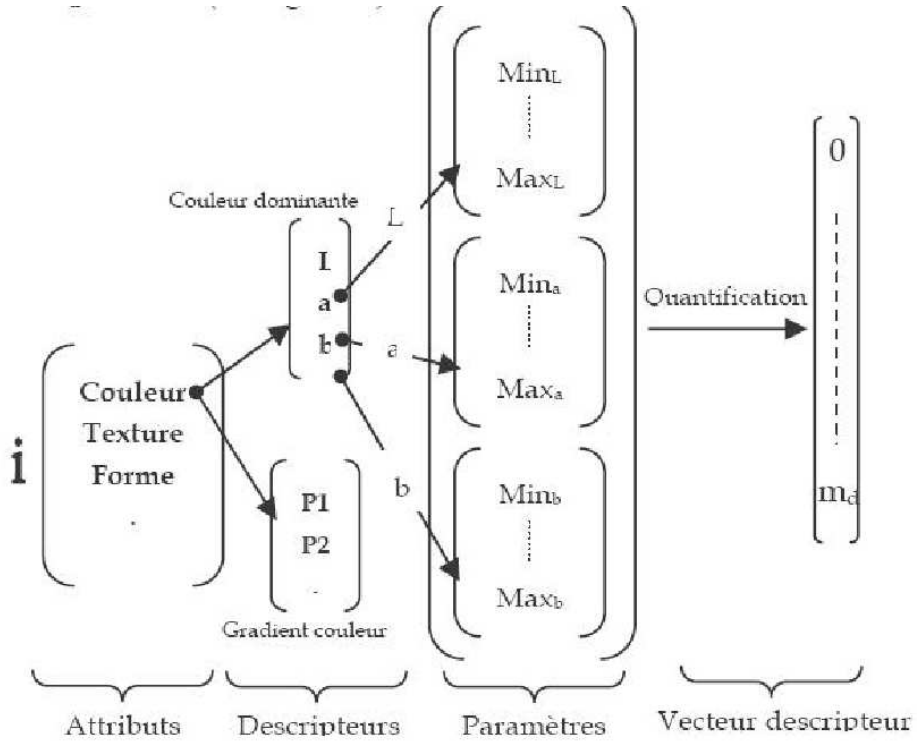


FIGURE 3.5 – Méthode d'adaptation permettant de passer de la fréquence d'apparition des mots-cés à une représentation pertinence des descripteurs

passer de la notion de fréquence d'apparition des mots-cmés à une représentation pertinente des descripteurs (voir figure 3.5) , qui ont habituellement des valeurs numériques. A chaque image est associé un ensemble d'attributs de bas-niveau (couleur, forme, texture), chaque attribut étant décrit par un ensemble de descripteurs (couleurs dominantes ,gradient couleur...). Ceux-ci sont composés chacun d'un certain nombre de paramètres (L, a, b...) 3.5. La méthode d'adaptation choisie est la suivante :

- Les p_d paramètres caractéristiques sont identifiés (exemple : les composantes L, a, b de l'espace couleur)
- D'identifier la plage de variation de chaque paramètre (exemple : pour la composante L l'intervalle $[Min_L...Max_L]$)
- Sélectionner un ensemble d'apprentissage, tiré aléatoirement des bases d'images utilisées, appelé base de test.
- D'efectuer, suite à une analyse statistique du comportement de tous les paramètres sur la base de test, une quantification de l'espace du descripteur sur m_d valeurs.

Dans [65], 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.

Dans [15], les auteurs exposent une approche pour indexer des scènes cinématographique 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. Étant 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.

Etiquettage non-supervisé Des efforts plus récents on été faits pour traiter le problème dans toute sa généralité. L'idée est d'introduire un lot de variables latentes qui codent les états cachés dans l'image, où chaque état définit une distribution jointe sur l'espace des 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 donné 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. Différentes méthodes existent qui diffèrent par la définition des états des variables latentes. Certaines associent un état à chaque image de la base de donnée, d'autres les associent à des clusters d'images, et d'autres modélisent des groupements de haut-niveau, comme des "topics". 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 $\mathbf{w}$ est l'étiquetage 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 du au couplage entre le vecteur χ qui a une valeur continue, et $\mathbf{w}$ qui est à valeur discrète, les caractéristiques visuelles et les étiquettes sont souvent supposées être des variables indépendantes conditionnellement à la variable latente :

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

Le modèle étant un modèle de mélange, l'algorithme expectation-maximization (EM) est la plupart du temps employé pour mener à bien l'apprentissage. Certaines

méthodes de régularisation bayésiennes permettent de réduire l'impact d'annotations faibles.

Les deux formulations ont chacune des avantages et des inconvénients. En général, l'annotation non-supervisée conduit à des procédures d'apprentissage plus abordables. D'un autre côté, elle ne traite pas exactement les sémantiques comme des classes d'images et apporte peu de garantie que les annotations sémantiques sont optimales dans une optique de recherche d'images. En effet, au lieu de fournir les annotations qui ont la probabilité la plus faible d'erreur de recherche d'images, elle produit simplement celles qui ont la plus grande probabilité jointe étant donné le modèle de mélange qui a été supposé.

Dans [13] , une procédure de "soft annotation" est en plus proposée. On annote tout d'abord un ensemble d'apprentissage où chaque image est annotée avec un seul mot (ex : forêt, ciel etc.). Un ensemble de classifieurs binaires est ensuite appris à partir de cet ensemble. Ensuite, sur de nouvelles images, des annotations multiples sont produites avec des mesures de confiance en utilisant un classifieur BMP (Bayes Machine Point) et un classifieur SVM (Support Vector Machine). Ces mesures de confiance seront prises en compte lors des requêtes de l'utilisateur, et peuvent également être améliorées lors d'une interaction avec lui du type "relevance feedback" (voir la sous-section suivante).

Annotation de portions d'image 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.

Dans [50], l'approche utilisée consiste à découper de grandes photos aériennes en imagettes de taille 64×64 et à extraire pour chacune les caractéristiques de Gabor. Chaque imagette du lot d'apprentissage est associée à une classe (route, végétation, parking etc.). Un modèle de mélange de gaussiennes (GMM) est calculé pour décrire la distribution statistique des caractéristiques pour chaque classe, et permet d'annoter les tuiles de l'image automatiquement par le maximum de vraisemblance. Cette méthode utilise des régions de taille prédéfinie afin de faire l'apprentissage et l'annotation automatique, ce qui est réducteur car les régions à décrire dans l'image de test ne sont pas connues à priori et peuvent avoir des formes variées.

Afin de pouvoir attacher des régions de taille non prédéfinie à des annotations, Mori et Takahashi présentent dans [49] 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. A 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 régions 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". Étant 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 [5], 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 [37], 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 [51], pour pallier le problème d'une segmentation préalable, les auteurs introduisent en plus d'une variable latent 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 :

$$P_{X|L}(x|l) = \sum_{i=1}^{5} \pi_i G(x, \mu_i^l \sigma_i^l)$$

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 déterminé 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.

3.2.2.5 Interactive learning

Beaucoup de systèmes utilisent un retour de pertinence pour améliorer les résultats d'annotation. Au début, les poids des descripteurs pour chaque image sont fixes et objectifs, puisque calculés de façon indépendante, tandis que les requêtes de l'utilisateur sont subjectives par nature. L'objectif est de faire des interactions entre le système et l'utilisateur afin de faire transparaitre cette subjectivité dans les poids des descripteurs pour la composition de la réponse. Beaucoup de travaux exploitent le relevance feedback pour améliorer les performances de leur système, mais il ne constitue pas nécessairement le coeur de la méthode. C'est pourquoi nous ne détaillerons pas trop ici ce point. Nous citons cependant l'approche décrite dans [52] où les auteurs proposent une méthode d'*interactive learning* pour relier les concepts subjectifs qui intéressent l'utilisateur aux valeurs symboliques caculées 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. Le niveau le plus bas correspond aux pixels de l'image (niveau 0). Des modèles stochastiques sont appliqués pour toutes sortes de caractéristiques (spectrales, texturales, etc.) et sont obtenues 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. A 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 un clustering non supervisé, en utilisant une classification bayésienne, ou un algorithme de k-means. Les niveaux 1 à 3 sont obtenus par une caractérisation complètement non-supervisée

de l'image. A 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" :

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

L'exemple donné ici utilise une combinaison de caractéristiques spectrale 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.

3.2.3 Analyse syntaxique de l'image

Inspirée de la sémantique des grammaires formelles, 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([56]). 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" and "top-down".

— Le "gap 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 ([21]), les pyramides d'images ([67]), les ondelettes ([18]), et les méthodes d'apprentissage ([63], [25]).

Après un nombre important d'avancées dans ces domaines, des travaux d'analyse syntactique d'images commencent à apparaître de nouveau dans la littérature. Nous citons ici les travaux de Ahuja ([71]), S. Geman ([27], [38]), Pollak ([75]) et Zhu ([31], [30], [29], [79]). 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 ([14]).

3.2.3.1 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 sont possibles :

$$A- > \beta_1|\beta_2|...|\beta_{n(A)}, \gamma_i : A- > \beta_i$$

Chaque règle γ_i est associée à une probabilité $P(\gamma_i) = P(A- > \beta_i)$ telle que : $\sum_{i=1}^{n(A)} P(\gamma_i) = 1$.

La probabilité de l'arbre syntaxique $pt(\omega)$ ("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 $\omega \in L(G)$ s'écrit :

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

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

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

3.2.3.2 Grammaires stochastiques sensibles au contexte.

Une grammaire (V, Σ, P, S) est dite sensible au contexte si les règles de production P sont de la forme :

$$\alpha A\beta -> \alpha\gamma\beta$$

avec $A \in N$, $\gamma \in V^+$, $\alpha, \beta \in V^*$, ou

$$S -> \epsilon$$

si

$$S -> \epsilon \in P$$

et S n'apparaît jamais dans le membre de droite d'une règle de production.

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

Passer de grammaires de langage en 1 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" ([26]), par Grenander ([41]), et plus récemment dans des graphes de grammaire pour l'interprétation de diagrammes ([58])
- 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 plus grandes que dans le langage. Les images peuvent comprendre des occlusions et des zones texturées dont les règles de production seront fortement stochastiques.

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

3.3.1 Utilisation d'ontologies pour la classification d'images

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.

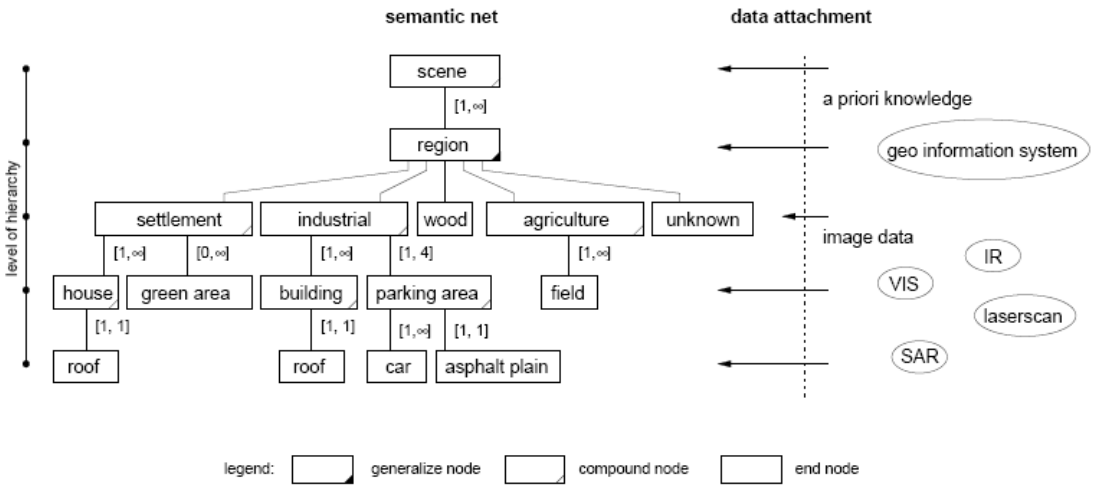


FIGURE 3.6 – Réseau sémantique utilisé par le système GeoAIDA

Ainsi, dans [42], la base de donnée topographique ATKIS (Amtliches Topographisch Kartographisches Informations System), créée par l'administration allemande, est modélisée par le réseau ERNEST (Elanger Semantisches Netzwerksystem [22]). 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", "traffic", "végétation", "eau", relief", et "zone"). Ces 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 [36], 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 3.9). Ce réseau sémantique contient deux types de nœud : le nœud de généralisation (relation sémantique d'hyponymie) et le nœud 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.

3.3.2 Construction automatique de réseaux sémantiques

En fouille d'images, les réseaux sémantiques sont essentiellement utilisés pour clarifier les termes employés et pour apporter une information à priori sur les relations sémantiques entre différents concepts. Cependant, une construction automatique

d'un réseau sémantique permettrait d'une part une souplesse accrue, et permettrait d'autre part d'aller plus loin dans l'extraction de sémantiques.

Ce premier point part d'un constat de rigidité de réseaux sémantiques comme connaissance à priori donnée par un expert. Certains utilisateurs peuvent vouloir introduire de nouveaux termes dans un réseau sémantique, qui pourraient ainsi y être intégrés automatiquement et itérativement.

Le deuxième point concerne l'ambition des travaux d'extraction de sémantiques dans les bases d'images. Une ambition plus grande que celle de mettre simplement en relation des images de test avec des concepts sémantiques introduits par l'utilisateur consiste à élaborer un système permettant de déterminer des liens sémantiques entre des concepts à partir d'une base d'images. En effet, comme cela a été expliqué en 2.1, une branche de la sémantique consiste a étudier les relations entre concepts, comme la synonymie,l'homonymie etc.

Ainsi, dans [66], le lien entre les caractéristiques de bas-niveau et les concepts de haut-niveau est déterminé par relevance feedback (RF). Le RF est utilisé pour capturer la perception subjective de l'utilisateur. Pour cela, un réseau sémantique dit "probabiliste" est construit à l'apprentissage et mis à jour itérativement à chaque boucle de pertinence. Ce réseau sémantique prend la forme d'une matrice d'affinité entre image A. Si $I = \{i_1, ..., i_N\}$ est le lot d'images d'apprentissage, $A = \{a_{m,n}\}(1 \geq m, n \geq N)$ note les proximités sémantiques entre les images. Chaque paramètre $a_{m,n}$ est mis à jour lors de l'interaction avec l'utilisateur en fonction du nombre de fois que les images m et n sont sélectionnées simultanément.

