Automatic semantic network construction for multi-layer annotation of satellite images

Construction de réseaux sémantiques pour l'annotation multi-couches d'images satellitaires

Jean-Baptiste Bordes
Henri Maître

2008D011

Août 2008

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

Construction de réseaux sémantiques pour l'annotation multi-couches d'images satellitaires.

Jean-Baptiste Bordes and Henri Maître
Institut TELECOM/TELECOM ParisTech LTCI CNRS
46 rue Barrault 75013 PARIS

August 28, 2008

Résumé: Une méthode probabiliste pour annoter des images satellitaires 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 à la modélisation optimale de la base de données. Cette approche exploite ainsi la puissance de description des réseaux sémantique tout en conservant la flexibilité des approches statistiques par apprentissage. La méthode a été évaluée sur des bases de données SPOT5 et Quickbird.

Abstract: A novel method is presented for annotating satellite images. The labels used for annotation are given by a user with a set of example images. A learning step is then applied to learn the model. The originality of the method is to formulate the problem of semantic annotation to a further extent than a mere probabilistic classification task. The method takes into account the semantical relationships between the concepts by considering a *duality* between the structure of the model and the structure of the set of

labels. The semantical structure of the labels is represented by a semantic network containing three semantical relationships: synonymy, meronymy, and hyponymy. The semantic network is constrained in a hierarchy induced by the links of hyponymy and meronymy. By a procedure of MDL model selection, it is possible to find the optimal semantical structure of the set of labels.

1 Introduction

1.1 State of the art

The last two decades have seen the emergence of large image databases of various types. The spread of digital cameras and the increase of power and archiving ability of the computers have resulted in a growing need for the efficient handling of large databases of personal images. As for the field of remote sensing, a large variety of space-borne and air-borne sensors provide every day a huge quantity of information about the surface of the Earth, and this amount is getting even more enormous with the arriving generation of high-resolution satellite sensors.

Reliable image retrieval and indexing has thus become a major problem to efficiently access this information, driving an important amount of studies on the topic of content-based image retrieval. As a response to this demand, the early works made a direct use of low-level features extracted from the images [36, 45, 9, 13] and focused on the "query-by-example" request. Low-level descriptors computed from the documents are compared to those extracted in the user provided example images, and the images returned to the user are those with the least distance in the feature space. But it is now widely acknowledged that significative improvements in indexing systems require to build a bridge over the so-called *semantic gap* existing between the low-level features extracted in the images and semantic concepts [16].

In recent systems, image indexation and retrieval are based most of the time on the annotation of the database by a set of words describing the content of each image, enabling the user to specify the query through a natural language description of the concepts of interest. The earliest researches focused mainly on supervised learning schemes of specific semantics: differentiating indoor from outdoor scenes [47], photographs from paintings [6], bodies of humans and animals [11], cities from landscapes [48], natural scenes [30]. In these studies, a set of training images with and without the concept of interest was collected and a binary classifier trained to detect this concept. This classifier was then applied to the images of the whole database which were therefore annotated with respect to the presence or absence of the concept.

More recent researches tackle the problem differently by posing the problem of annotation as the inference of latent variables that encode the hidden semantic classes. They were encouraged by the results of topics extraction in textual data [17, 4, 41] and draw an analogy with retrieval in textual data by considering *visual words*. Visual words belong to a discrete collection obtained by quantization of low-level features and processed as the words of

a text [20]. In [24, 8], the annotation process is considered as translating the content from a visual language to a set of textual words and the learning process is identical to learning a lexicon from an aligned bitext. In [23], annotation is made through a cross-media relevance model, inspired by the relevance models introduced for cross-lingual retrieval [25]. In [32, 33], an exchangeability assumption is applied for the visual words and the pLSA model is applied for annotation. In [3], three probabilistical models of annotation are presented, inspired from various text models like LDA [4].

During training, a set of labels is assigned to each image, the image is segmented into a collection of regions (either using a regular block-based decomposition [34], or using a traditional segmentation method using directly the low-level features computed in the image [2]) and an unsupervised learning algorithm is run over the entire training set to estimate the joint density of semantic labels and visual features.

In these studies, the semantic labelling is thus seen as a problem of classification. Classes correspond to latent variables which have no semantic relationship between them. The semantic is introduced by the user during the annotation stage when providing a vocabulary for annotation and example images for each annotation label. Stochastic models for each model are built independantly for each label, and the semantic relationships among labels are not taken into account. Moreover, the labels are treated similarly for learning and annotation even if they are often very different in terms of semantic complexity and specificity and even if they can correspond to areas of various typical sizes.

In the KIM system [7], this hierarchy of the information is taken into account and modelized by a hierarchy distributed in five layers. Symbolic values, free of semantic, are infered by unsupervised learning at the four first layers. The semantic labels, introduced by the user, lie at the fifth layer and are linked to the symbolic values by the user interaction. The relation between layers is estimated using information-theoretic quantities. However, in this system, all the concepts lie on a single semantic layer and have no semantic relationship between them. There has been only few attempts to exploit the relationships between the semantic labels. Barnard and Forsyth, motivated by the statistical models proposed by Hoffman and Puzicha [18], adapted the hierarchical clustering for image annotation [2]. The hierarchy of models for generating words and image segments is derived from clustered images in the training set. Clusters capture contextual similarities while nodes capture generality of concepts. Words and blobs are then represented over the nodes of a hierarchy. Hierarchies induced by image clusters provide semantic interpretation for the models.

More generally, an efficient way to take into account the relationships

between labels and to obtain a high-level scene description is the use of semantic net [46]. Semantic networks contain nodes and links and are defined as directional acyclic graphs. They have been widely used for multimedia annotation. Naphade proposed the MultiNet as a way to represent high-level dependencies between concepts [35] and to take into account the mutual information between concepts. However, both classes and structure of the classification framework were decided by experts. Moreover the structure becomes very large when the number of classes increases. Benitez proposed a more general approach with MediaNet for video annotation in which the salient classes are automatically selected from annotated images and the relationships between concepts are discovered by using external knowledge resources from WordNet [1]. The relationships between concepts in MediaNet are divided in perceptual relationships (such as *looks similar to*) and semantic relationships (such as *meronymy/holonymy*).

For image interpretation, semantic networks describe stuctural relationships between the objects which are expected to be found in a scene and contain *prior* knowledge given by experts. For remote sensing images, literature provides several references of semantic networks application [27, 10, 42, 37, 22]. Devoted to the understanding of remote sensing data, the GeoAIDA system [21] represents explicitly the knowledge given by photointerpreters in a hierarchical semantic network. This network contains two different types of node: the generalization nodes (comparable to logic *xor*) and the compound nodes (comparable to logic *and*). Each node carries information about the type of area it represents and possesses attributes which are also a part of the knowledge (for example, a road junction consists of three to six interesecting roads). The interpretation strategy relies on fixed control rules. In [14], semantic concepts are structured by *and/or* graphs [52, 51] which are used for annotation of various types of images: outdoor scenes, faces, remote sensing data. An *and/or* graph is set in correspondence with the structure of a stochastic grammar. Given a test image, a parsing graph is formed by the production rules of the stochastic grammar and is defined as the image interpretation.

In such systems, adding a new concept requires an expert to add manually a new node in the structure of the semantic network and to build the edges linking this concept to others. However, some studies tried to solve the problem in greater generality and with higher flexibility by extracting these links automatically. In [38], when the user introduces a new concept, the concept is linked by some probability distributions to the low-level features and a distance is computed between this distribution and the other concepts distributions. If the distributions are too similar, the new concept is not added to the system. In [50], an algorithm is elaborated to learn the structure

of the semantic network based on Bayesian decision.

1.2 Paradigmatic Semantic Labelling

In this work, we show that it is possible to infer automatically the links between labels from their properties learnt during the training stage. Moreover, we show that this influence results in a hierarchy of labels well described by a semantic net. The nodes of the network contain the semantic labels and the arcs contain paradigmatic relationships between the labels ("meronymy/holonymy", "hyponymy/hyperonymy", "synonymy"). The structure of the network is constrained in a hierarchy naturally induced by the paradigmatic relationships of meronymy and hyponymy and decomposed in a finite set of layers. This *Paradigmatic Semantic Labelling* (PSL) makes possible to annotate the test images in several layers, corresponding to different levels of generality and complexity. It maintains an efficient correspondance between the structure of the semantic networks and the structure of a stochastic image model.

Each label of the network is associated with a probabilistic model used to compute the likelihood of an image annotated by this label. The models depend on each other according to the paradigmatic relationhips between the labels. At the training stage, an annotated database is provided by the user and the model optimaly fitting this database is estimated. Both parameters and structure are determined using the Minimization of Stochastic Complexity criterion [40], and the semantic network is thus deduced without ambiguity from this optimal model. We prove that these paradigmatic links can be infered with a training database consisting in a set of example images provided for each semantic label. A substantial amount of semantic information can thus be extracted from a weakly labelled training set. Given a new concept, the structure of the model is updated easily and adding new concepts increases the descriptive power of the database by the model (property of *extensivity*). Each model is fitted by taking into account the whole knowledge acquired by the system. We prove that this increases greatly the descriptive power of the images by the model.

As in [33, 3] and many other works previously cited, the PSL is based on modelisations inspired from text retrieval. The statistical models which will be introduced are based on visual words lying at the lowest level of the hierarchy. This layer contains a discrete collection of textons obtained from quantization of the low-level features extracted from the image. Moreover, the generation models which are used to estimate the likelihood of the textons are inspired from text retrieval methods. On top of this texton layer, three layers are devoted to express the semantic description. The hierarchy between

these upper layers derives naturally from the relationships of hyponymy and meronymy [29, 5].

From a theoretical point of view, we have to answer three questions:

- First, *which semantic concepts will be used to describe the image?*

  We have chosen the usual vocabulary of remote sensing, as familiar to any application expert: "city", "forest", "fields", etc.

- Second, *which relationships do we expect between nodes ?*

  We have chosen to use the paradigmatic relationships which link concepts in semantic languages. As the hyponymy relationship orders the words by their generality, the meaning of the words will be all the more general that they lie in the highest layer of the hierarchy. For instance, "forest" will lie in a lower layer when "vegetation" will lie in a higher layer. The meronymy relationship corresponds to a *global/part-of* hierarchy, and the words describe regions which are all the more structured that they lie in the highest layer of the hierarchy. For instance, "factory" will lie in a lower layer when "suburb" will lie in a higher layer.

- The third question is *How to derive automatically a correspondance between the structure of the model and a semantic network?*

  Each label of the semantic network is associated with a probabilistic model which is used to express the likelihood of the database of example images associated to the label. The concepts which lie in the lowest layer of the network are associated to a Naive Bayes model of the texton generation. If a label lies in a higher level of the hierarchy and is linked to a set of other words, the likelihood of the database is expressed with the models associated to these words, according to the nature of the link. By model selection, it is possible to determine the semantic network corresponding to the optimal model in terms of the database coding.

Following Smeulders, when attempting to bridge the semantic gap, we should also take into account the context [44]. In this work, since we only deal with remote sensing images, we make the simplifying assumption that all the images of the database correspond to the same context.

The paper is organized as follows: Section 2. details the global formalism used for the approach. Section 3. deals with the probabilistic modelling. Section 4. explains the coding of the model. Section 5. details the annotation of a test image using the model. Section 6. is devoted to experimental results.

2 Semantic representation

2.1 Semantic relationships

Semantic, as a field of linguistic, is based on the assumption that the words lie in their own space which is structured by the semantic relationships. It is now commonly acknowledged that the lexical units are structured by four different kinds of paradigmatic relationships [29]: the synonymy, the antonymy, the hyponymy/hyperonymy, and the meronymy/holonymy. We describe here briefly these relationships in their linguistical meaning:

The relationship of synonymy It is based on the possibility to exchange two lexical units in a minimal context while keeping a stable meaning.

The relationship of antonymy It determines a relationship of opposition between two terms. Like synonymy, it links lexical terms of the same grammatical category. However, at a semantic level, antonymy differs from synonymy by its binarity and the four kinds of dichotomical oppositions it corresponds to: contradictory ("inside"/"outside"), polar ("short"/"long"), inverse ("go up/ go down"), reciproc ("buy/sell").

The relationship of hyponymy/hyperonymy J. Lyons creates these terms to specify the vertical structuration of the lexical units [29]. The relationship of hyponymy corresponds to a specification: "cat" is hyponym of "animal". Reciprocally, the relationship of hyperonymy corresponds to a generalization. The logical implication which is linked to this relationship is: "if it is a cat, then it is an animal". Hyponymy is often refered to, in semantic networks, by a *kind-of* relationship.

The relationship of meronymy/holonymy The term of meronymy has been introduced by A. Cruse [5] to differentiate the "overall/part-of" hierarchical lexical relationship with the hyponymic relationship. Indeed, even if these two relationships share the properties of inclusion and asymmetry, their semantical meaning is different and their hierarchies are not compatible. Meronymy is often refered to, in semantic networks, by a *part-of* relationship.

For PSL, we choose not to take into account antonymy.

2.2 Considered links

2.2.1 Link of Synonymy

Two words are supposed to be synonyms if they describe the same type of region. Abusively, we will consider it here as an equivalence relationship as we suppose it here to be symmetric, reflexive and transitive.

2.2.2 *kind-of* link

The *kind-of* link corresponds to the semantical relationship of hyponymy. If $\{c_1, \ldots, c_k, c\}$ is a set of labels belonging to a vocabulary Ω, the semantical link *kind-of* KO is denoted as:

- $KO(c, \{c_1, \ldots, c_k\})$ means that "c_1, c_2,..., et c_k are kinds of c" (lexical semantic link of hyponymy), and conversely that c is more general than c_1, c_2,..., and c_k (lexical semantic link of hyperonymy).

The *kind-of* relationship is defined here in an *exclusive* meaning: $KO(c, \{c_1, ..., c_k\})$ means that a region is annotated by the label c *if and only if* it is annotated by the label c_1 or c_2 ... or c_k.

From this relationship, a set of semantic networks S_Ω^{ko} is defined as a set of semantic networks whose nodes are the elements of Ω submitted to the following constraints:

- The nodes of the semantic networks are assumed to be located in a hierarchy of a finite number of layers. The layer containing a label c is denoted $l(c)$. The set of concepts such as $l(c) = i$ is denoted $C_i(S_\Omega^{ko})$.
- The only semantic link existing between two links is the link *kind-of*
- A label c_a can be linked to a label c_b by the *kind-of* relationship only if $l(c_b) = l(c_a) + 1$.
- If c is a label such as $l(c) = i, i > 1, \exists\{c_1, \ldots, c_k\} \in C_i(S_\Omega^{ko}) \backslash KO(c, \{c_1, \ldots, c_k\})$

2.2.3 *part-of* link

The *part-of* link corresponds to the semantical relationship of meronymy/holonymy. If $\{c_1, \ldots, c_k, c\}$ is a set of labels belonging to a vocabulary Ω, the semantical link *part-of* PO is denoted:

- $PO(c, \{c_1, ..., c_k\})$ means that: "c_1, c_2,..., et c_k are *parts of* c", and conversely that c is *composed of* c_1, c_2,..., et c_k.

More precisely, $PO(c, \{c_1, ..., c_k\})$ means that, given a set of images, if a region R is annotated by the concept c, R admits a partition in one or several subregions which are annotated by concepts belonging to the set $\{c_1, ..., c_k\}$. It is important to note that the *part-of* relationship is defined here in a weak sense as all the concepts $\{c_1, ..., c_k\}$ do not have to be necessarily represented in the partition of R, and, contrary to the *kind-of* network used in [21], no constraint is imposed on the occurrence number of a same label inside region R.

Given a vocabulary of labels Ω, we define S_Ω^{po} as a set of semantic networks whose nodes are the elements of Ω submitted to the following constraints:

- The nodes of the semantic networks are supposed here to be located in a hierarchy of a finite number of layers. The layer a concept c lies in is noted $l(c)$. The set of concepts such as $l(c) = i$ is denoted $C_i(S_\Omega^{po})$.
- The only semantic link existing between two links is the link *part-of*
- A concept c_a can be linked to a concept c_b by the *part-of* relationship only if $l(c_b) = l(c_a) + 1$.
- If c is a label such as $l(c) = i, i > 1, \exists \{c_1, \ldots, c_k\} \in C_i(S_\Omega^{po}) \backslash PO(c, \{c_1, \ldots, c_k\})$

2.3 General structure of the global semantic network

To tackle the problem of semantic representation of the labels to an extent that incorporates both *kind-of* and *part-of* networks presented in Section 2.2, a structure of a semantic network integrating the links of hyponymy and meronymy is defined here. But as seen in Section 2.1, the hierarchies induced by meronymy and hyponymy are of completely different types. The solution proposed here to integrate the links *part-of* and *kind-of* within a single network is to consider a *kind-of* network which is superimposed on a *part-of* network (cf fig 1).

More formally, the global semantic network S_Ω whose nodes are the elements of Ω is supposed to contain two partial networks S_Ω^{ko} and S_Ω^{po}. The first one verifies the network structure of type *kind-of*, and the second one verifies the network structure of type *part-of* previously described. To simplify the notations, each one of these networks is supposed not to contain more than two layers. Moreover, the nodes of the first layer of S_Ω^{ko} are supposed to be the nodes of S_Ω^{po}:

$$C_1(S_\Omega^{ko}) = C_1(S_\Omega^{po}) \cup C_2(S_\Omega^{po})$$

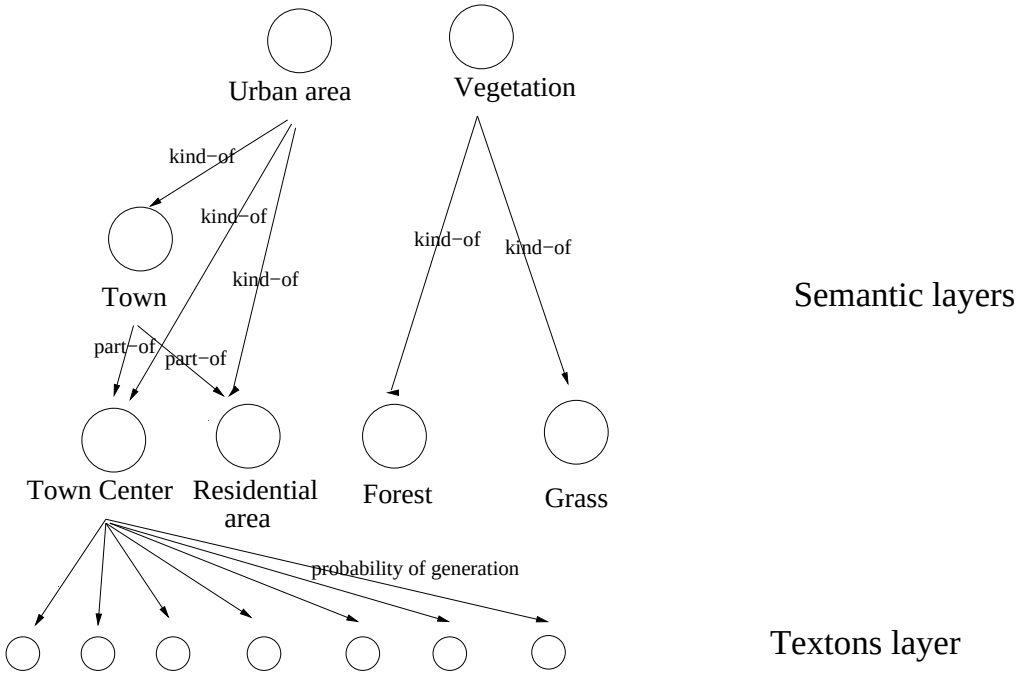


Figure 1: Illustration of the global semantic network

The choice to superimpose the *kind-of* structure on the two layers of the *part-of* structure is based on the fact that the words of the second layer of the *kind-of* relationship correspond to very general structures which can be of various complexity. For example, the word "urban" can correspond to small scale structures like "group of buildings", but can also correspond to a very large scale and complex areas which could be labelled by "suburb".

3 Probabilistic modelisation

3.1 Formalism

Let $\Omega = \{c_1, ..., c_n\}$ be the codebook of labels used for annotation. Each label c_i is assumed to be linked to a set of example images X_i provided by the user. The whole dataset is noted $X = \{X_1, ..., X_n\}$, where X_{ij} stands for the jth learning image provided for label i. A semantic network S_Ω whose nodes are the labels $\{c_1, ..., c_n\}$ and the corresponding probabilistic model M_Ω used to express the dataset likelihood $P(X|M_\Omega)$ are defined. The set of networks the nodes of which are the labels $\{c_1, ..., c_n\}$ and verifying the requirements defined in section 2.2 is denoted $\mathcal{S}_\Omega$, and the set of possible models to describe X is denoted $\mathcal{M}_\Omega$.

The global model M_Ω is decomposed in a set of models $\{M_1, ..., M_n\}$, M_i modelling the database X_i of the label c_i. Each model M_i contains a set of parameters θ_i fitted on the database X_i, and a location in the structure of M_Ω. In Section 2.3, a surjective function defined from $\mathcal{M}_\Omega$ to $\mathcal{S}_\Omega$ associating each model M_Ω to a semantic network S_Ω will be specified.

3.2 Different layers of the model

The main idea of the association of a semantic network S_Ω to an image model M_Ω is to link each layer to a specific kind of probabilistic modelisation. For the sake of clarity, the number of layers of each partial network $C_2(S_\Omega^{po})$ and $C_2(S_\Omega^{ko})$ is constrained to be less than 2. The training sets X_i are supposed to be generated independently conditionnaly to the image model M. The likelihood of the training database X can be written as:

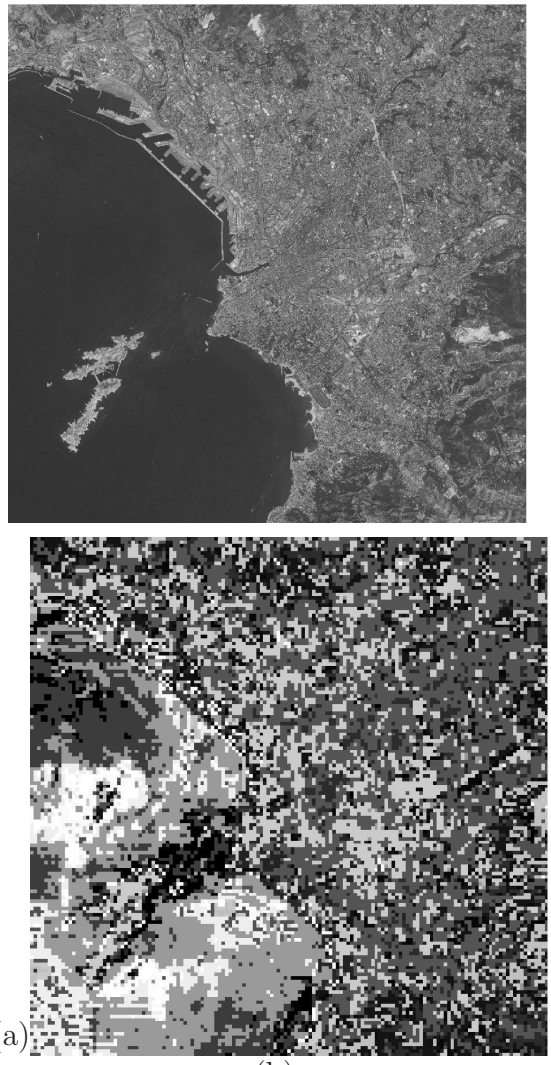
$$P(X_1, X_2, ..., X_n | M_\Omega) = \prod_{i=1}^{n} P(X_i | M_\Omega) \tag{1}$$

Each label c_i in S_Ω is associated to a model M_i used for description of the learning set X_i.

- If $c_i \in C_1(S_\Omega^{po})$, M_i is a naïve Bayes model over the low-level features of the image.
- If $c_i \in C_2(S_\Omega^{po})$, M_i is a model which has some similarity with LDA. The image is decomposed in regions which are then described by the models associated to the labels of $C_1(S_\Omega^{po})$.
- If $c_i \in C_2(S_\Omega^{ko})$, M_i is defined as a mixture of unigrams over the models associated to the concepts of $C_1(S_\Omega^{ko})$.

3.2.1 Layer 0: Low-level description of the images

The method detailed in this work takes place after a first stage of image processing where features have been extracted on a regular grid in the images of the database. A clustering algorithm is then applied on these features and a codebook of size n_0 is computed. Given an image of the database, the features previously extracted from it are quantized using the computed codebook to represent this image with a discrete collection of textons. A map, the pixels of which are the index of the textons, is thus obtained (cf image 2). This reduced image is used as the input of the modelling and the original image will no longer be used in the following lines.



(a) (b)

Figure 2: (a) SPOT5 image of size $6000 * 6000$ at 2,5m of resolution @CNES (b) image of size 150×150 whose pixels are the index of the textons. Each texton is computed on a 40×40 window. The codebook of textons contains 90 labels.

3.2.2 Layer 1

Given a map I annotated by the concept a belonging to $C_1(S_\Omega^{ko})$, its likelihood is expressed by a naive Bayes model [28] and the number of textons in the map is coded by a Poisson's law. The likelihood of the map I conditionally to M_a is thus written as:

$$P(I|M_a) = Poiss_{\lambda_a}(\sum_{j=1}^{n_0} x_j) \prod_{j=1}^{n_0} (\theta_{aj})^{x_j} \tag{2}$$

where θ_{aj} stands for the probability for a texton of value j to be generated with model M_a, and x_j is the occurrence of texton j in map I. As this expression only depends on M_a:

$$P(I|M) = P(I|M_a) \tag{3}$$

3.2.3 Layer 2 of the *part-of* network

Given a *part-of* network S_Ω^{po}, let a be a label in $C_2(S_\Omega^{po})$ linked to a set of labels $\{c_1, ..., c_k\}$ lying in $C_1(S_\Omega^{po})$. If I is a map annotated by a, the generative model is detailed as:

- a number m_{ai} is sampled with probability $Poiss_{\Lambda_a}$.
- a partition of the map $P = \{R_1, R_2, ..., R_{m_{ai}}\}$ is sampled with uniform probability.
- for j from 1 to m_{ai}, a label $c(R_j)$ is sampled among $\{1, ..., k\}$ with probability $\{\pi_{a1}, ..., \pi_{ak}\}$ and the likelihood of the region is computed conditionally to the label c: $P(R_j|c(R_j))$.

The likelihood of the map is thus written as:

$$\begin{aligned} P(I, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(I|M_a, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}) \\ P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) \end{aligned} \tag{4}$$

The first term of this product is expressed as:

$$\begin{aligned} P(I|M_a, \{R_1, R_2, ..., R_{m_{ai}}\}, \{c(R_1), c(R_2), .., c(R_m)\}) = \\ Poiss_{\Lambda_a}(m_{ai}) \prod_{j=1}^{m_{ai}} P(R_j|M_{c(R_j)}) \end{aligned} \tag{5}$$

where the term $P(R_j|M_{c(R_j)})$ can be written using eq 2, and $c(R_j) \in C_1(S^{po}_{\Omega_{po}})$.

Assuming independancy between the annotations and the partition of the image conditionally to the model M_a, the second term of eq 4 can be written as:

$$P(\{R_1, R_2, ..., R_m\}, \{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \\ P(\{R_1, R_2, ..., R_m\}|M_a)P(\{c(R_1), c(R_2), .., c(R_m)\}|M_a) \quad (6)$$

The labels are also assumed independant conditionally to the model M_a: $P(\{c(R_1), c(R_2), .., c(R_m)\}|M_a) = \prod_{j=1}^{m} P(c(R_j)|M_a)$.