3.4 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 correspondre 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 ex-

plicité 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 inférés efficacement par une modélisation directe des caractéristiques de bas-niveau, on remarque que pour d'autres concepts, le *gap 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.

Le modèle basé sur un réseau sémantique hiérarchique proposé dans ce chapitre, outre le fait qu'il permette de prendre en compte les relations 'kind-of' et 'part-of', permet d'améliorer considérablement la représentation et la prise en compte des relations sémantiques. Il permet de tirer partie de la richesse descriptive d'un réseau sémantique tout en conservant la souplesse des modèles statistiques à variables latentes et permet un apprentissage statistique par présentation d'images exemples fournies pour chaque concept. L'idée fondamentale est de mettre en bijection un ensemble de réseaux sémantiques et un ensemble de modèles statistiques permettant d'estimer la vraisemblance de la base d'apprentissage. La forme de la fonction de distribution de chaque concept dépend de la place du concept dans la hiérarchie du réseau sémantique et peut s'exprimer en fonction des distributions associées aux autres concepts. Ainsi, étant donné une base d'apprentissage constituée d'images annotées, une procédure de sélection de modèle permet de définir un réseau sémantique.

Le critère de sélection du modèle qui a été choisi est celui de minimisation de la longueur de description (MDL : Minimum Description Length). Ce critère stipule que le modèle optimal pour décrire une base de données est celui qui fournit le meilleur codage de cette base de données. Le réseau sémantique qui est ainsi obtenu est celui qui correspond au modèle statistique permettant le plus court codage, et donc la meilleure compréhension de la base de d'apprentissage.

Chapitre 4

Modélisation stochastique associée à un réseau sémantique

Le postulat fondamental de la sémantique est que les concepts mènent une existence qui leur et propre dans un univers qui est structuré. S'intéresser au problème de la sémantique dans les bases d'images nécessite donc de prendre en compte les liens sémantiques qui les lient entre eux. L'idée principale de la modélisation introduite dans ce chapitre est de mettre en relation une structure sémantique, qui sera représentée ici par un réseau sémantique, avec un modèle stochastique.

La sémantique lexicale considère 4 types de relations sémantiques : la synonymie, l'antonymie, l'hyponymie et la méronymie. Parmi ces 4 relations, l'antonymie est la seule qui puisse difficilement être prise en compte dans notre approche. En effet, rappelons (voir 2.1.2) que l'antonymie est une relation binaire qui se décompose en 4 types : polaire, inverse, réciproque, et contradictoire. étant donné que ce travail prend en compte des concepts annotant des régions d'une image, aucun de ces trois premiers types d'antonymie n'a de sens ici. En ce qui concerne l'antonymie contradictoire, si l'on considère un concept tel que "zone résidentielle", l'antonyme de ce concept serait donc "tout ce qui n'est pas une zone résidentielle". Or ici, on suppose que l'on fournit au système un certain nombre de concepts associés à des images exemples qui permettent au système de modéliser efficacement la base de données d'images. Il y a ainsi peu d'intérêt à fournir au système un concept qui se définit comme étant "tout sauf" un concept donné, c'est pourquoi nous ne prenons pas en compte ici la relation d'antonymie. Les relations de synonymie, de méronymie et d'hyponymie feront par contre chacune l'objet d'une modélisation spécifique.

Si le "niveau de sémantique" d'un concept donné est une notion pour nous intuitive, elle n'en reste pas moins difficile à définir avec précision. En effet, le niveau de sémantique peut être défini de façon différente en fonction de l'application envisagée et de l'objectif poursuivi. De notre point de vue, l'information, qui est une propriété statistique du signal, n'a en effet rien de commun avec la signification. Nous préférons représenter le sens comme étant sous-jacent au signal, et nous modélisons un concept sémantique annotant une image par un modèle à partir duquel est calculée la vraisemblance du signal de l'image. Les signifiés sont donc représentés comme des entités appartenant à un système structuré situé en amont du signal, mais ils ne sont pas

confondus avec l'information. Pour représenter la structure de ce système, nous utilisons un réseau sémantique dont la hiérarchie est organisée par des liens sémantiques. Le niveau sémantique par rapport à ce lien est donc exprimé naturellement par le niveau du concept dans la hiérarchie. Dans un premier cas, nous utilisons des liens paradigmatiques de sens : les liens d'hypéronymie et d'hyponymie. Dans un deuxième cas, nous introduisons le lien "part-of" que l'on peut considérer comme étant syntagmatique. La définition précise de ces deux liens sera explicitée ultérieurement.

4.1 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 liens sémantiques qui seront utilisés par la suite.

4.1.1 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 et 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 appellera 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 4.1.1. Ainsi, un image annotée par le concept "Végétation" est aussi annotée soit par le concept "Toundra", soit par le concept "Forêt". Les nœuds de ce type des réseaux sont d'autant plus généraux qu'ils appartiennent à une couche élevée dans la hiérarchie.

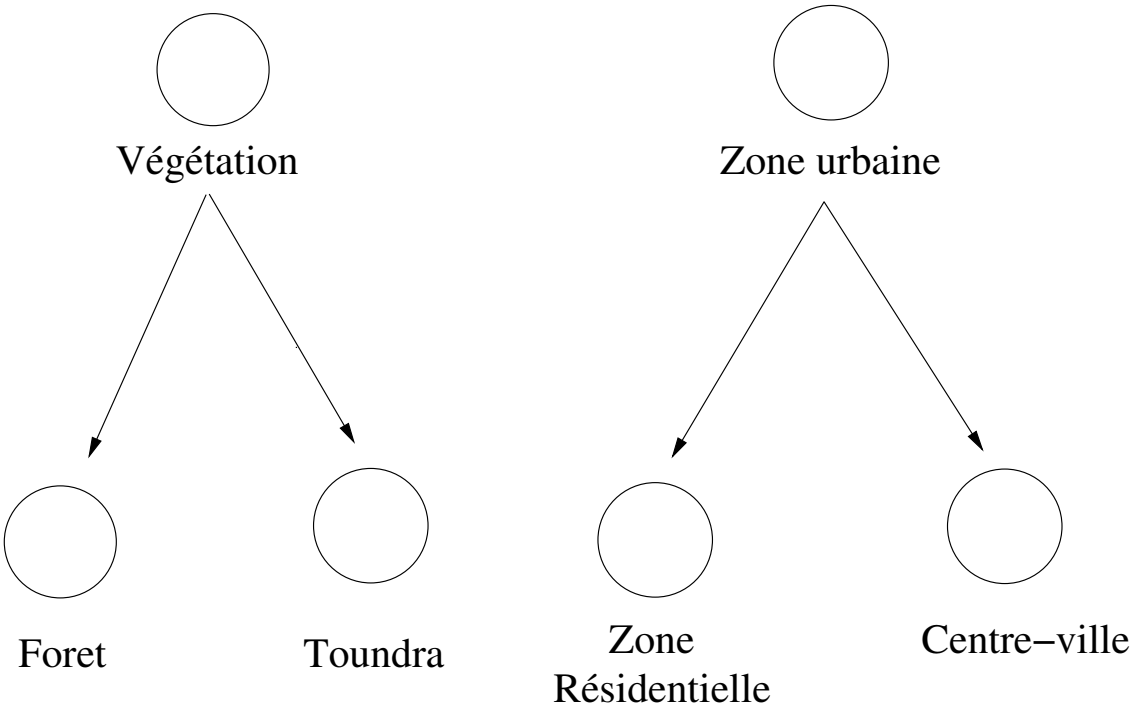


Figure 4.1 – Exemple de réseau sémantique hiérarchisé par la relation *kind-of*. Une image est annotée par le concept “Toundra” ou “Forêt” sera aussi annotée par le concept ”Végétation”. Une image annotée par le concept “Zone urbaine” est soit annotée par le concept “Zone résidentielle”, soit par le concept “Centre-ville”.

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 en effet par définition à combiner des mots, qui sont les unités porteuses de sens, pour aboutir à une nouvelle signification généralement plus complexe. Un syntagme étant par définition un groupe de mots, intermédiaire entre le mot et la phrase, les mots combinés par cette relation sont alors accolés les uns aux autres. Par analogie, plusieurs régions annotées par des concepts peuvent se combiner 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 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.

4.1.2 Réseau sémantique et ontologie

Dans les chapitres 3 et 4, la sémantique est introduite à travers les concepts fournis par l'utilisateur définissant chaque lot d'images exemples. Les images sont ensuite soumises à une segmentation en régions dont chacune est annotée par un concept. Ainsi, de même que dans la plupart des systèmes d'annotations actuels ([20],[37],[51],[74]), la sémantique est extraite uniquement à travers l'annotation d'images par les concepts introduits par l'utilisateur. Or, le principe fondateur de la sémantique en tant que domaine d'étude indépendant est que les concepts mènent une existence qui leur est propre. Et un des buts de la sémantique structurale est de trouver des regroupements pertinents en classes sémantiques ou de relier les mots associés aux concepts par des notions de polysémie, synonymie, etc. On souhaite donc, pour aller plus loin dans notre étude de l'extraction de sémantique, définir une approche pouvant déterminer automatiquement un certain nombre de liens sémantiques qui peuvent exister entre les concepts introduits par l'utilisateur. Ainsi, une ontologie consistant en un réseau sémantique décrivant un domaine complet, nous souhaitons établir une méthode permettant la construction d'une "onto-

logie partielle" dont les concepts sont ceux introduits par l'utilisateur et dont les liens sont déterminés automatiquement par une analyse du signal contenu dans les images exemples fournies par l'utilisateur pour chaque concept.

Une ontologie est supposée par définition décrire complètement un domaine. Or, dans notre cas, les concepts introduits par l'utilisateur ne décrivent pas nécessairement la totalité des domaines que nous pouvons considérer : agronomie, cartographie, etc. L'approche que nous détaillons dans cette section n'a donc pas la prétention de construire une véritable ontologie, mais un réseau sémantique dont les concepts ne sont qu'un sous-ensemble du domaine qui nous intéresse.

4.1.3 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.

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 relation sémantique entre les concepts : "part-of" et "kind-of", et nous supposons une structure hiérarchisée en couches où des liens sémantiques ne peuvent exister qu'entre 2 couches successives. Nous proposons dans ce chapitre deux réseaux sémantiques correspondant à chacun de ces deux liens sémantiques, qui seront chacun associés à deux modélisation probabilistes différentes du signal de l'image.

4.1.4 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 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.

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

4.1.5.1 Caractéristiques de Haralick

Les caractéristiques de Haralick se basent sur les matrices de co-occurrence. 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. On préfère donc les représenter 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, foret, nuages, neige. Les résultats de classification avec les coefficients de Haralick peuvent être considérés comme très satisfaisants, et surpassent notamment les résultats obtenus avec les caractéristiques de Gabor [11]. Les caractéristiques de Haralick ont également montré qu'elles étaient très robustes aux changements de contraste et de luminosité.

4.1.5.2 Clustering des caractéristiques de Haralick

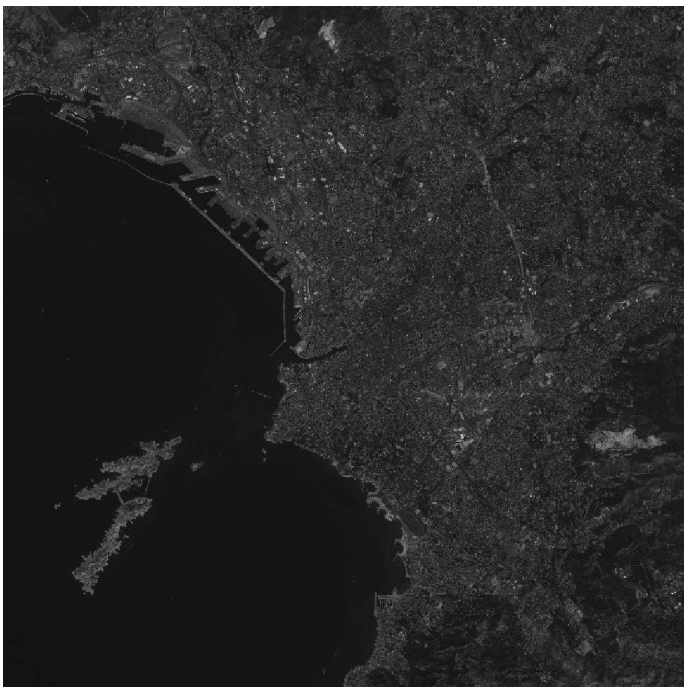
Afin de déterminer la taille optimale du codebook à utiliser pour effectuer le clustering des caractéristiques de Gabor sur la base d'images SPOT5 que nous considérons, nous utilisons l'approche proposée dans [34]. Un modèle de mélange de gaussiennes est utilisé pour modéliser cet ensemble de caractéristiques 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 4.2).