$P(c(R_i) = c_j)$ is denoted π_{aj} for every $j \in \{1, ..., k\}$. π_{aj} and Λ_a are parameters of the model a. A uniform law is assumed on the set of the map partitions. $P(\{R_1, R_2, ..., R_m\}|M_a) = \frac{1}{K}$, where K is the number of possible partitions in the image with 4-connex regions. This number depends on the image, it is untractable for most images.

3.2.4 Layer 2 of the *kind-of* network

Given a vocabulary of labels $\Omega_{ko} \subset \Omega$ and a *kind-of* network $S^{ko}_{\Omega_{ko}}$, let a be a concept in $C_2(S^{ko}_{\Omega_{ko}})$ linked to a set of concepts $\{c_1, ..., c_k\} \in C_1(S^{ko}_{\Omega_{ko}})$ The concept a is associated with a latent variable L_a taking its value in a finite vocabulary of size k: $\{1, ..., k\}$. The process of semantic specification from concept a to concept c_j is modelled by the assignment of latent variable L_a to value j.

Given an image I annotated by a:

$$P(I, L_a = j|M) = P(L_a = j)P(I|L_a = j, M) \quad (7)$$

Conditionnaly to the fact that the concept a has been specified to concept c_j, the probability of the image is computed with the submodel M_j:

$$P(x, L_a = j|c(I) = a) = P(L_a = j)P(I|M_j)$$

Thus, the probability is computed as a mixture model:

$$P(x|c(I) = a) = \sum_{j=1}^{k} P(L_a = j)P(I|M_j)$$

The set of values $\{P(L_a = j)\}_{j=1}^{k}$ are the parameters of the concept a and are noted: $\pi_{aj} = P(L_a = j)$. The fact that each concept a in $C_2(S^{ko}_{\Omega_{ko}})$ is linked to at least one concept of $C_1(S^{ko}_{\Omega_{ko}})$ guarantees that $\sum_{j=1}^{k} \pi_{aj} = 1$.

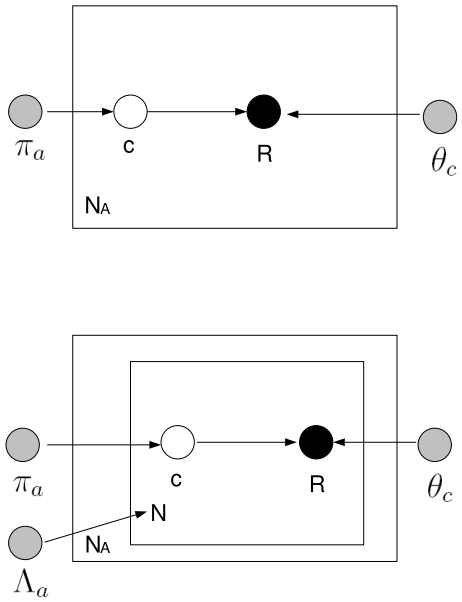


Figure 3: Representation of the models *kind-of* (upper graph) and *part-of* (lower graph). A box corresponds to the iterative and independant sampling of random variables which are inside the box. The number of sampling is wirtten at the bottom-left of the box. A colored disc corresponds to an observable random variable. A non-colored disc corresponds to a non observable random variable.

$$P_a(x) = \sum_{j=1}^{k} \pi_j P(I|M_j) \tag{8}$$

3.3 Synonymy relationship

Let $\Omega = \{c_1, ..., c_n\}$ be a vocabulary and S_Ω a semantic network the nodes of which are the elements of Ω. Let M_Ω be a stochastic model which is set in relationship with S_Ω. Let c_i and c_j be two labels, if they are synonyms in S_Ω, they are linked to the same model in M_Ω. Thus, if M_i is a model, $c(M_i)$ is not a singleton but a set of words which are all synonyms.

3.4 Correspondance between a global stochastic model and a semantic network

Given a model $M = \{M_1, \ldots, M_k\}$, the associated stochastic model S is built in two steps:

- The labels are divided into the different sets $C_1(S^{po})$, $C_2(S^{po})$ and $C_2(S^{ko})$ according to the model they correspond to.
- For all the labels c set in $C_2(S^{po})$ and $C_2(S^{ko})$, the semantic relationship *part-of* (if $c \in C_2(S^{po})$ or *kind-of* (if $c \in C_2(S^{ko})$ with the label corresponding to index i is created if $\pi_{ci} > 0$.

3.5 Extensivity

Let M be a model the parameters and structure of which have been optimized by a maximum of likelihood of the training set $X = \{X1, ..., X_n\}$ among the set $\mathcal{M}$ of possible models:

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

If a new concept c_{n+1} is added to the initial set of concepts $c_1, ..., c_n$, and X' is the training set composed of X and a training set X_{n+1} associated to the concept c_{n+1}.

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

Let M' be the model estimated by maximum of likelihood on the dataset X' among the set of models **M'**:

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

The following property holds:

$$P(X|M') \geq P(X|M) \tag{9}$$

Proof: Let $i \in \{1, ..., n\}$, $\mathcal{M}_i$ et $\mathcal{M}'_i$ two sets of models fitting the dataset X_i in each case. Models M_i and M'_i can be located in the first or the second layer of the network:

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

The models of layer 1 are defined by direct modelisation on the textons of the image and their expression are not related to other models of the network (cf equation 3). The following property can thus be deduced:

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

As long as the second layer is concerned:

$$\{C2(\mathcal{M'}_i)/\pi_{i,n+1} = 0\} = C2(\mathcal{M}_i) \tag{10}$$

the following inclusion relationship can be infered:

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

Thus

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

The following property can be deduced:

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

This property means that, by adding a new concept to the vocabulary, the likelihood of the rest of the database can only increase. Thus, this method takes advantage of all the knowledge given by the user to improve the description of the database. By analogy with Statistic Physics, this property is called 'extensivity property" [26] as a physical state variable describing a system is said to be "extensive" if this variable grows with the size of the system.

4 Coding of the model

The principle of minimization of the stochastic complexity has been introduced by Rissanen in 1978 [40]. The notion of stochastic complexity substitutes to the complexity of Kolmogorov [12] the number of bits necessary to code a sequence with an entropic code, and also the number of bits which are necessary to code the probabilistic model.

4.1 Coding of the system

The code length is traditionnaly made in two parts ([39]):

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

the first term corresponding to the coding of the model, the second term corresponding to the length of the code which is necessary to code the data using the model. The set of example images are supposed to be independant for each label. The description length can thus be summed on the labels:

$$C(X, M) = \sum_{c=1}^{n} [C(X_c|M) + C(M_c)] \tag{11}$$

4.1.1 Layer 1

The layer 1 models are assumed to generate directly the textons of example images. Thus, $C(X_c|M) = C(X_c|M_c)$ and if X_{cj} is the j-th example image associated to the concept c, by using Shannon formula [43], the term $CS(X_c|M_c)$ is written as:

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

the images X_{cj} of the database provided for concept c are supposed to be independant. The stochastic complexity of the database can be written as:

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

by defining $card(X_{cj}) = \sum_{j=1}^{n_0} x_{cj}$ as the total number of textons in X_{cj}, and by introducing the expression 2 in the last equation:

$$CS(X_c|M_c) = \lambda_c - card(X_{cj}) \log \lambda_c + \sum_{j=1}^{card(X_{cj})s} \log j - \sum_{j=1}^{n_0} x_{cj} \log(\theta_c) \quad (12)$$

To code the model M_c, the layer it belongs to has to be coded first. Here, a model which can have a maximum of two layers is presented, the index of the layer can be coded by a single bit with value 0 or 1. Then, the generation parameters of textons θ_c and the size parameter Λ_c have to be coded. Rissanen formula [39] linking a vector of parameters of size T estimated with N_{ech} samples provides the code length:

$$\frac{T}{2} \log N_{ech} \quad (13)$$

Vector θ_c has size n_0. The number of samples it is estimated with is equal to the total number of textons of the database X_c. The vector Λ_c has size 1 and is estimated with a number of samples equal to the total number of images in the database X_c, defined as N_c.

$$CS(M_c) = \frac{n_0}{2} \log(\sum_{j=1}^{N_c} \sum_{k=1}^{n_0} x_{cjk}) + \frac{1}{2} \log N_1 \quad (14)$$

The stochastic complexity $CS(X_c|M_c)$ is therefore:

$$CS(X_c|M_c) = \lambda_c - card(X_{cj}) \log \lambda_c + \sum_{j=1}^{card(X_{cj})s} \log(j) - \sum_{j=1}^{n_0} x_j \log(\theta_c) + \frac{n_0}{2} \log(\sum_{j}^{N_c} \sum_{k}^{n_0} x_{cjk}) + \frac{1}{2} \log N_1 + 1 \quad (15)$$

4.1.2 Layer 2 of the 'Kind-of" network

Let c stand for a label lying in $C_2(S_\Omega^{ko})$, the term $C(X_c|M)$ can be written as:

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

Applying the independence assumption of the images of the database X_c and by introducing the expression of $P(X_c|M_c)$ written in eq 8, the following expression can be written as:

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

For the coding of the model, it is necessary to code the generation parameters π_c. Using expression 13:

$$CS(M_i) = \frac{n_1}{2} \log N_c \quad (16)$$

4.1.3 Layer 2 of the *part-of* network

Let c stands for a label lying in $C_2(S_\Omega^{po})$, and i for the index of a map in the database X_c and $P_i = \{R_{i1}, R_{i2}, ..., R_{im_{ci}}\}$ stand for an annotated partition of this image, the term $C(X_{ci}|M, P_i)$ is written:

$$C(X_{ci}|M, P_i) = -\log P(X_{ci}|M, P_i)$$

This probability being expressed as a sum on the joint probability of the image and the labels on all possible annotations of the image given a partition, becomes:

$$P(X_{ci}|M, P_i) = \sum_{\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}} P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i) \quad (17)$$

The term $P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i)$ is decomposed as:

$$P(X_{ci}, \{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M, P_i) = P(X_{ci}|\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}, M, P_i) P(\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}|M) \quad (18)$$

The first term of this product is expressed by equation 5, and the second by equation 6.

$$P(X_{ci}|M, P_i) = \sum_{\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (19)$$

However, this expression may become untractable for a high number of regions. Some approximations can be used to find a lower bound of this expression. In this work, we just use the following very coarse bound:

$$P(X_{ci}|M, P_i) = \max_{\{c(R_1), c(R_2), ..., c(R_{m_{ci}})\}} Poiss_{\Lambda_c}(m_{ci}) \prod_{j=1}^{m_{ai}} \pi_{c(R_j)} P(x(R_j)|c(R_j)) \quad (20)$$

For each model, the index of the layer and the generation parameters of the labels of $C_1(S_\Omega^{po})$ π_c have to be coded.

$$CS(M_i) = \frac{n_1}{2} \log N_c \quad (21)$$

The point is, as for the concepts of layer 1, to code the values of the textons with a minimal code length. Thus, the partition P_i for each image of index i is not coded and the code length $C(X, M)$ is expressed as:

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

4.2 Optimization procedure

Given a database X, we wish to find the model M minimizing the stochastic complexity $C(X, M)$ among the set of possible models in terms of parameters and structure. Given a vocabulary Ω of cardinal n, the labels are separated in three subsets: $C_1(S_\Omega^{po})$, $C_2(S_\Omega^{po})$ and $C_2(S_\Omega^{ko})$. The number of possible structures thus corresponds to the total number of possible repartitions of

the n labels in these three subsets, that is to say 3^n. As a global exploration is untractable for high values of n, we propose here a greedy algorithm leading to a local minimum of the stochastic complexity $CS(X, M)$.

Initialization The initialization configuration of the algorithm is the one containing all the n concepts in layer 1.

Iterations At each step, and for each label c_j of layer 1, the stochastic complexity associated with the configuration where the label c_j is located in layer 2 is computed. The model parameters of the first layer are first estimated by likelihood maximisation.

For each label c of layer 1, the parameters are estimated using the following equations:

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

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

$card(X_{cj})$ stands for the number of textons in the image X_{cj}, $card(X_c) = \sum_{j=1}^{N_c} card(X_{cj})$. $occ_{X_c}(L_c = j)$ stands for the number of occurrences of the texton of value j in the database X_c.

For every concept c of layer 2, we have the following expressions:

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

Once the parameters are estimated, the global stochastic complexity is computed using equation 11. The model minimizing the stochastic complexity is set in layer 1 if the corresponding complexity is less than the complexity obtained at last step.

Stopping the algorithm The algorithm stops when the stochastic complexity algorithm increases. The last model is taken as the optimal model.

Discussion To put a label c from layer $C_1(S_\Omega^{po})$ to layer $C_2(S_\Omega^{po})$ or $C_2(S_\Omega^{ko})$ has two main impacts: an impact on the complexity term $C(X_c|M_c)$ as the likelihood of X_c will be expressed by a different model, and an impact on the complexity terms corresponding to the labels belonging to layers $C_2(S_\Omega^{po})$ or $C_2(S_\Omega^{ko})$. This latter impact results in an increase of the stochastic complexity

due to the decrease of the number of possible modelisations according to the models belonging to $C_1(S_\Omega^{po})$.

5 Annotation of a new image

Once the structure of the model has been learnt and the parameters have been estimated, test images can be annotated using the model. The step of annotation extracts semantic information from the image by annotating it with the vocabulary Ω provided by the user. Our method is focused here on large databases of remote sensing images. As these images are very large, annotate a whole image to semantic labels seems not relevant, and we present a method which provides annotated regions of the test image.

5.1 Annotation method

The first step of the annotation stage is to decompose the test image I in regions annotated by concepts existing in the model. This process is modelled as a *part-of* network where the third layer is constituted of a virtual label c_{scene} annotating the test image. In the stochastic model corresponding to this semantic network where c_{scene} is added, the generation probability of each label $c \in C_2(S_\Omega^{po})$ is unknown. Thus, we consider a uniform distribution on the elements of $C_2(M)$:

$\forall c \in C_2(M), P_{c_{scene}}(c) = \frac{1}{|C_2(M)|}$

The test image is represented as a region R annotated by the virtual concept c_{scene}. This region is decomposed as an annotated partition $G_2 = \{R_1^2, R_2^2, ..., R_{N_2}^2\}$ where each concept of annotation $c(R_i^2) \in C_2(S_\Omega^{po})$. Then, each region R_i^2 is decomposed in a partition of regions annotated by concepts of $C_1(S_\Omega^{po})$. The result is an annotated partition $G_1 = \{R_1^1, R_2^1, ..., R_{N_1}^1\}$.

Given an image I, the annotations G_1 et G_2 are taken as maximizing the probability $P(G_1, G_2|I)$:

$$\max_{G_1,G_2} P(G_1, G_2|I) \tag{22}$$

Let us write:

$$P(G_1, G_2|I) = \frac{P(I, G_1, G_2)}{P(I)}$$

As the maximisation does not depend on I, we just operate the maximization of the joint probability $P(G_1, G_2, I)$:

$P(G_1, G_2, I) = P(G_2)P(G_1|G_2)P(I|G_1)$

5.2 Optimization algorithm

As the space of joint configurations of G_1 and G_2 is too large, the complexity is reduced by applying two steps of inference. The maximum $G_{1,opt}$ of $P(I|G_1)$ is determined in the space of all the possible annotated partitions $\mathcal{G}_1^I$. Then, given the partition $G_{1,opt}$, the maximum $G_{2,opt}$ of $P(G_{1,opt}|G_2)P(G_2)$ is determined in the space of all the possible annotated partitions $\mathcal{G}_2^{G_{1,opt}}$ where each region is a union of regions of $G_{1,opt}$.

For each inference step, the optimal annotated partition has to be found among a huge set of configurations. A path is thus explored through the space of annotations starting from a complex partition and finishing to a single region containing the whole image. Initialization is performed by annotating each texton with a label considering its value and the value of its neighbours on a small window.

5.2.1 First inference step

- Initialisation of the algorithm:

  A first partition is created using the estimated models M_c associated with the concepts $c \in C_1(S_\Omega^{po})$ in the following way: at each texton of the image of coordinates (k, l), the following histogram vector of size n_0 is computed:

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

  where E_i stands for the i-th base vector, $g_{m_1,m_2,\sigma}(x,y)$ stands for the 2D Gaussian function of mean m_1, m_2 and of variance σ^2. σ is a parameter of the algorithm. This vector is a representation of the neighbourhood of the texton (k, l).

  Then, for $i \in \{1, \ldots, n_1\}$, the following probability functions is computed:

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

  and texton (k, l) is annotated with label c verifying:

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

An annotated partition G_1^0 is then created by building a region annotated with concept c for each 4-connex area of textons which has been linked to the concept c during the previous step.

- Let i be the number of iterations done in the loop. While the number of regions in the image contained in G_1^i is more than 1:
  - for all possible merging of adjacent regions:
    - we consider the n_1 possible annotations for the new region. If two regions are adjacent and are annotated with the same label, they are merged. For these n_1 configurations G_1, the likelihood $P(I|G_1)$ is computed.
  - the configuration maximizing the likelihood is kept and denoted G_1^i
- The final annotated partition G_1^{opt} is the configuration verifying:

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

At each step of the loop, at least two regions are merged. As a consequence, the algorithm finishes in less iterations than the number of regions in the initial partition G_1^0. The higher σ, the fewer regions are in G_1^0.

The process of the second step of the algorithm is similar to the first step. A path is explored in the space of the annotated partitions by creating an initial partition G_2^0 and by merging regions iteratively until just one region remains in the image.

5.2.2 Second step of inference

- Initialisation of the algorithm:

  A first partition is created using the estimated models M_c associated with the concepts $c \in C_2(S_\Omega^{po})$ in the following way: for each region R_k of G_1^0, the following histogram vector of size n_1 is computed:

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

  Then, for $i \in \{1, \ldots, n_2\}$, the following probability functions are computed:

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

and region R_k is annotated by the label c verifying

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

An annotated partition G_2^0 is then created by building a region annotated by the concept c for each 4-connex area of textons which have been linked to the concept c during the previous step.

- Let i be the number of iterations done in the loop. While the number of regions in the image contained in G_1^i is over 1:
  - for all possible merging of adjacent regions
    - the n_2 possible annotations for the new region are considered. If two regions are adjacent and are annotated with the same label, they are merged. For these n_2 configurations G_2, the likelihood $P(G_1^{opt}|G_2)$ is computed.
  - the configuration maximizing the likelihood is kept and noted G_2^i
- The final annotated partition G_2^{opt} is the configuration verifying:

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

5.3 Semantic representation of the image

The previously detailed algorithm output consists in two annotated partitions of the test image. the fusion of these two partitions in a single set of regions is denoted $P_{po} = \{G_1, G_2\}$. This is the minimal interpretation of the image, to get it richer:

- For every region R belonging to P_{po}, of the word $c(R)$ is linked by an hyponymic relationship to the word c' belonging to $C2(S_{ko})$, we create a new region the localisation of which is the same as R and th annotation of which is c'.
- For every pair of regions R and R' belonging to P_{ko} and annotated by the same concept, if R and R' are adjacent or if their intersection is not empty, these regions are merged.