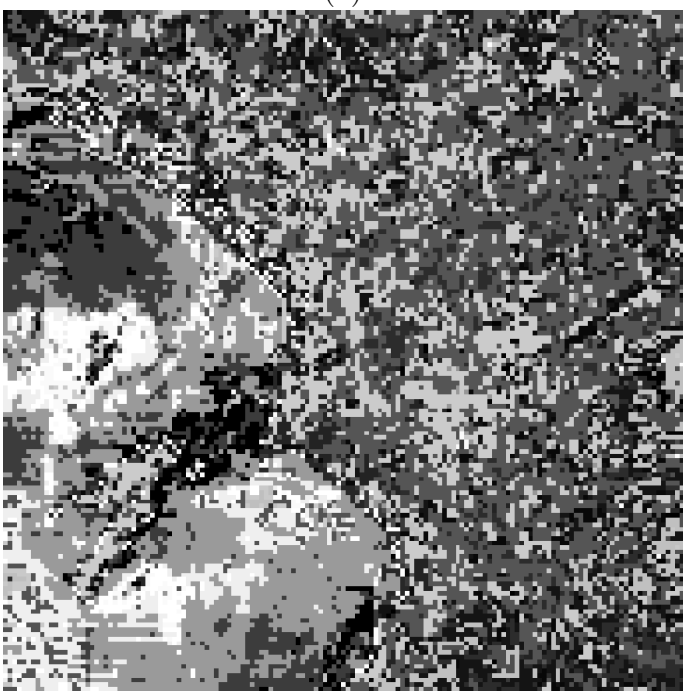
4.1.6 Description de bas-niveau d'images Quickbird

4.1.7 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



(a)



(b)

Figure 4.2 – 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 pixels. L'image (b) de taille 148 × 148 correspond à la quantification de ces caractéristiques : la valeur de chaque pixel est l'indice du cluster dans lequel a été classé le vecteur correspondant à cette fenêtre. ©CNES

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 R_Ω d'une manière qui sera détaillée dans les sections suivantes. Nous considérons ici que chaque concept est associé à un modèle statistique 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 le seconde couche.

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 pixels de cette image. On suppose par ailleurs que chaque image X_{ij} de la base d'apprentissage correspond elle même à une région. 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 discrétisées et étant extraites selon une grille régulière, elles forment les pixels d'une nouvelle image sur lesquelles s'appliquent les algorithmes d'inférence de sémantique mis en œuvre dans ce travail. On appellera donc *pixels*, par commodité, ces caractéristiques de bas-niveau. On ne s'intéresse ici qu'à l'histogramme de ces pixels dans une région donnée, qu'on notera pour une région quelconque $x = (x_1, ..., x_{n_0})$.

4.2 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 un nœud de la couche 2 doit être relié à au moins un nœud.

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 pixels. Les seules régions qui sont considérées ici sont celles définies par l'ensemble des pixels 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

pixels qui y sont présents.On note par ailleurs $C(M_i)$ le numéro de couche auquel appartient un modèle, $C1(M)$ l'ensemble des modèles de la couche 1 du modèle M, $C2(M)$ l'ensemble des modèles de la couche 2 du modèle M.

4.2.1 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.

4.2.1.1 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 pixels dans l'image est codé par une loi de Poisson, et où chaque pixel est tiré indépendemment avec probabilité θ_{cj} si sa valeur est j. Ainsi, la vraisemblance de l'image x sachant M_c s'écrit :

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

4.2.1.2 nœuds de la deuxième couche

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

$$P_a(x) = \sum_{i=1}^{k} \pi_i P_{c_i}(x) \tag{4.3}$$

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 pixels, 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.

En effet, cela revient à dire que a est associé à une variable latente L_a, et que cette variable latente peut prendre ses valeurs dans $\{1, .., k\}$ avec probabilité $P(L_a = i) = \pi_i$. Ainsi, la spécification correspond au choix d'une valeur de la variable aléatoire.

4.2.1.3 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 :

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

4.2.1.4 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 $X = \{X1, ..., X_n\}$ 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'** :

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

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 [4] (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 $i \in \{1, ..., n\}$, $\mathbf{M}_i$ et $\mathbf{M'}_i$ les deux ensembles de modèles décrivant le lot d'apprentissage X_i pour chacun des deux cas de figure. Les modèles M_i et M'_i peuvent se situer en première ou en deuxième couche du réseau, ce que l'on écrit de la manière suivante :

$$\mathbf{M}_i = C1(\mathbf{M}_i) \cup C2(\mathbf{M}_i)$$

$$\mathbf{M'}_i = C1(\mathbf{M'}_i) \cup C2(\mathbf{M'}_i)$$

Les modèles de la couche 1 correspondants étant définis par une modélisation directe sur les pixels 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 :

$$C1(\mathbf{M}_i) = C1(\mathbf{M'}_i)$$

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

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

Ainsi, on a la relation d'inclusion :

$$\forall i, \mathbf{M}_i \supset \mathbf{M'}_i$$

Donc

$$\mathbf{M'} \supset \mathbf{M}$$

On en déduit la propriété à démontrer :

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

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" supplémentaire 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.

4.2.2 Codage des différents modèles

Le principe de minimisation de la complexité stochastique a été introduit par Rissanen en 1978 [59]. Basé sur des concepts issus à la fois de l'estimation statistique et de la théorie de l'information, il apparaît, à un niveau intuitif, relativement naturel. Il a été appliqué dans différents types de problèmes en traitement d'images et a offert de bons résultats.

4.2.2.1 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 [59] 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 posée par Shannon en 1943 [64]. 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 indépendemment 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.

C'est pour cela que 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.

4.2.2.2 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 [60] 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 modèle $M = \{M_k\}_{k=1}^{K}$ 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 ([59]) :

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

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 sachant le 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 :

$$C(X, M) = \sum_{c=1}^{n} [C(X_c|M) + C(M_c)] \tag{4.7}$$

Couche de niveau 1 Les modèles de la couche 1 sont supposés générer directement les pixels 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 exemple associée àu concept c, on utilise la formule proposée par Shannon [64] 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 :

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

Cette dernière quantité est la longueur de code nécessaire pour coder un par un les pixels 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 pixels θ_c, et le paramètre de taille Λ_c. Pour cela nous utilisons la formule introduite par Rissanen ([59]) 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 pixels 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 :

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

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 $CS(M_i)$ l'expression :

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

4.2.3 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 A chaque étape, et pour chaque concept C_j de la couche 1, on calcule la complexité stochastique associée à la configuration dans laquelle le concept C_j 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 :

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

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

$card(X_{cj})$ est le nombre de pixels 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 pixel de valeur j dans la base X_c.

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

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

Les paramètres du modèle étant ainsi estimés, la complexité stochastique globale est alors calculée avec la formule 4.7. 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 :

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

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

4.2.4 Analyse de l'algorithme d'optimisation

A chaque itération, la complexité globale du modèle est calculée pour chaque itération. Ayant au plus N itération et ayant, à l'étape k, $N-k$ configurations à étudier, l'algorithme est donc de complexité N^2.

Cet algorithme est un algorithme dit "glouton" et, à ce titre, il peut fournir un minimum local de la complexité stochastique qui n'est pas un minimum global.

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, sélectionner 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 sélectionner 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 modifiée, 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 $CS(X_c|M)$ entraîné par le passage du concept c sur la couche 2 sera donc nécessairement moins important en e_2 quand 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 susceptible impliquant le modèle c est plus important.

Ainsi, si Δ_c^e est la variation de CS entraînée 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 sélectionner 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 pixels dans la base X (voir équations 4.9 et 4.25), tandis que la complexité associée à l'attache aux données $CS(X|M)$ augmente linéairement (voir équation 4.18). 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.

Ainsi, l'heuristique qui guide la recherche de la structure optimale du modèle semble correcte.

4.3 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 1.2). 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 pixels 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 pixels à l'intérieur de cette région.

Par soucis de simplification, nous supposons que le nombre maximal de couches du modèle global est de deux. Cette modélisation qui est détaillée ici peut aisément se généraliser à un nombre de couches quelconque.

4.3.1 nœuds de la première couche

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

$$P(x|M_c) = Poiss_{\lambda_c}(\sum_{j=1}^{n_0} x_j) \prod_{j=1}^{n_0} \theta_c^{x_j} \tag{4.13}$$

4.3.2 nœuds de la deuxième couche

Si 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 pixels 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 :

$$\begin{aligned} P(X_{ai}, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(X_{ai}|M_a, \{R_1, R_2, ..., R_{mai}\}, \{c(R_1), c(R_2), .., c(R_m)\}) \\ P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) \end{aligned} \quad (4.14)$$

Le premier terme de ce produit s'écrit :

$$\begin{aligned} P(X_{ai}|M_a, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}) = \\ Poiss_{\Lambda_a}(m_{ai}) \prod_{j=1}^{m_{ai}} P(x(R_j)|c(R_j)) \end{aligned} \quad (4.15)$$

où le terme $P(x(R_j)|c(R_j))$ s'écrit avec la formule 4.13, $c(R_j)$ 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 4.14 de la manière suivante :

$$\begin{aligned} P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(\{R_1, R_2, ..., R_m\}|M_a) P(\{c(R_1), c(R_2), .., c(R_m)\}|M_a) \end{aligned} \quad (4.16)$$

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.

4.3.3 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

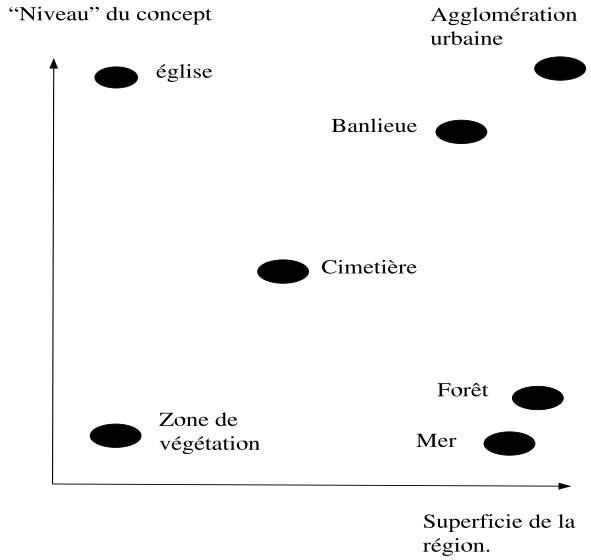


Figure 4.3 – Représentation de différents concepts que l'on attache à des superficies et des "niveaux de sémantique" différents

M. L'expression 4.5 démontrée dans la section précédente s'applique toujours et la vraisemblance de la base de données s'écrit donc :

$$P(X_1, X_2, ..., X_n|M) = \prod_{c \backslash M_c \in C1(M)} P(X_c|M_c) \prod_{c \backslash M_c \in C2(M)} P(X_c|M) \qquad (4.17)$$

4.3.4 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.

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 2.1.4 est rigoureusement valable et peut être réutilisée telle quelle car l'argument d'inclusion explicité dans l'équation 4.6 s'applique encore.

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

Comme en section 2.2, la longueur de codage nécessaire pour coder la base de données est également séparée en deux termes : $C(X, M) = C(X|M) + C(M)$. 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 :
$C(X, M) = \sum_{c=1}^{n}[C(X_c|M) + C(M_c)]$

4.4.1 Codage de la couche de niveau 1

Les modèles de la couche 1 sont, comme dans la section 2.4, supposés générer directement les pixels 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 ([64]), 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 4.13. 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 pixels dans l'image X_{cj}, et en introduisant la formule 4.2 dans l'equation précédente :

$$CS(X_c|M_c) = \lambda_c - card(X_{cj}) \log \lambda_c + \sum_{j=1}^{card(X_{cj})s} \log(j) - \sum_{j=1}^{n_0} x_{cj} \log(\theta_c) \qquad (4.18)$$

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 pixels θ_c, et le paramètre de taille Λ_c. Pour cela nous utilisons la formule introduite par Rissanen ([59]) qui attribue au codage d'un vecteur de paramètres de taille T estimé avec N_{ech} la longueur de codage :

$$\frac{T}{2} log N_{ech} \qquad (4.19)$$

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 pixels 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.

$$CS(M_c) = \frac{n_0}{2}\log(\sum_{j=1}^{N_c}\sum_{k=1}^{n_0} x_{cjk}) + \frac{1}{2}\log(N_1) \quad (4.20)$$

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

$$CS(X_c|M_c) = \lambda_c - card(X_{cj})log\lambda_c + \sum_{j=1}^{card(X_{cj})s} logj - \sum_{j=1}^{n_0} x_j \log(\theta_c) + \frac{n_0}{2}\log(\sum_{j}^{N_c}\sum_{k}^{n_0} x_{cjk}) + \frac{1}{2}\log N_1 + 1 \quad (4.21)$$

4.4.2 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 :

$$P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i) = P(X_{ci}|\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}, M, P_i)P(\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M) \quad (4.22)$$

Le premier terme de ce produit est exprimé par la formule 4.15, et le deuxième par la formule 4.16. On écrit donc :

$$P(X_{ci}|M, P_i) = \sum_{\{c(R_1),c(R_2),...,c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (4.23)$$

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 :

$$P(X_{ci}|M,P_i) =$$

$$max_{\{c(R_1),c(R_2),...,c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (4.24)$$

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 :

$$CS(M_i) = \frac{n_1}{2} \log(N_c) \quad (4.25)$$

Il s'agit pour les concepts appartenant à la couche 2, comme pour les concepts de la première couche, de coder les valeurs des pixels 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 :
$C(X,M) = min_{P,M}(C(X|M,P) + C(M))$

4.4.3 Algorithme d'optimisation utilisé

Le même algorithme que celui détaillé en 2.3 est utilisé, seules les formules d'estimation des paramètres sont différentes pour les modèles de la couche 2.

Étant donné un modèle c de la deuxième couche, et i un index d'une image, on note $P(X_{cj} = \{R_{1j}, ..., R_{m_jj}\}$ 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 :

$$\forall j \in \{1, ..., n_1\}, \pi_{cj} = \frac{occ_{P(X_c)}(c(R_i) = c_j)}{card(P(X_c))} \quad (4.26)$$

$$\Lambda_c = \frac{1}{N_c} \sum_{j=1}^{N_c} card(P(X_{cj})) \quad (4.27)$$

$occ_{P(X_c)}(c(R_i) = c_j)$ 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 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.