- For every region R belonging to the set $P = P_{ko} \cup P_{po}$, all the synonyms of $c(R)$ are added to the set of annotations of R

Each region has an outer boundary. Regions with holes also have inner boundaries to represent the holes. Fuzzy modeling is used for pairwise relationships between regions to represent the following high-level user concepts:

Perimeter-class relationship

- *disjoined*: Regions are not bordering each other
- *bordering*: Regions are bordering each other
- *invaded by*: Smaller region is surrounded by the larger one at around 50% of the perimeter
- *surrounded by*: Smaller region is almost completly surrounded by the larger one

Surface-class relationship

- *overlapping*: Regions are overlapping each other
- *contained by*: Smaller region is almost completly surrounded by the larger one

Distance-class relationship

- *Near*: Regions are close to each other
- *Far* : Regions are far from each other

Given two regions R_i et R_j, to find the relationship between a pair of regions R_i and R_j, the following quantities are first computed:

- Perimeter of the first region π_i,
- Perimeter of the second region π_j,
- Common perimeter between the two regions π_{ij},
- Ratio of the common perimeter to the perimeter of the first region: $r_{ij}^1 = \pi_{ij}/\pi_i$,
- Surface of the first region σ_i,

- Surface of the second region σ_j,
- Common surface between the two regions σ_{ij},
- Ratio of the common surface to the surface of the first region: $r_{ij}^2 = \sigma_{ij}/\sigma_i$,
- Closest distance between the boundary of the first region and the boundary of the second region d_{ij}.

Then, each pair is assigned a degree of their spatial relationship using the fuzzy class membership functions given in figure 4. Then, these pairwise relationships are combined using an attributed relational graph [15]. The attributed relational graph represents regions by the graph nodes and their spatial relationship by the edges between such nodes. Nodes are labeled by concepts and the corresponding confidence values (likelihood) for the label assignment. Edges are labelled with the spatial relationship classes (pairwise relationship names) and the corresponding degrees (fuzzy membership values) for these relationships.

6 Experimentations

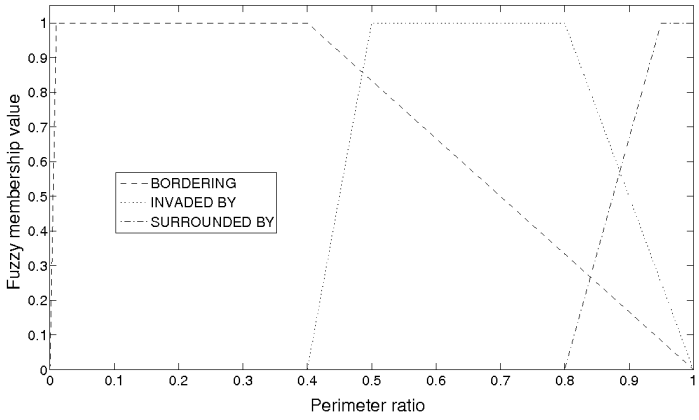
6.1 Construction of semantic networks

Two different sets of data will be used: synthetic data will confirm case by case the adequacy of the PSL algorithm and real data will demonstrate its robustness.

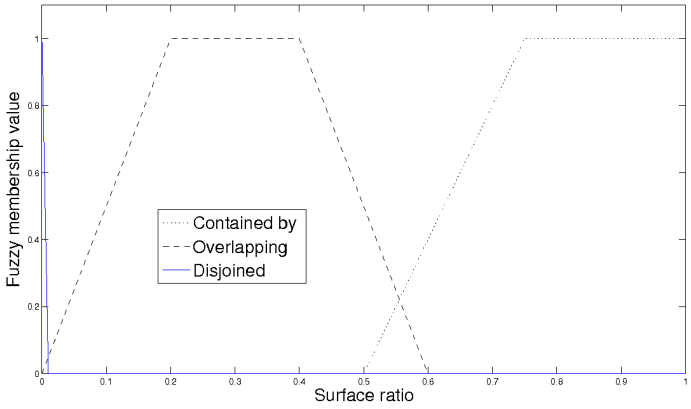
6.1.1 Synthetic data

Synonymy relationship A Gaussian probability function on gray values of range from 1 to 256 is defined. Two databases X_1 and X_2 associated with two concepts c_1 and c_2 and containing N images of size 200×200 are generated pixel by pixel using this distribution. These images are supposed to correspond to the textons map mentioned in Section 3.2.1. Two models M and M' are estimated considering two different cases:

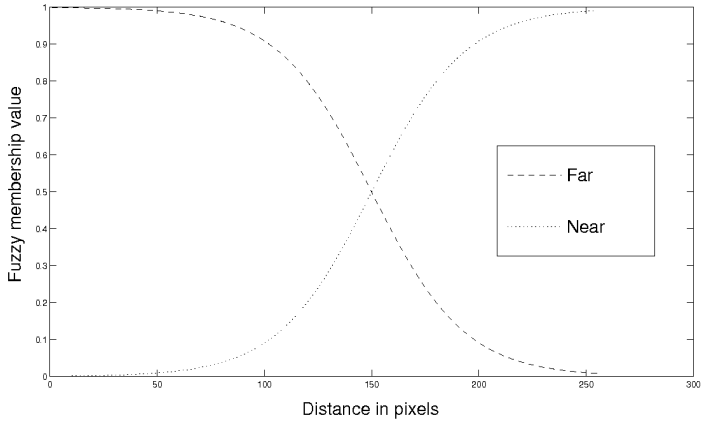
- model M corresponds to the case where c_1 and c_2 are supposed to be not synonyms. Two vectors of parameters θ_1 and θ_2 of dimension 256 were estimated over X_1 and X_2 respectively.



(a)



(b)



(c)

Figure 4: Fuzzy spatial relationships (a) Perimeter-related relationships (b) Surface-related relationships (c) Distance-related relationships

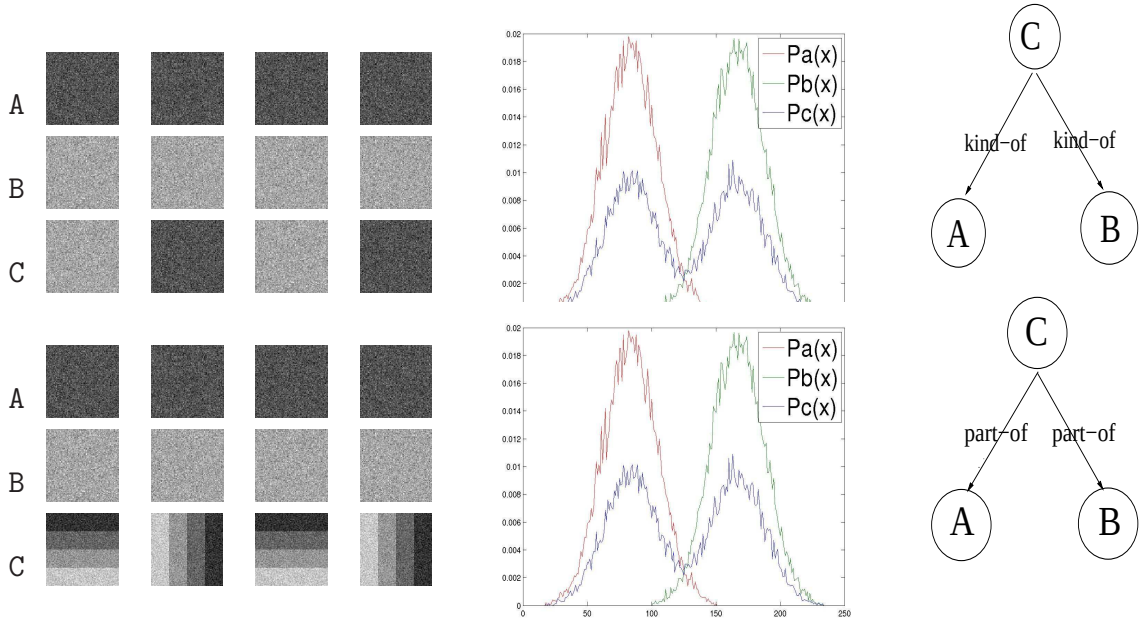


Figure 5:

- model M' corresponds to the case where c_1 and c_2 are supposed to be synonyms. A single vector θ' of dimension 256 is estimated over $X_1 \cup X_2$.

The stochastic complexity was computed in each case:

$$C(X, M) = -\log P(X_1|M_1) - \log P(X_2|M_2) + C(M_1) + C(M_2)$$

$$C(X, M') = -\log P(X_1|M') + C(M')$$

The ratio $R_{syn} = \frac{C(X,M}{C(X,M')}$ obtained according to the size of the database is shown on figure 6. As expected, R_{syn} converges to 1 when the size of the database increases. Indeed, $C(X, M)$ contains the coding of two vectors of parameters as $C(X, M')$ contains one. As the databases X_1 and X_2 are generated by a single distribution, the vectors θ_1 and θ_2 converge to θ asymptotically when the size of the database increases. Though, the terms $C(\theta_1)$, $C(\theta_2)$, $C(\theta')$ have a logarithmic dependence in N but the terms $C(X_1|\theta_1)$, $C(X_2|\theta_2)$ evolve in a linear way.