4.5.1 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. Soit 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 pixels 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 proportionnel 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 base de données d'images exemples fournies aux concepts est trop réduite, lors de la procédure de sélection de modèles, deux concepts peuvent être reliés par un lien de synonymie, 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 pixels pouvant prendre n valeurs différentes. Le modèle M_1 génère avec probabilité 1 le pixel de valeur 0, et le modèle M_2 génère avec probabilité 1 le pixel 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 pixels, le modèle M' qui est estimé sur la base $X_1 \cup X_2$ génère avec probabilité 0.5 le pixel 0 et avec probabilité 0.5 le pixel 1. Ecrivons 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 entraînera 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'entraînera pas la création d'un lien de synonymie (voir 4.4). Ainsi, la fonction $f_n(N)$ étant continue, elle passe nécessairement par 0. Les zéros de la fonction f_n en fonction du paramètre n sont illustrés figure 4.5

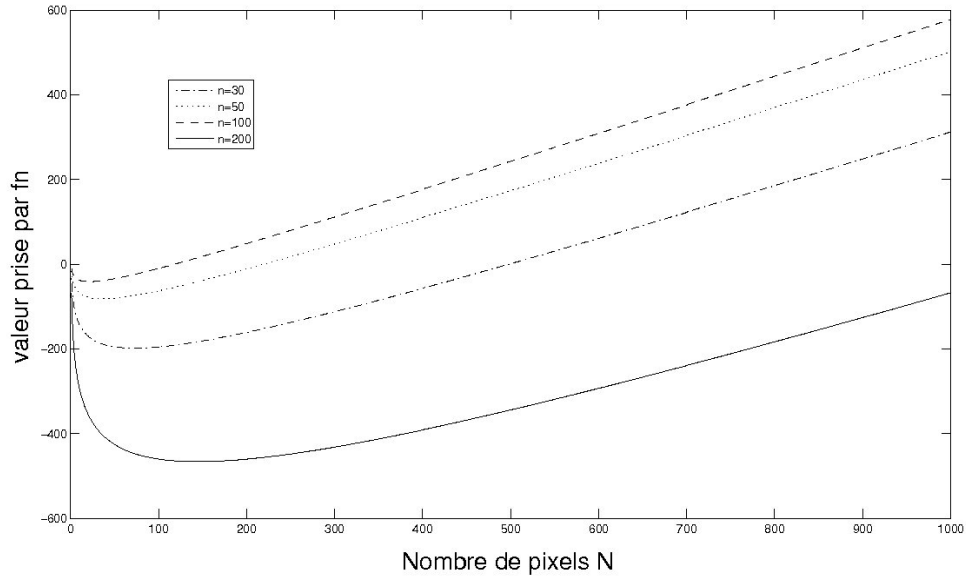


Figure 4.4 – Graphe de la fonction $f_n(N)$ pour différentes valeurs de n

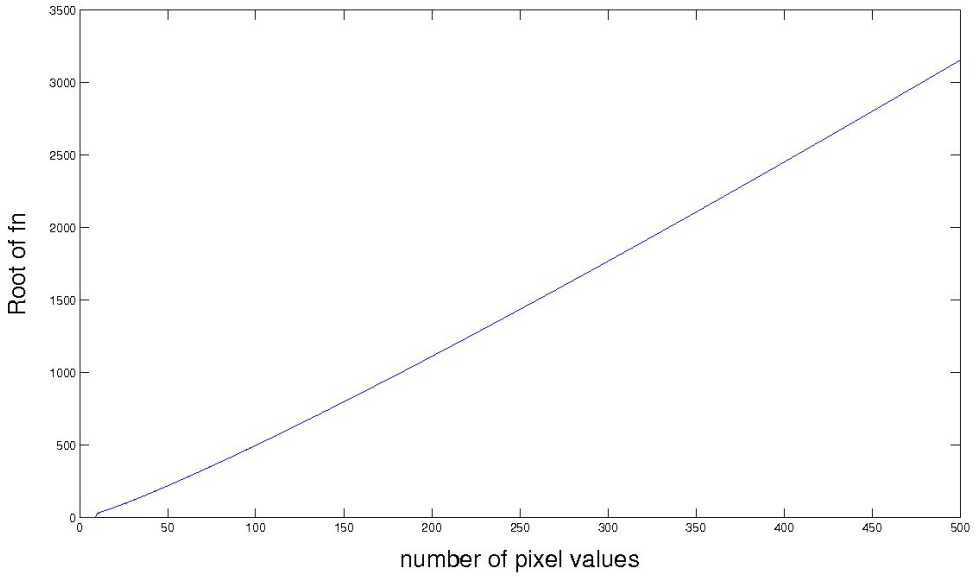


Figure 4.5 – Zeros de la fonction $f_n(N)$ pour $N \geq 1$ en fonction du paramètre n.C es valeurs correspondent, en fonction du nombre de valeurs possibles des pixels, à une valeur plancher de la taille d'apprentissage nécéssaire à un concept pour en faire l'apprentissage.

4.5.2 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 [23] [77]). 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} coïncident 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" .

4.5.3 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 $CS(X, M_\Omega)$.

Etat 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 réseau 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 : $M_\Omega^{initial} = argmax P(X|M_\Omega)$

Evolution

- 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 4.7.
- 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 4.7.
- 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 4.7.

Le modèle qui minimise la complexité est retenu 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 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 $P(L_a = b)$.

4.6 Expériences

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

4.6.1 Données synthétiques

4.6.1.1 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.

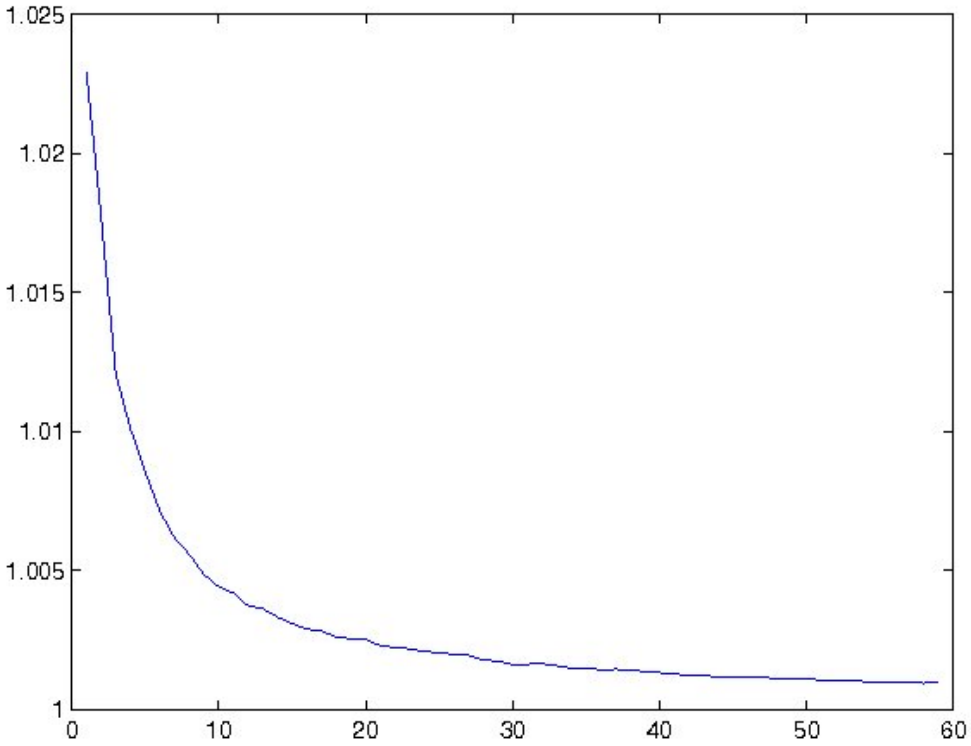


Figure 4.6 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction du nombre N d'images présentes dans X_1 et X_2

— 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 4.6 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 4.6 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.

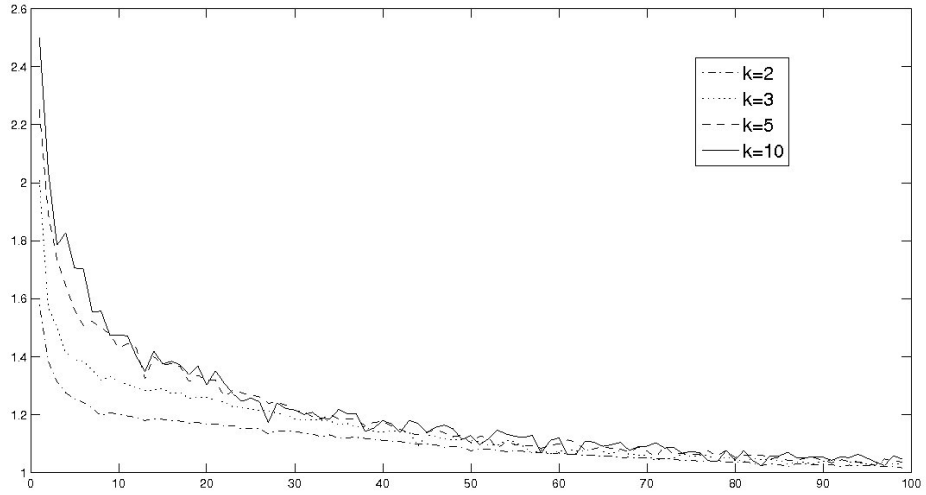


Figure 4.7 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction de σ et pour différentes valeurs de k

4.6.1.2 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 :

- 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 indépendamment les pixels avec probabilité g_i.

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 (voir section 6.4), et le vecteur de paramètres λ_{k+1} est estimé à partir de X_{k+1}.

La figure 4.7 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.

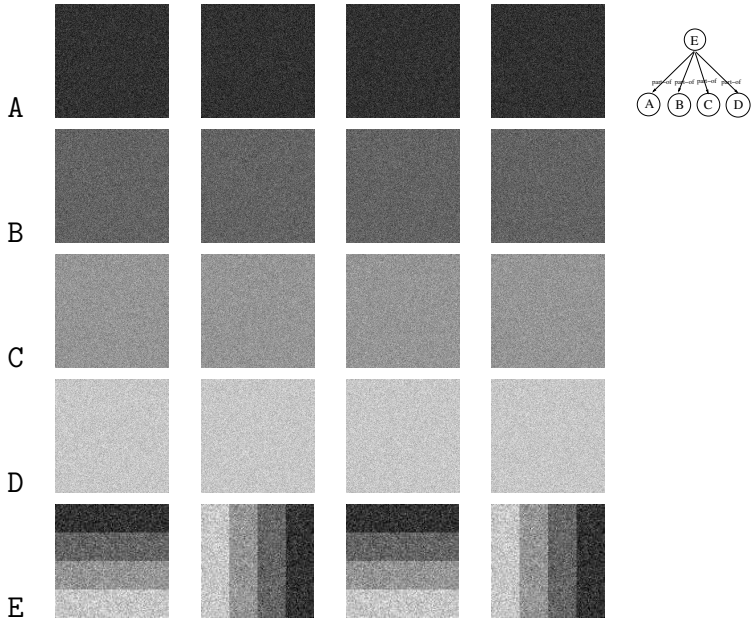


Figure 4.8 – Illustration des expériences de validation du processus de construction automatique d'un réseau de type "part-of" à partir d'une base d'images synthétiques.

Figure 4.9 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction de σ et pour différentes valeurs de k

4.6.1.3 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 4.6.1.3) :

- 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 indépendamment les pixels 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 (voir section 6.4), et le vecteur de paramètres λ_{k+1} est estimé à partir de X_{k+1}.

La figure 4.9 montre le rapport $R_{syn} = \frac{C(X,M}{C(X,M')}$ en fonction de l'écart type des

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

Figure 4.10 – 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

gaussiennes. On voit que le rapport R_{syn} tend vers 1, ce qui signifie qu'introduire la relation d'hyponymie est moins bien reconnue par le modèle. 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.

4.6.2 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 4.14.

4.6.2.1 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 introduits 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 $C(X_c, M')$.

Le rapport $\frac{C(X_c,M)}{C(X_c,M')}$ obtenu pour différents concepts est listé sur le tableau 4.6.2.1. On constate que ce rapport est inférieur 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.

4.6.2.2 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.

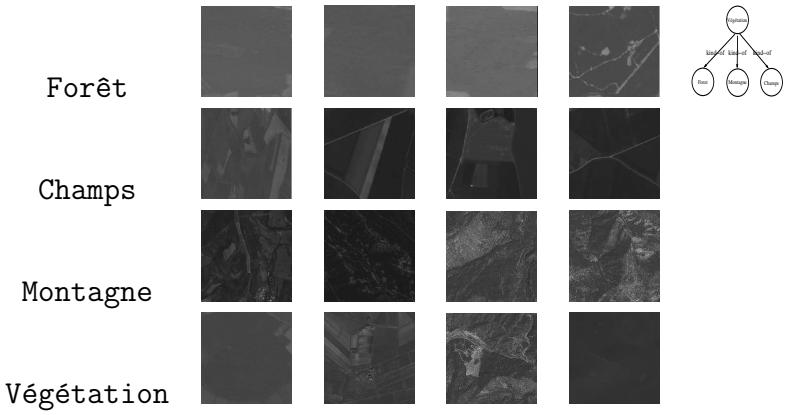


Figure 4.11 – Des images exemple de forêt, de champs et de montagne sont fournis pour l'apprentissage et constituent des concepts correspondant à un seul type de zone. La classe de végétation contient par contre des zones de champs, de montagne, et de forêt fournis en proportions égales. On constate 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

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 : *prairie*, *champs* et *forêt* (voir figure 4.6.2.2). 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 pixels de la manière suivante : étant donné b un pourcentage de bruit, pour tous les pixels, avec une probabilité b, la valeur du pixel va être changée en une nouvelle valeur tirée avec une probabilité uniforme sur l'ensemble des valeurs possibles du pixel. Le rapport $\frac{C(X,M}{C(X,M')}$ est calculé pour différentes valeurs de b. La courbe de résultat est affichée en 4.12. 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

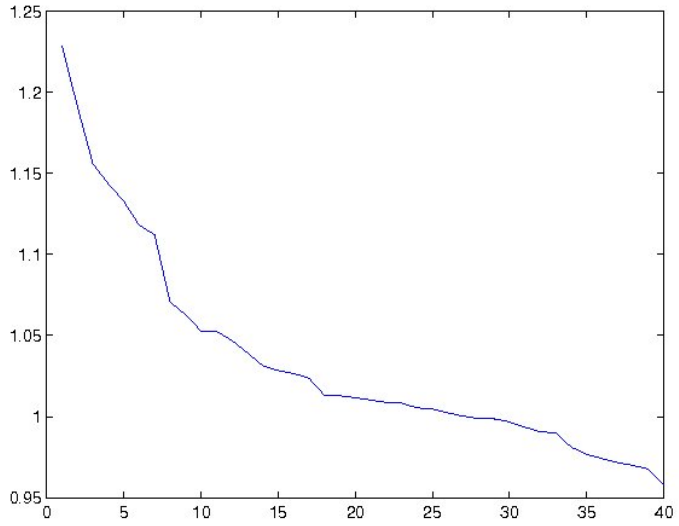


Figure 4.12 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction du pourcentage de bruit montrant la diminution de complexité stochastique entrainée par l'ajout d'un lien d'homonyie enre le concept "végétation" et les concepts : "montagne", "champs", et "forêt".

de bas-niveau.

4.6.2.3 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".

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 X_{k+1} dont les images correspondent à un regroupement de régions qui peuvent être annotées par les 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 méronymique avec les concepts $\{c_1, \ldots, c_k\}$.

Une deuxième 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*. 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 pixels de la même manière que dans le cas de la relation d'hyponymie. Les résultats sont montrés figure 4.13 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.

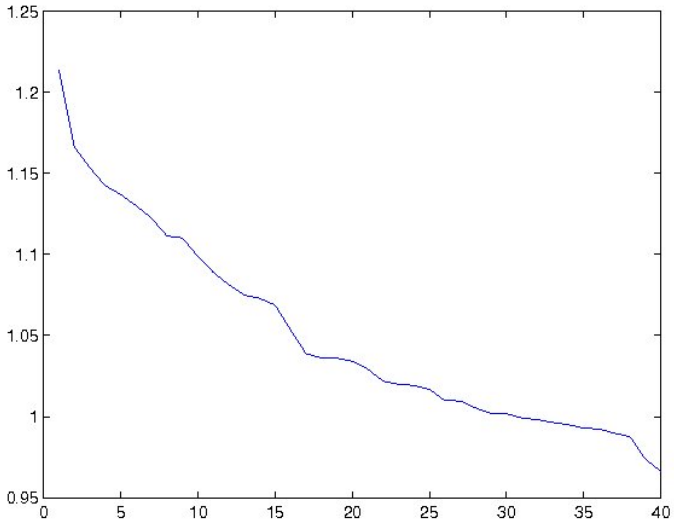


Figure 4.13 – Rapport $\frac{C(X,M}{C(X,M')}$ en fonction du pourcentage de bruit montrant la diminution de complexité stochastique entrainée par l'ajout d'un lien de méronymie enre le concept "zone rurale" et les concepts : "habitations éparses", "champs", et "zone résidentielle".

4.6.2.4 Construction d'un réseau sémantique complet

Expérience sur image 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 4.14). 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 60 (voir chapitre 3, section 1,1). 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 2 couches au maximum.

La complexité stochastique diminue pendant 4 itérations puis remonte (voir figure 4.16). En 4.3.4, 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 Pékin. La base d'apprentissage comprend 32 concepts listés dans le tableau 4.17. La construction du réseau sémantique est limitée à 2 couches au maximum. Le seuil sur la probabilité de génération d'un concept pour la création d'un concept sémantique

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

Figure 4.14 – Présentation de la base de données utilisée pour faire l'apprentissage du réseau sémantique

Concepts	**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

Figure 4.15 – Liens existants entre les concepts de la deuxième et de la première couche

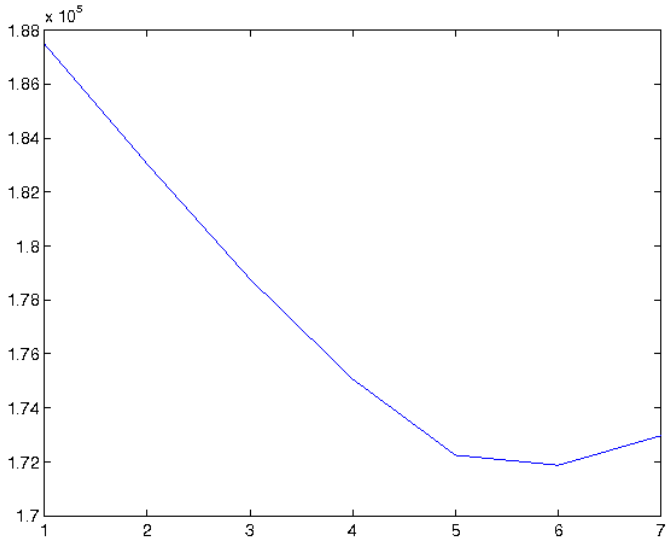


Figure 4.16 – Evolution de la complexité stochastique en fonction du nombre d'itérations de l'algorithme d'apprentissage du réseau sémantique sur la base d'images SPOT5

est imposée à 0.01. La complexité stochastique diminue pendant 5 itérations puis remonte (voir figure 4.18). 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 concepts qui sont mis en deuxième couche sont : Aéroport, Zone rurale, Complexe industriel, Jardin public, Zone urbaine éparse, Zone résidentielle, Zone pavillonnaire.

4.7 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. Le critère de minimisation de la complexité stochastique permet ainsi, étant donné une base de données annotée par un lot de concepts, de déterminer automatiquement un réseau sémantique contenant ces concepts. 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.

Jardin	Zone commerciale
Grandes tours	Zone résidentielle
Serres	Zone pavillonnaire
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

Figure 4.17 – Présentation des concepts utilisés pour faire l'apprentissage du réseau sémantique

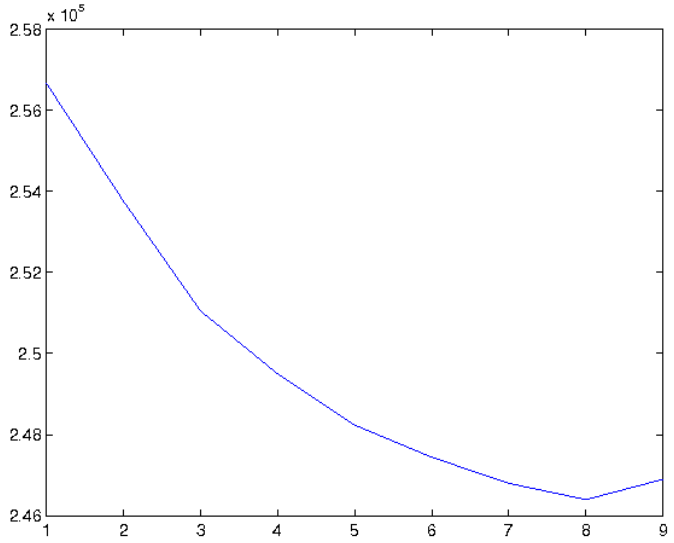


Figure 4.18 – Evolution de la complexité stochastique en fonction du nombre d'itérations de l'algorithme d'apprentissage du réseau sémantique sur la base d'images Quickbird

Chapitre 5

Annotation d'images tests.

5.1 Méthode d'annotation sémantique d'une image test

Nous étudions dans cette section comment obtenir une représentation pertinente de l'image à l'aide des réseaux sémantiques, en étudiant l'apport d'une représentation multi-couche de l'image. Cette représentation sera obtenue par inférence statistique d'annotations en utilisant le modèle stochastique construit au chapitre précédent.

Avant de décrire le mécanisme d'annotation sémantique que nous allons mettre en œuvre, essayons de raisonner en rapport avec 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 donné 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 présent 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 Algorithme d'inférence

On souhaite trouver deux lots d'annotations 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 la couche 1, et G_2 avec des concepts de la couche 2. Étant donné une image I, on souhaite trouver les partitions annotées G_1 et G_2 maximisant la probabilité

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

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

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

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 $P(G_1|I)$. Puis, on optimise le terme $P(G_2|G_{1,opt})$.

L'image est tout d'abord annotée avec les concepts de la première couche selon l'algorithme décrit dans la section 3.4.1. 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 3.4.1 :

Première étape d'inférence

— Initialisation de l'algorithme :
Une partition annotée initiale est créée en utilisant les modèles M_c associés aux modèles $c \in C_1(S^{po}_\Omega)$ de la manière suivante :
Pour chaque pixel 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 pixel (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 pixel (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 pixels 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 :

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

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 pixels 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 :

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

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

5.1.2 Représentation sémantique de l'image

L'algorithme détaillé précédemment fournit 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'. Soit P_{ko}
— 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^{po} un autre ensemble de régions qui ne définit pas nécessairement une partition et qui contient les concepts du réseau sémantique S^{ko}. En effet, la relation sémantique "A est kind-of de 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 la première région π_i.
— Périmètre de la deuxième région π_j.
— Périmètre commun entre les régions π_{ij}

— Ratio du périmètre commun au périmètre de la première région : $r_{ij}^1 = \pi_{ij}/\pi_i$
— Surface de la première région σ_i.
— Surface de la deuxième région σ_j.
— Surface commune entre les régions σ_{ij}
— Rapport de la surface commune et de la surface de la première région : $r_{ij}^2 = \sigma_{ij}/\sigma_i$

Les relations spatiales entre R_i et R_j se définissent alors par une représentation floue selon des fonction d'appartenance particulières. 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 labellisés par les relations spatiales énoncées précédemment.

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 scientifiques 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 (Évaluation des systèmes de Transcription des émission Radiophoniques). 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 Co-

rel5K [53], [62], [72]. 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 pixels 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 de segmentation avec une vérité terrain peuvent être employés comme la distance de Vinet (voir [12]). 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 :

$$WER = \frac{Elisions + Ajouts + Substitutions}{NombreTotalDeMots} \tag{5.1}$$

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 :

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

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

En référence à la métrique WER, on définit alors une métrique que l'on appelera IAWER (Image Adapted World Error Rate) :

- Une élision comme 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 5.1 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 WER. 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 le taux de recouvrement des régions. Soit la région $R_{i,opt}^{jr}$ de la vérité terrain correspondant au taux de recouvrement optimal.
  - En cas de conflit : 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 optimal.
  - Sinon, R_i^j est appariée avec $R_{i,opt}^{jr}$.

Dans l'évaluation de l'annotation qui sera faite en 5.2.3, deux métriques seront utilisées : la distance de Vinet, et la distance IAWER.

5.2.2 Expériences

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 concepts utilisés pour l'annotation sont ceux listés en 4.14 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 concepts, 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 4.26 et 4.27. Les paramètres du modèle markovien sont appris sur cette base d'apprentissage en utilisant la méthode du gradient stochastique (voir section 5.3.2.5).

Résultats Les images sont annotées en utilisant la méthode décrite en 6.4. 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 5.3.1 avec un coefficient de regroupement fixé à 0,85 pour permettre un appariement entre régions. Comme nous l'avons vu en 5.3.1, 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.

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 5.1). 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 5.2), ce qui peut s'interpréter par le fait que cette optimisation au niveau des sites individuels tend à entrainer la création d'un certain nombre de petites régions qui ne correspondent pas à des régions sémantiques identifiables. Tandis que la méthode bayésienne naïve, utilisant une force de rappel pour créer les régions qui est la loi de Poisson, obtient ainsi de meilleures performances.

Les résultats sont nettement plus contrastés en ce qui concerne les résultats du processus d'annotation avec les concepts appartenant à la deuxième couche (voir tableau 5.3), 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. Des images exemples sont montrés sur les images 5.4 et 5.5.

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

FIGURE 5.1 – Résultats d'annotation estimés avec la mesure de Vinet sur la base SPOT5

	élisions	Additions	Substitutions	IAWER
Modélisation markovienne	15	63	5	9,79 %
Modélisation hiérarchique	31	29	16	8,91 %

FIGURE 5.2 – Résultats d'annotation estimés avec la IAWER pour la première couche d'annotation sur la base SPOT5

5.2.2.1 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 4.17 et sont placés dans un réseau sémantique à deux couches, qui est celui obtenu en 4.7.2.4.

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

Résultats Comme pour la base de données d'images SPOT5, les images sont annotées en utilisant la méthode décrite en 6.4. 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 5.3.1. Les résultats quantitatifs sont présentés tableaux 5.6, 5.7 et 5.8. Des résultats visuels sont montrés figures 5.9, 5.10 et 5.11.

Des conclusions relativement analogues à celles tirées dans la section 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.

	élisions	Additions	Substitutions	IAWER
Modélisation markovienne	11	24	3	23.14 %
Modélisation hiérarchique	9	8	3	12,34 %

FIGURE 5.3 – Résultats d'annotation estimés avec la IAWER pour la deuxième couche d'annotation sur la base SPOT5