Hyponymy relationship k Gaussian distributions g_i of mean $m_i = \frac{i}{256}, i \in \{1, \ldots, 256\}$ and of same variance σ^2 are considered. k training sets X_i associated with each label c_i are generated pixel by pixel from distribution g_i and contain N images of size 200×200. These images are supposed to correspond

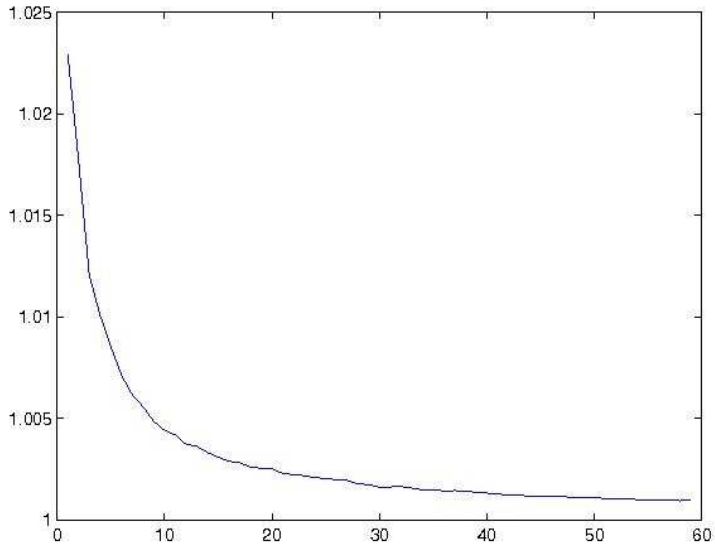


Figure 6: Ratio R_{syn} versus the number N of images in X_1 et X_2. R_{syn} converges to 1 when N increases.

to the textons map mentioned in Section 3.2.1. Then, a training set X_{k+1} associated to a label c_{k+1} is generated in the following way:

- For every $i \in \{1, \ldots, N\}$
  - a number j is sampled with uniform probability in $\{1, \ldots, 256\}$.
  - the image $X_{(k+1)i}$ is generated by sampling independantly each pixel with probability g_i.

Two models M and M' are estimated for the two cases corresponding respectively to two different structures:

- $\{c1,..., c_k\}$ and c_{k+1} are all located on layer 1. The model M_{k+1} is estimated by maximum of likelihood on X_{k+1} as the other models.
- $\{c1,..., c_k\}$ are supposed to be hyponyms of c_{k+1}, and c_{k+1} is linked with $\{c1,..., c_k\}$ by *kind-of* links. The model M_{k+1} is thus built as a mixture of models over the distributions P_{c_j} and the vector of parameters λ_{k+1} is estimated on X_{k+1}.

Figure 7 shows the ratio $R_{hyp} = \frac{C(X,M}{C(X,M')}$ according to the variance σ of the Gaussian distributions. As expected, R_{hyp} converges to 1. Indeed, increasing

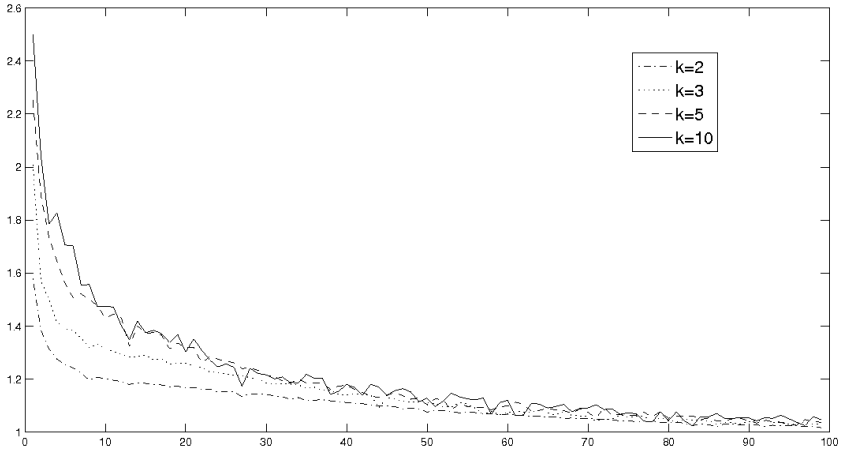


Figure 7: Ratio R_{hyp} versus σ (the higher σ, the less discriminative the features) for different values of k for the hyponymy relationship. R_{hyp} is always greater than 1, and decreases for increasing σ

the variance of the Gaussians amounts to saying that the features are less discriminative. This experience shows that the *kind-of* link between labels provides all the more reduction of stochastic complexity that the features describe efficiently the data.

Meronymy relationship k Gaussian distributions g_i of mean $m_i = \frac{i}{256}, i \in \{1, \ldots, 256\}$ and of same variance σ^2 are considered. k training sets X_i each associated with label c_i are generated from distribution g_i and contain N images of size 200×200. These images are supposed to correspond to the textons map mentioned in Section 3.2.1. Then, a database X_{k+1} associated to a label c_{k+1} is generated in the following way:

- for every $i \in \{1, \ldots, N\}$, the image X_i, of size $200 \times (200 * k)$ is separated in k regions $R_1(X_i), ..., R_k(X_i)$ of size 200×200
  - for every $i \in \{1, \ldots, k\}$
  - each region $R_j(X_i)$ is generated by sampling independantly the textons with distributions g_j.
- $\{c1, ..., c_k\}$ and c_{k+1} are all located on layer 1. The model M_{k+1} is estimated by maximum of likelihood on X_{k+1} as the other models.

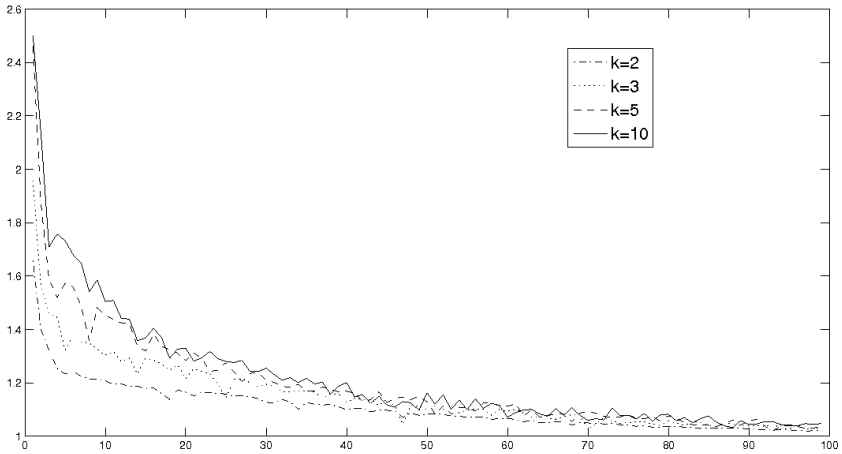


Figure 8: Ratio R_{mer} versus σ for different values of k for the meronymy relationship. Similar conclusions can be drawn than in fig 7.

- $\{c1,...,c_k\}$ are supposed to be meronyms of c_{k+1}, and c_{k+1} is linked with $\{c1,...,c_k\}$ by *part-of* links. The model M_{k+1} is thus built using the corresponding model described in Section 3.4.

Figure 8 shows the ratio $R_{mer} = \frac{C(X,M}{C(X,M')}$ according to the variance σ of the Gaussian distributions. As expected, R_{mer} converges to 1. Indeed, increasing the variance of the Gaussians amounts to saying that the features are less discriminative. This experience shows that the *part-of* link between labels provides all the more reduction of stochastic complexity that the features describe efficiently the data.

6.1.2 Real data

The reliability of the automatic learning of the model structure has been tested for real data on a database of SPOT5 images at 2,5m resolution. A training set has been created by annotating extracted images with a set of concepts listed on Table 2.

Synonymy relationship To evaluate the role of synonymy on stochastic complexity, the following protocol is applied for every label c listed on Table 6.1.2:

- The training set X_c associated to label c is split in two subsets denoted X_a and X_b of similar sizes (for us, X_a and X_b are therefore synonyms)

Table 1: Ratio of stochastic complexity obtained by the introduction of the synonymy relationship on image databases corresponding to different concepts

Labels	Forest	Town Center	Mountain	Residential Area	Sea
R_{syn}	1.00031	1.00042	1.00061	1.00012	1.0064

- Two models M_a and M_b are estimated on X_a and X_b respectively. The stochastic complexity $C(X_c, M)$ is computed, where M is defined as $M = \{M_a, M_b\}$.
- A single model M' is also estimated on $X_a \cup X_b = X_c$, and the stochastic complexity $C(X_c, M')$ is computed.

The ratio $\frac{C(X_c,M)}{C(X_c,M')}$ is computed for the different labels and the results are shown on Table 6.1.2. The ratio is higher than 1 for all the labels. This means that the synonymy link is detected for every label.