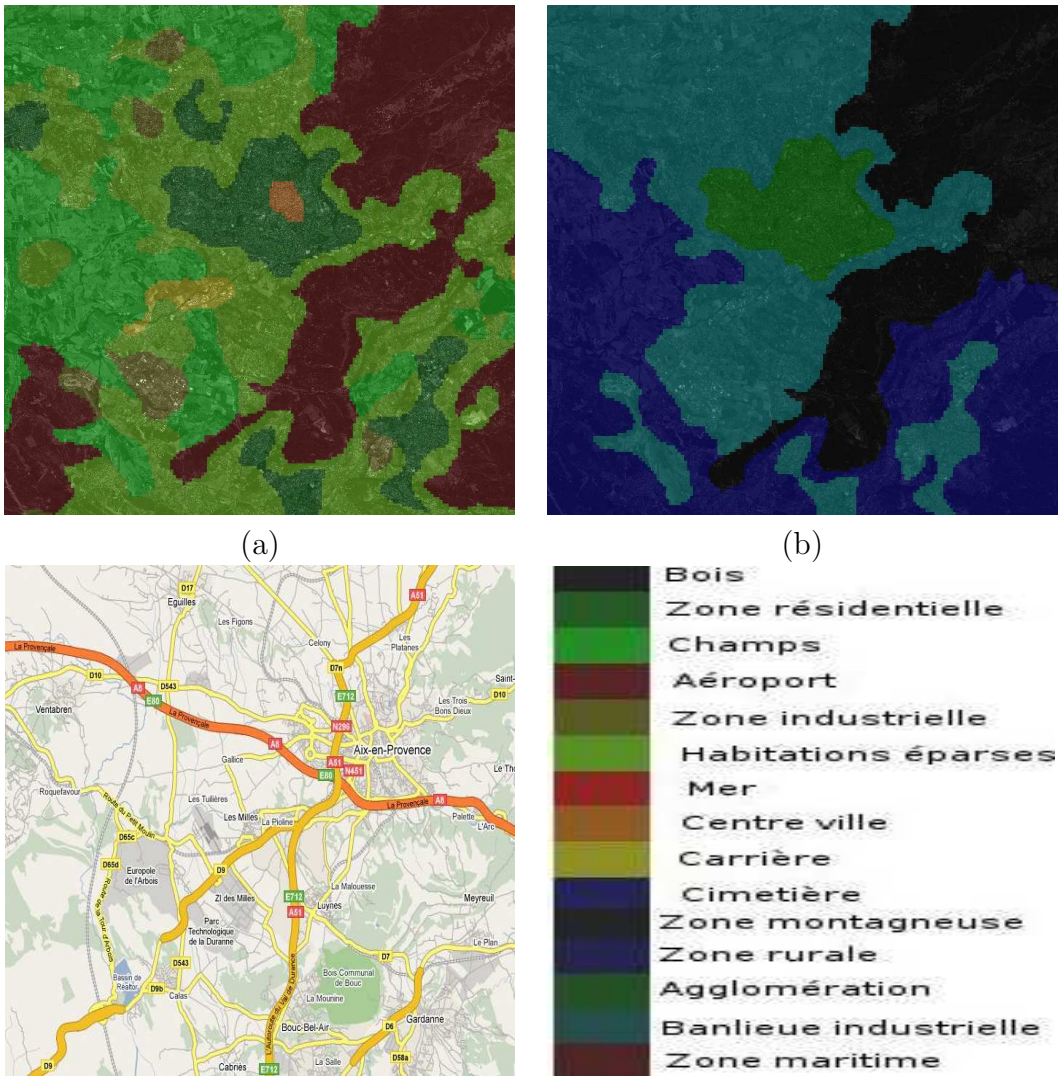
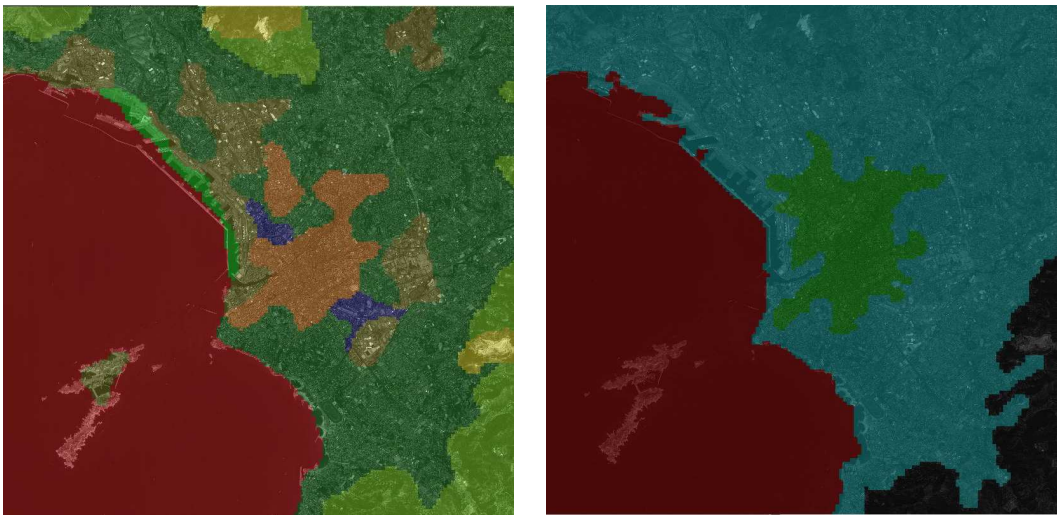
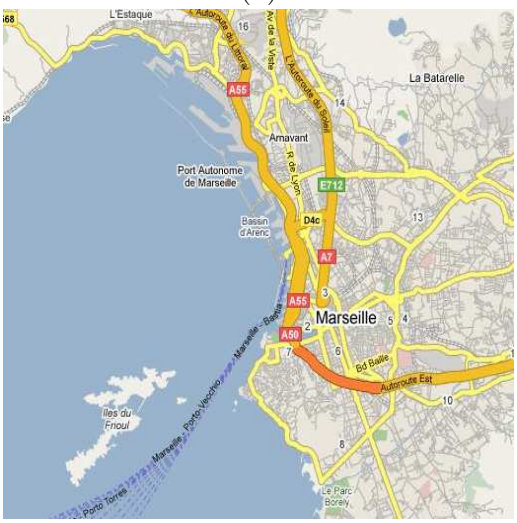


Figure 5.4 – Image SPOT5 à 2,5m de résolution 6000×6000 de la région d'Aix en Provence. (a) Première couche d'annotation (b) Deuxième couche d'annotation (c) Carte google-map de la zone correspondante



(a)

(b)



(c)

FIGURE 5.5 – 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) Carte *google-map* de la zone correspondante

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

Figure 5.6 – Résultats d'annotation estimés avec la mesure de Vinet sur la base Quickbird

	élisions	Additions	Substitutions	IAWER
Modélisation hiérarchique	20	25	6	21,39 %
Modélisation markovienne	14	40	7	25,02 %

Figure 5.7 – Résultats d'annotation estimés avec la IAWER pour la première couche d'annotation sur la base Quickbird

Ceci conforte l'idée d'utiliser une inférence basée sur des concepts sémantiques pour franchir le fossé sémantique.

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. On suppose que cette requête contient une règle sur les concepts recherchés et l'agencement spatial des régions (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 sélectionner les meilleures hypothèses parmi une ensemble de groupes de régions annotées en réponse à cette requête.

5.3.1 Fonction de cohérence

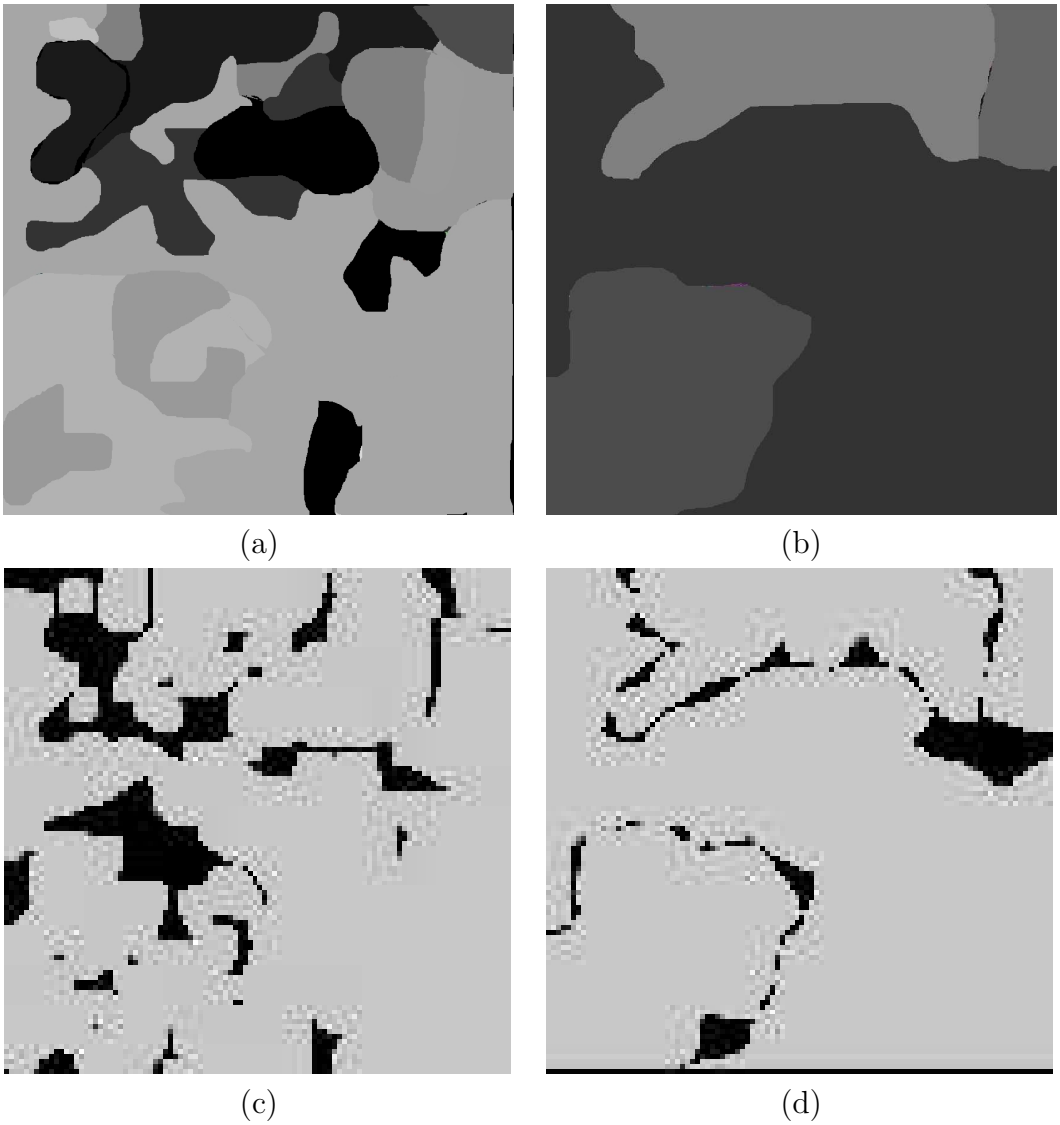
étant donné un graphe de N régions correspondant à l'annotation d'une zone d'une image de la base de données, on mesure l'adéquation entre une partition annotée P_s et une requête Req formulée par un utilisateur en utilisant la fonction de cohérence définie dans [47] :

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

où :

	élisions	Additions	Substitutions	IAWER
Modélisation hiérarchique	9	8	0	20,99 %
Modélisation markovienne	13	15	2	37,04 %

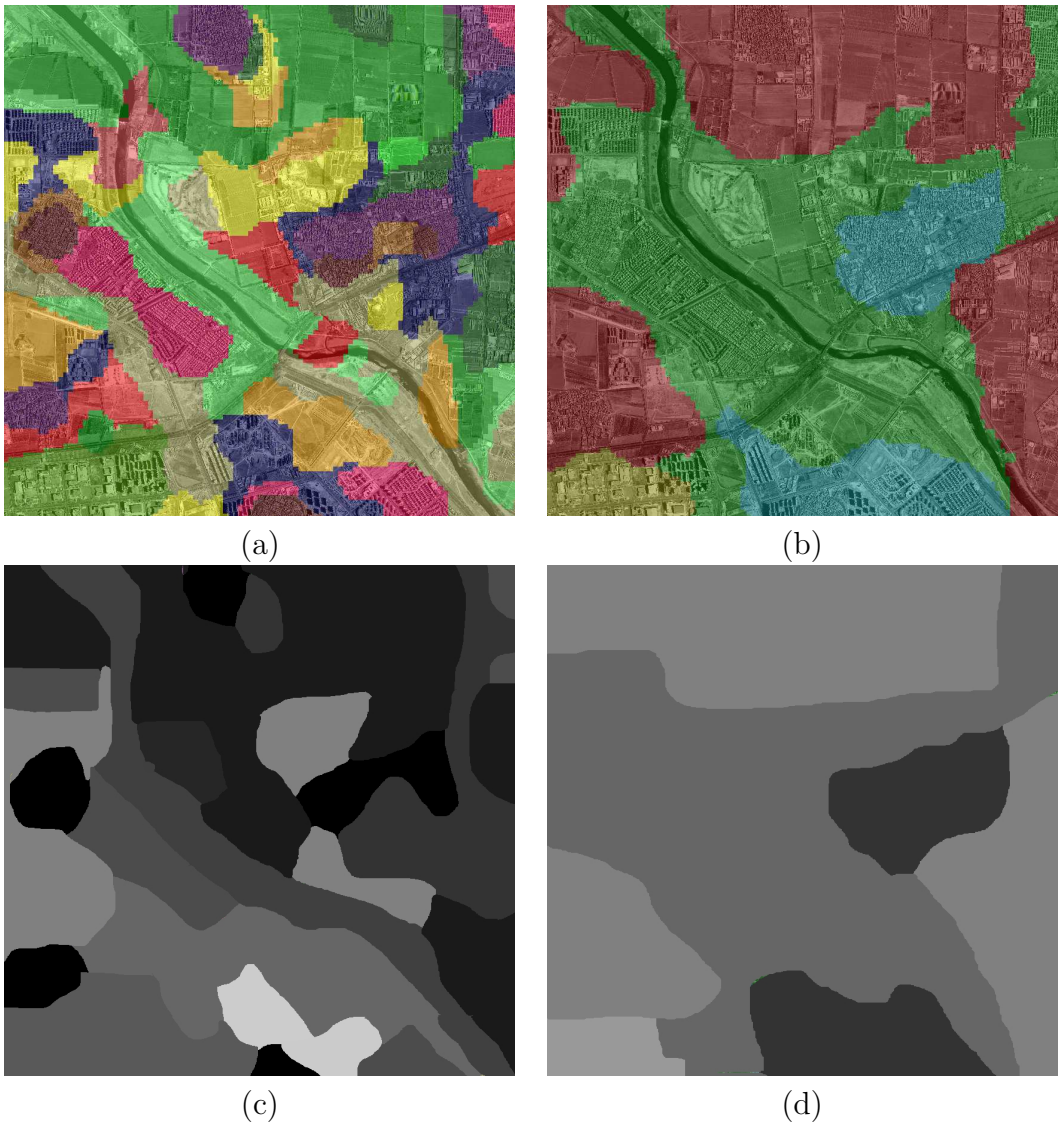
Figure 5.8 – Résultats d'annotation estimés avec la IAWER pour la deuxième couche d'annotation sur la base Quickbird



(a) (b)

(c) (d)

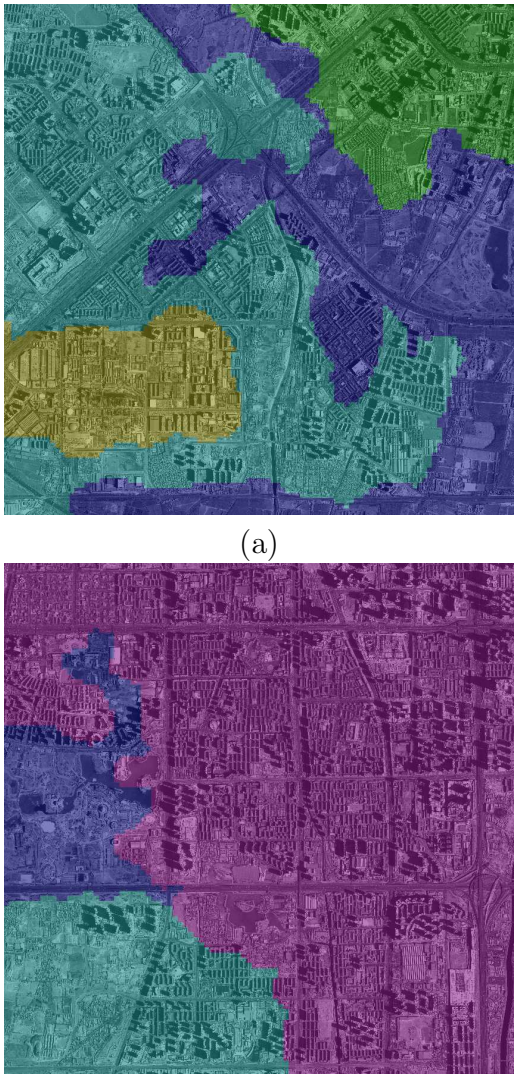
Figure 5.9 – Résultats d'annotation sur une image Quickbird de Pékin (a) Masque de la vérité terrain, couche 1 (b) Masque de la vérité terrain, couche 2 (c) Pixels mal annotés de la couche 1 (d) Pixels mal annotés de la couche 2



(a) (b)

(c) (d)

Figure 5.10 – (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 pavillonnaires, le bleu foncé aux jardins publics, le bleu clair aux zones urbaines éparses.



(a)

(b)

Figure 5.11 – Annotation produite par le système pour la couche 2 sur deux images Quickbird. 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.

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

— $P(R_i)$ 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 décroît avec le nombre de pixels 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 pixels 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"), $Eval(R_1, \mathbf{R}, R_2) = 1$.

5.4 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 Ω de taille i.

— Initialization : $\omega_0 = \emptyset$

— 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 : $-logP(D|M_{\omega'_{i+1,j}})$

  — We define $SC_i = -logP(D|M_{\omega_{i+1}}) = \min_{\omega'_{i+1,j}}(-logP(D|M_{\omega'_{i+1,j}})$

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

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 %

Figure 5.12 – Pourcentage de couverture des zones couvertes par les régions annotées par les différents concepts dans la base de données considérée.

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 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'Anger 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. 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

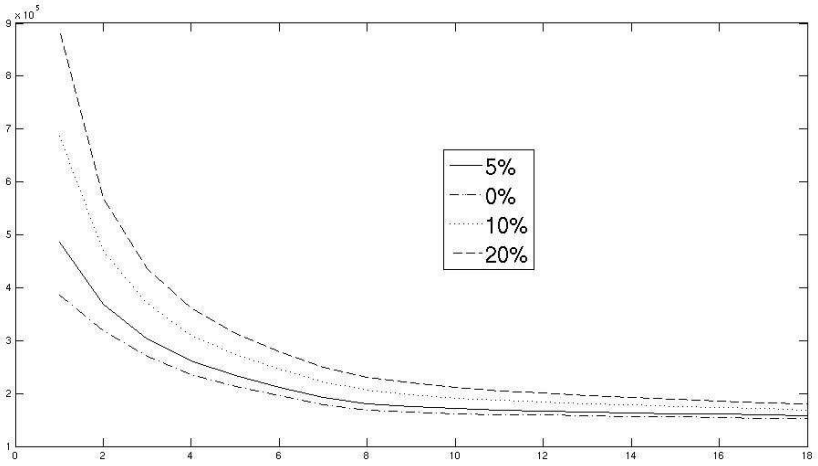


Figure 5.13 – Log vraisemblance de la base de données en fonction du nombre de concepts inclus dans le modèle statistique pour différentes valeurs du bruit ajouté sur les caractéristiques.

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 l'information présente dans l'image qui est essentielle pour l'utilisateur, à savoir l'information sémantique. La représentation présentée en 7.1.2 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 supplémentaires 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 nécessitent 192 octets. Ainsi, une représentation en une couche nécessite donc 4,8ko. Ce qui fait une compression d'un facteur 7500 conservant l'essentiel de l'information sémantique présente dans l'image.

5.6 Conclusion

5.6.1 Part d'innovation dans le travail effectué

Dans ce rapport, nous avons présenté 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 concepts servant à l'annotation, et d'un modèle statistique permettant de coder les images de la base d'apprentissage. Les outils utilisés sont des outils relativement classiques dans le domaine du traitement d'images et de la reconnaissance de formes : modèles à variables latentes, critère de Minimisation de la Complexité Stochastique, réseaux sémantiques etc. Cependant, l'innovation apportée dans ce travail se situe essentiellement au niveau de la modélisation, 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. La méthode proposée se situe 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.

— 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 donné par un expert.

Notre approche permet de combiner flexibilité et prise en compte des relations sémantiques en utilisant un réseau sémantique mis en relation avec un modèle statistique. Le critère de minimisation de la complexité stochastique est employé de manière à trouver le meilleur modèle dans l'espace des paramètres et des structures admissibles du modèle.

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

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

Afin d'effectuer 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 relativement restrictif. Il est possible d'envisager un réseau de relations entre les concepts qui serait beaucoup plus complexe, en s'affranchissant par exemple de la modélisation en couches successives et qui pourrait ainsi décrire la réalité de manière beaucoup plus fine. La modélisation qui en résulterait s'en trouverait 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. Parfois, 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 sure 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. Il semble donc qu'il faille attendre des progrès significatifs dans le domaine de l'estimation des paramètres avant de pouvoir prendre en compte des structures moins contraintes.

5.6.2.2 Introduction d'information spatiale

Dans le modélisation qui est proposée, le modèle de base est le modèle bayésien naïf. Ce modèle ne prend pas en compte le contexte de chaque pixel. Or, même si, dans notre exemple, les pixels correspondent à des caractéristiques de bas-niveau qui contiennent une information sur les relations spatiales entre les pixels, il est fort probable qu'une information très importante soit contenue dans les relations spatiales entre les sites où sont extraites les caractéristiques de bas-niveau. Il serait ainsi intéressant de pouvoir améliorer notre modélisation en prenant en compte cette information. Cependant, une difficulté importante de cette prise en compte de l'information spatiale 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 pixels dans les régions. Si l'on souhaite introduire, des relations spatiales, les relations spatiales existantes entre régions situées en deuxième couche seraient différentes que les relations de voisinages définies entre pixels, et une information différente serait alors codée selon le placement dans la hiérarchie du concept d'annotation.

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 ([43], [46]). 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 pixels. En effet, notre objectif n'est pas de faire de matching d'objets, mais simplement d'avoir une caractérisation 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.0.0.1 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 pixels 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 pixels d'une nouvelle image dont les niveaux de gris sont les indices des codewords dans le codebook (voir figure 4.2). Nous appellerons dans la suite ces vecteurs quantifiés "textons".

A.0.0.2 Regroupement en patchs de descripteurs SIFT

Modélisation utilisée Les descripteurs SIFT apportant à eux seuls une information beaucoup trop locale, ils sont regroupés en ce que nous appellerons 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.

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 :

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 :

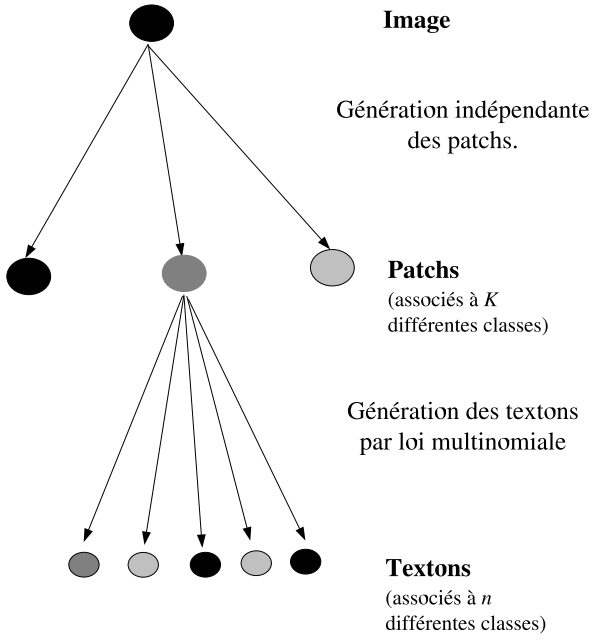


Figure A.1 – Schéma de classification des patchs utilisé

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

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 :

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

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 :

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

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

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 $P(Z_i|O_i)$ indépendamment 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 pixels. Un patch a une taille de 40 pixels de côté, il contient ainsi 64 textons, un texton pouvant avoir 25 valeurs possibles.

Annexe B

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 discrétisées comme une nouvelle image dont les niveaux de gris de chaque pixel sont les indices auxquels est attaché chaque vecteur de caractéristiques. Notons qu'aucune comparaison n'est possible entre les "pixels" 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 pixels par un loi de mélange. Étant donné une région, une hypothèse d'indépendance des pixels sachant le concept sémantique est supposée. Seul l'histogramme des valeurs des pixels est ainsi pris en compte. Si la spatialité entre les pixels n'est pas prise en compte, notons cependant que chacun de ces pixels est un vecteur de caractéristiques quantifié. A ce titre, le pixel contient en lui même une information sur la répartition spatiale des pixels 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 tâche 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.

Étant 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 pavillonnaire, 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'image. Nous évaluerons ainsi cette approche sur des images panchromatiques SPOT5 en utilisant des caractéristiques de Haralick.

B.1 Principe de la méthode

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

B.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 pixels que nous allons relier aux concepts sémantiques. L'ensemble des pixels 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 pixels qui sont supposés être générés par cette région et que nous notons ici $S_l(O_I)$. Nous supposons pour le moment que chaque pixel 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 pixels sachant l'annotation de l'image.

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

$P(O_I)$ 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 pixels sont supposés ne dépendre que de la région qui les a générés, par conséquent, on peut écrire :

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

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 pixels dans l'image d'indice i du lot d'apprentissage associé à un concept
N_{ij}	Nombre de pixels 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 pixel de valeur j dans le modèle d'indice k du mélange associé à un concept s
N_s	Nombre de pixels de la région sémantique S
j	indice de la valeur d'un pixel
O	Ensemble de pixels 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

Figure B.1 – Notations utilisées dans ce chapitre

B.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 pixels 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 pixel 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 pixels O_i de la façon suivante :

- Le modèle k est choisi avec probabilité π_{ks}.
- Le nombre N_i de pixels de la région est généré avec une loi de Poisson de paramètre λ_s.
- Chaque pixel de la région est choisi indépendamment des autres avec probabilité p_j^{ks}, où j correspond à la valeur du pixel.

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

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

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 pixels étant donné le concept s et la variable latente.

En remplaéant cette probabilité par son expression dans l'équation B.3, on peut écrire :

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

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

B.2 Apprentissage du modèle

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

B.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 :

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

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 B.5, et en conditionnant par rapport aux valeurs possibles de la variable latente, on obtient, N_i étant le nombre de pixels dans le patch i :

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

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 pixels 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 B.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 pixels 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 :

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

Notons que l'estimation de p_j^{ks} correspond, comme on pouvait s'y attendre, au rapport des occurrences de pixels de valeur j présents dans le lot d'apprentissage sur le nombre total de pixels, 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.

B.2.2 Complexité Stochastique

Le principe de minimisation de la complexité stochastique a été introduit par Rissanen en 1978 [59]. 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. Il a été appliqué dans différents types de problèmes en traitement d'images et a offert de bons résultats.

B.2.2.1 Principe de la minimisation de la CS

Le principe de minimisation 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 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 [59] 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 n'apparaét pas tellement facile 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 posée par Shannon en 1943 [64]. 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 connaître la loi de probabilité $P(X)$, c'est à dire le modèle, pour reconstruire cette suite.

C'est pour cela que 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.

B.2.2.2 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 [60] 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 modèle $M = \{M_k\}_{k=1}^{K}$, 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.

B.2.3 Modélisation utilisée

Soit M_s le modèle utilisé pour décrire le lot d'apprentissage X_s pour le concept s. La longueur de code pour décrire les données peut être séparée en deux termes :

$$C(X_s, M_s) = C(X_s|M_s) + C(M_s) \tag{B.8}$$

M_s étant un lot de paramètres réels, la longueur de code est en principe infinie. Cependant, les paramètres sont estimés avec un nombre fini d'échantillons, Rissanen suggère dans [60] l'expression suivante pour $C(M_s)$:

$$C(M_s) = \sum_{i=1}^{n_s} \frac{\alpha}{2} log(N_i)$$

où α correspond au nombre de paramètres à coder, et N_i correspond au nombre de pixels avec lequel a été estimé les paramètres du modèle i.

Pour les paramètres de probabilité de génération des pixels, notons que la propriété $\sum_{i=1}^{N} p_i^{js} = 1$ est vérifiée pour tout $j\epsilon\{1, ..., k_s\}$, étant donné que p_i^{js} est un lot de probabilités. Ainsi, $N-1$ paramètres doivent être codés pour chaque modèle d'indice k pour les probabilités de génération des pixels. De même, la probabilité à priori de chaque modèle π_k, la relation suivante est vérifiée : $\sum_{i=1}^{K_s} \pi_{is} = 1$. Ainsi, $K_s - 1$ paramètres seulement doivent être codés.

En ce qui concerne le terme d'attache aux données $C(X_s|M_s)$, Shannon propose la formule suivante, liant directement la longueur de codage de la séquence à sa probabilité d'apparition ([64]) :

$$C(X_s|M_s) = -logP(X_s|M_s)$$

En utilisant l'expression de $P(X_{si}|M_s)$ écrite dans B.4, l'équation B.8 peut être écrite de la façon suivante :

$$C(X_s, M_s) = -log(\sum_{i=1}^{n_s}\sum_{j=1}^{k_s}\pi_{js}\prod_{i=1}^{N_i} p_{m(l_i)}^{js}) + \sum_{i=1}^{K_s}\frac{N-1}{2}log(\sum_{k=1}^{n_s}\gamma_k N_k) + \frac{K_s-1}{2}log(n_s) \tag{B.9}$$

Le lot de paramètres que nous choisirons sera ainsi celui qui minimise cette expression.

B.2.4 Apprentissage non supervisé des paramètres

B.2.4.1 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 B.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 :

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

B.3 Annotation d'images

B.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 pixel l de l'image I est lié à un concept en prenant en compte sa valeur, ainsi que la valeur des pixels de son voisinage $NE(l)$ en choisissant le concept s qui minimise la quantité (cf équation B.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 pixel l. Le voisinage $NE(l)$ est défini comme l'ensemble des pixels 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 pixels 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 B.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.

B.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 précisément les paramètres. Nous avons fait des évaluations visuelles pour les deux bases d'images à disposition.

B.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 côté centrée sur la municipalité de Pékin. Les concepts suivants sont utilisés pour faire une évaluation visuelle : zone urbain dense, zone résidentielle pavillonnaire, 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.

B.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 franéaises : 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 B.3.2.2), de taille variant de 250×250 à 1000×1000.

$$\begin{pmatrix} & ZI & ZR & CV & ZC & ZB & C & Ch & ZM \\ ZI & 29 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ ZR & 1 & 25 & 0 & 0 & 0 & 0 & 0 & 0 \\ CV & 0 & 1 & 15 & 0 & 0 & 0 & 0 & 0 \\ ZC & 0 & 0 & 0 & 9 & 0 & 0 & 0 & 0 \\ ZB & 0 & 0 & 0 & 0 & 12 & 0 & 1 & 0 \\ C & 0 & 0 & 0 & 0 & 0 & 8 & 0 & \\ Ch & 0 & 0 & 0 & 0 & 0 & 0 & 35 & 0 \\ ZM & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 17 \end{pmatrix}$$

Figure B.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 B.3.2.2. On obtient au total $96,7\%$ de bonne classification.

Test d'auto-cohérence 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 extraites 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 comme apprentissage pour estimer les modèles de mélange de ces 5 concepts. Chacun des 5 modèles de mélange ne contient évidemment qu'un seul lot de paramètres, et les tailles sont toutes fixées à une même taille de 200 pixels. Le résultat est exhibé figure B.4. Les différentes régions sont retrouvées tout à fait correctement dans l'image.

Protocole d'apprentissage Nous choisissons pour effectuer l'apprentissage un ensemble de 60 imagettes parmi la base de données extraites présentées précédemment

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

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 B.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.

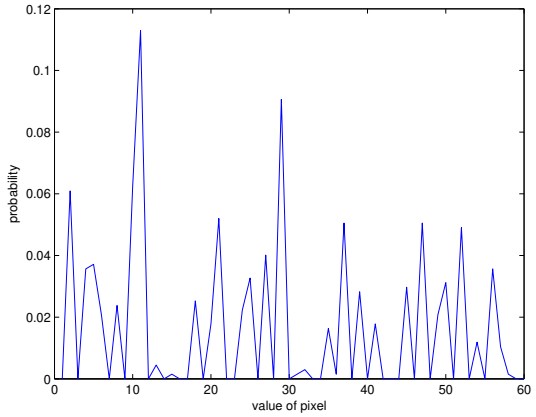
Analyse des résultats

- 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.
- 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

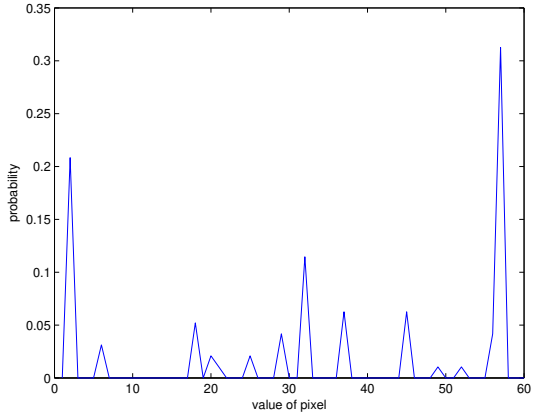
Catégorie	**concept**	**nombre d'imagettes**
Végétation	carrière	13
	bois	18
	champs	37
	montagne	17
	prairie	3
Eau	mer	3
	lac	2
	bassin	2
Urbain	aeroport	4
	centre ville	16
	cimetière	1
	marais salant	1
	port de plaisance	10
	port industriel	1
	raffinerie	4
	village	29
	zone commerciale	9
	zone industrielle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

Figure B.3 – Base de données constituée à partir de la base d'images SPOT5

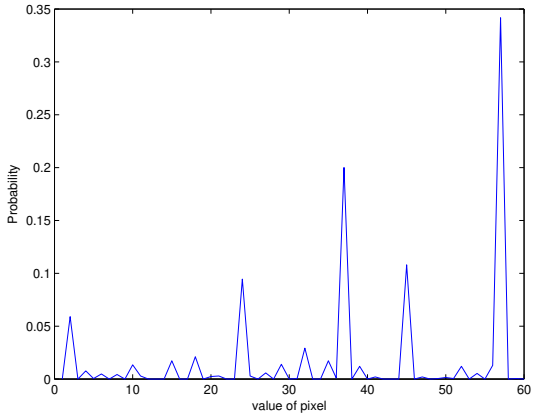
associée au concept aéroport (voir figure 3.17) est beaucoup plus proche de la densité uniforme que d'autres densité. Par conséquent, ce concept a tendance à annoter par défaut des zones correspondant relativement à aucune autre sémantique. On peut en conclure que le concept "aéroport" correspond à des 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 chacune 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 concept "aéroport" est probablement trop important dans la modélisation que nous employons dans ce chapitre.



(a)



(b)



(c)

Figure B.4 – (a) : densité de probabilité associée au concept 'aéroport'. (b) : densité de probabilité associée au concept 'centre ville'.(c) : densité de probabilité associée au concept 'zone résidentielle'.

Bibliographie

[1] K.S. Pedersen A.B. Lee and D. Mumford. The non-linear stastics of high-contrast patchesin natural images. *IJCV*, 54 :83–103, 2003.

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

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

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

[5] 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.

[6] J.R. Bellagarda. Latent semantic mapping. *IEEE Signal processing magazine*, 22(5) :1349–1380, Sept 2005.

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

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

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

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

[11] Marine Campedel, Bin Luo, Henri Maitre, Eric Moulines, Michel Roux, and Ivan Kyrgyzov. *Indexation des images satellitaires*. December 2004.

[12] Sebastien Chabrier, Bruno Emile, Christophe Rosenberger, and Helene Laurent. Unsupervised performance evaluation of image segmentation. *EURASIP J. Appl. Signal Process.*, 2006(1) :217–217, uary.

[13] E. Chang, G. Kingshy, G. Sychay, and G. Wu. Cbsa : content-based soft annotation for multimodal image retrieval using bayes point machines, 2003.

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

[15] International conference in computer vision (ICCV03), editor. *Video google : A text retrieval approach to object matching in videos*, volume 2, May 2003.

[16] D.A. Cruse. *Lexical Semantics*. Cambridge Univ Press, 1986.

[17] F. Cutzu, R. Hammoud, and A. Leykin. Distinguishing paintings from photographs. *Comput. Vis. Image Underst.*, 100(3) :249–273, 2005.

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

[19] 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.

[20] H. Elghazel and A. Baskurt. Approches textuelles pour la recherche d'images. In *SETIT, 3rd international Conference : Sciences of Electronic, Technologies of Information and Telecommunications. Sousse. Tunisia*, March 2005.

[21] H. Murase et S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of computer vision*, 14 :5–24, 1995.

[22] G. Sagerer F. Kummert H. Niemann and S. Schrder. *Werkzeuge zur modellgesteuerten Bildanalyse und Wissensakqusition–Das System ERNEST*, pages 556–570. Springer-Verlag, 1987.

[23] Song-Chun Zhu Feng Min, Jin-Li Suo and Nong 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.

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

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

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

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

[28] G. Giacinto and F. Roli. Nearest prototype relevance feedback for content based image retrieval. In *Proc. of international conference on pattern recognition (ICPR)*, 2004.

[29] Z.Q. Liu H. Chen, Z.J. Xu and S.C Zhu. A high resolution grammatical model for face representation and sketching. *Proc IEEE conf on CVPR*, June 2005.

[30] Z.Q. Liu H. Chen, Z.J. Xu and S.C Zhu. Composite templates for cloth modeling and sketching. *Proc of International conference of Pattern Recognition on Computer Vision*, June 2006.

[31] F. Han and S.Chun Zhu. Primal sketch : integrating texture and structure. *Proc of International conference on computer vision*, 2005.

[32] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR-99, pp 35-44*, 1999.

[33] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning journal*, 42(1) :177–196, 2001.

[34] H. Maitre I. Kyrgyzov and M. Campedel. Kernel mdl to determine the number of clusters. submitted to MLDM 2007.

[35] Lyons J. *Eléments de sémantique*. Larousse Université, Paris, 1978.

[36] 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.

[37] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th international ACM SIGIR Conference*, pages 119–126, 2003.

[38] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *Proc IEEE Conference on computer vision and pattern recognition*, June 2006.

[39] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290 :91–97, 1981.

[40] G. Kanisza. *Organization in vision*. Praeger, 1974.

[41] G. Kanisza. General pattern theory. 1993.

[42] D. Kunz, K. Schilling, and T. ogtle. A new approach for satellite image analysis by means of a semantic network, 1997.

[43] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[44] D. Marr. *Vision*. Freeman Publisher, 1983.

[45] A. Meillet. Comment les mots changent de sens. *L'Annociologique*, 9, 1905.

[46] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR*, 2003.

[47] C. Millet. *Annotation automatique d'images : annotation cohnte et crion automatique d'une base d'apprentissage*. PhD thesis, École nationale supeure des tcommunications, 2008.

[48] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings ACM International Conference on Multimedia, Berkeley*, pages 271–274, November 2003.

[49] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[50] 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.

[51] P. Moreno Nuno Vasconcelos, Gustavo Carneiro. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions Pattern Intelligence and Machine Analysis*, 29(3) :394–410, 2007.

[52] 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.

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

[54] Z. Pecenovic. *image retrieval using latent semantic indexing.* PhD thesis, Ecole polytechnique frale de Lausanne, 1997.

[55] Pottier. *Vers une sntique moderne.* P.U.F, 1964.

[56] Proc. 4th Int'l Joint conference on Pattern Recognition. *An analysis for scenes containing objects with substructures*, Kyoto, 1978.

[57] A. Puissant and C. Weber. Une démarche orienté-objets pour extraire des objets urbains sur des images thr. *Bulletin de la Société Francaise de Photogrammétrie et Télédétection*, 43(3) :993–1022, 2004.

[58] J. Rekers and A. Schurr. A parsing algorithm for context sensitive graph grammars. *Leiden Univ*, 1995.

[59] J. Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978.

[60] J. Rissanen. *Stochastic complexity in statistical inquiry.* World Scientific, 1989.

[61] Y. Rui, T.S. Huang, M. Ortega, and S. Machotra. Relevance feedback : a powerful tool for intereactive content-based image retrieval. *IEEE transactions on circuits and video technology*, 8(5), Sept 1998.

[62] R. Manmatha S. Feng and V. Lavrenko. Mutliple bernouilli relevance models for image and video annotation. *Proc IEEE CS Conference Computer Vision and Pattern Recognition*, 2004.

[63] E. Schapire. The boosting approach to machine learning : an overview. *MSRI Workshop on nonlinear Estimation and Classification*, 2002.

[64] C. Shannon. A mathematical theory of communication. *Bell Syst Technology*, 27 :379–423, 1948.

[65] Jonathon S.Hare, Paul H. Lewis, Peter G.B Enser, and Christine J. Sandom. Mind the gap : another look at the problem of the semantic gap in image retrieval. *Management and retrieval*, 6073, 2006.

[66] Mei-Ling Shyu Shu-Ching Chen Min Chen Chengcui Zhang Chi-Min Shu. Probabilistic semantic network-based image retrieval using mmm and relevance feedback. In *Multimedia Tools Application.* Springer-science, August 2006.

[67] E. Simoncelli and W.T. Freeman. Shiftable multi-scale transforms. *IEEE Transactions Information Theory*, 38(2) :587–607, 1992.

[68] A.W.M Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349–1380, 2000.

[69] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *CAIVD '98 : Proceedings 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.

[70] I. Tamba. *La sntique*. Presses universitaires de France, 2005.

[71] S. Todorovic and N. Ahuja. Extracting subimages of unknown category from a set of images. *CVPR*, 2006.

[72] R. Manmatha V. Lavrenko and J. Jeon. A model for learning the semantic of pictures. *Proc. Conf. Advances in Neural Information Processing Systems*, 2003.

[73] A. Vailaya, A. Jain, and H. J. Zhang. On image classification : City vs. landscape. In *CBAIVL '98 : Proceedings of the IEEE Workshop on Content - Based Access of Image and Video Libraries*, page 3, Washington, DC, USA, 1998. IEEE Computer Society.

[74] Nuno Vasconcelos and Gustavo Carneiro. Formulating semantic image annotation as a supervised learning problem. *CVPR*, 5 :163–168, 2005.

[75] W. Wang and I. Pollak. Hierarchical stochastic grammars for classification and segmentation. *IEEE transactions on image processing*, 15(10) :3033–3052, Oct 2006.

[76] C. Weber and A. Puissant. Une drche orientobjets pour extraire des objets urbains sur des images thr. *Soci Franse de Photogrammetrie et de Tdction*, 2004.

[77] Benjamin Yao, Xiong Yang, and Song-Chun 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.

[78] S.C. Zhu. Embedding gestalt laws in markov random fields. *IEEE Trans on PAMI*, 21, 1999.

[79] A.L. Yuille Z.W Tu, X.R Xhen and S.C Zhu. Image parsing : unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2) :113–140, 2005.