Hyponymy/hyperonymy relationship To evaluate the evolution of stochastic complexity caused by introducing hyponymy (a label corresponds to the generalisation of a set of other concepts), the following protocol is applied:

- A set of k concepts $\{c_1, \ldots, c_k\}$ associated respectively to k training sets $X_1, \ldots, X_k$ is considered. Then, each set X_i is split in two subsets X_i^1 and X_i^2. A concept c is introduced and is supposed to annotate a database X_{k+1}^1 defined by $X_{k+1}^1 = \bigcup_{i=1}^k X_i^2$.
- The $k+1$ models M_i associated to each concept c_i are supposed to lie on a single layer and are estimated separately on each database X_i^1.
- The concept c_{k+1} is linked with $\{c1,..., c_k\}$ by *kind-of* links. The model M_{k+1} is thus built using the corresponding model described in Section 3.4.

An experience has been made with the label "vegetation", which is built as a generalization of the labels "grass", "fields", and "forest". A second experience has been made with the label "urban", which is built as a generalization of the labels "residential area", "town center", and "industrial area".

To test the ratio $R_{hyp} = \frac{C(X,M)}{C(X,M')}$ according to the discriminative power of the low-level features, a noise is added on the textons of X_i in the following way:

Every texton of the map is changed with probability $p \in [0, 1]$ into another texton of value randomly chosen in $\{1, \ldots, n\}$. The ratio R_{hyp} is computed for p varying from 0 to 0,4. The results are shown on figure 9. As on the synthetic data, the ratio is decreasing with the value of the noise. With about 15% of noise, the ratio R_{hyp} goes under 1, this means that the hyponymic relationship is no longer identified by the system. The reduction of stochastic complexity induced by the introduction of a hyponymic relationship in the network thus directly depends on the discriminative power of the low-level features. Under a certain level of discriminative power, the hyponymic relationship cannot be infered.

Meronymy/Holonymy relationship To evaluate the evolution of stochastic complexity caused by introducing a link of hyponymy (a label corresponds to the generalisation of a set of other concepts), the following protocol is applied:

- A set of k concepts $\{c_1, \ldots, c_k\}$ associated respectively to k training sets $X_1, \ldots, X_k$ is considered. Then, a concept c is introduced and is supposed to annotate a training set X^1_{k+1} .
- The $k + 1$ models M_i associated to each concept c_i are supposed to lie on a single layer and are estimated separately on each database X^1_i.
- The concept c_{k+1} is is linked with $\{c1,..., c_k\}$ by *part-of* links. The model M_{k+1} is thus built using the corresponding model described in Section 3.4

An experience has been made with the label "urban area", which is built as a generalization of the labels "town center", "residential area", and "cemetery". A second experience has been made with the label "rural area", built as a generalization of the labels "sparse housing", "fields", "carrier", and "residential area".

To test the ratio $R_{mer} = \frac{C(X,M)}{C(X,M')}$ according to the discriminative power of the low-level features, a noise is added on the datasets on the textons of the datasets X_i in the same way as in previous Section. The ratio R_{mer} is computed for several values of p. The results are also shown on figure 9. As on the synthetic data, the ratio is decreasing with the value of the noise. With about 20% of noise, the ratio R_{mer} goes under 1, meaning that the hyponymic relationship is not identified by the system. The reduction of stochastic complexity induced by the introduction of a hyponymic relationship in the network thus directly depends on the discriminative power of the low-level features. Under a certain level of discriminative power, the hyponymic relationship cannot be infered.

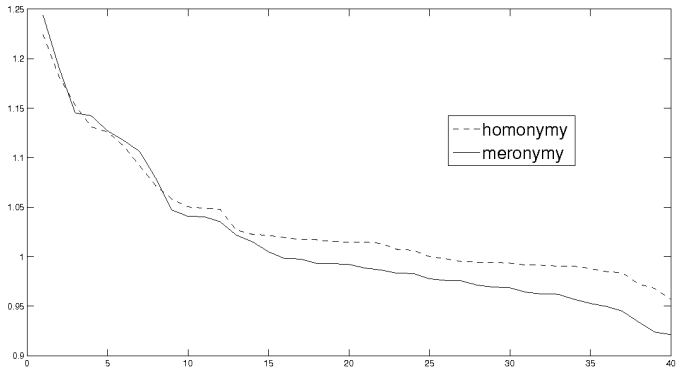


Figure 9: Ratio R_{hyp} and R_{mer} versus the percentage of noise. As long as the ratio is above 1, the relationships are recognized. When the noise increases, the recognition is lost.

Construction of a complete semantic network Experiences have been made on the database of SPOT5 images.

6.2 Annotation

In this section, the annotation performance of our method is evaluated. In the literature, a number of proposals for semantic image annotation and retrieval have appeared but it is quite difficult to compare the relative performances of the resulting algorithms with our method due to the lack of a proper experimental protocol. We thus decided to compare our method to the classical MPM segmentation method which has proved successfull for remote sensing segmentation [31].

6.2.1 Database

SPOT5 images of Paris, Marseille, Nice and Angers at 2,5m of resolution have been manually annotated and used to apply quantitative evaluation of the performance of semantic annotation. This database contains 24 images of size 6000×6000. The concepts used for annotation are those listed in Table 2 and are put in a two layer semantic network. Each image of the database is associated with two annotated partitions corresponding to each layer.

Industrial area
Residential area
Fields
Mountainous area
Forest
Water
Sparse housing area
Town center
Rural area
Refinery
Urban area
Industry
Cemetery
Carrier
Moutain
Airport
Maritim area

Table 2: Concepts used for semantic network

Concepts of second layer	**Concepts of first layer**
Rural area	Carrier, Forest, Sparse housing area, Fields, Residential area
Urban area	Town center, Residential area, Cemetery
Inudstrial complex	Refinery, Residential area, Industrial area
Moutainous area	Mountain, Carrier, Forest , Sparse housing area
Maritim area	Water, Residential area , Forest, Industrial area

Figure 10: *Part-of* links between first and second layer

6.2.2 Learning stage

The database has been split in three parts where each part contains a representative sampling of the landscapes existing in the whole database. A cross validation has been performed by applying three permutations on the three sets. Each time, two sets of the annotated database are used for the learning stage and one set for the test stage. Notice that as the structure is fixed, only the parameters are to be estimated. The parameters of the Markovian model are learnt using the method of the stochastic gradient.

6.2.3 Used metric

To evaluate the quality of the annotation of test images, the concepts of $C_2(S_{ko})$ are not taken into account as they depend deterministically of the annotations using S_{po}. The result of the annotation process using layer j of the model is an annotated partition $P_{sys} = \{R_1^j, R_2^j, ..., R_{m_j}^j\}$ of this image where each region R_i^j is annotated using the concept $c(R_i^j) \in C_j(S_{po})$. To evaluate the quality of a multi-layer annotation, we compare the annotation result for each layer to the ground truth $P_{gt} = \{R_1^{jr}, R_2^{jr}, ..., R_{m_j^r}^{jr}\}$ where each region R_i^{jr} is annotated with the concept $c(R_i^{jr}) \in C_j(S_{po})$. We assume that, in each layer, the set of annotation concepts used for ground truth is the one used by the system. For multi-layer annotation, we compare individually the segmentation of each layer. If the Vinet's distance [49] can be used to compare the produced segmentation, it is also relevant to use natural language processing tools as the goal of image annotation is to set images in relationship with words from natural language.

Indeed, the *World Error Rate* metric [19], used for evaluation of speech recognition systems, can be applied here by posing that a region R_1 of the system partition P_{sys} and a region R_2 of the ground truth partition P_{gt} can be matched if their overlap satisfy some given threshold:

$$\frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} > 0.8 \tag{23}$$

where $|R|$ corresponds to the number of textons of region R.

Given a matching between P_{sys} and P_{gt}, we define the following metric:

$$IAWER = \frac{|El| + |Add| + |Sub|}{N}$$

Where N is the number of words in the reference.

We define here:

- An elision (El) is a region of P_{gt} that has not been matched to any region of P_{sys}.
- An adding (Add) is a region of P_{sys} that has not been matched to any region of P_{gt}.
- A substitution (Sub) is a region of P_{sys} that has been matched to a region of P_{gt} but is not annotated by the same label.

For each layer, the matching is processed in order to minimize the IAWER. We use a greedy algorithm to find a minimum IAWER:

- For $i \in \{1, \ldots, m_j\}$. For each region R_i^j of P_{sys}, the overlap (cf eq 23) with all the regions of P_{gt} annotated with $c(R_i^j)$ is computed. Let $R_{i,opt}^{jr}$ be the region of P_{gt} with optimal overlap.
  - In case of conflict, if $R_{i,opt}^{jr}$ is matched already with a region R_k^j, $R_{i,opt}^{jr}$ is matched with the region corresponding to the higher overlap.
  - In other case, R_i^j is matched with $R_{i,opt}^{jr}$.

6.2.4 Results

The images are annotated using the method described in Section 5.1. The performances are evaluated by comparing the ground truth with the output of the system using the two metrics: Vinet's measure and IAWER. These two metrics are complementary as the Vinet measure evaluates the output of the system as a segmentation result, and the IAWER evaluates the output of the system as an annotation result.

As it can be seen on Table 3, the two algorithms provide rather similar results with the Vinet metric for the annotation with the labels of the first layer. A similar conclusion holds for the evaluation with IAWER criterion. The annotation results are much different as far as the second layer is concerned. The Markovian method is outperformed by PSL. A clear difference can also be noted with the results obtained by the Markovian method compared with the annotation of the first layer. This can be interpreted by the fact that the semantic gap cannot be crossed by a direct inference from low-level to high-level. This leads us to conclude with the relevance of our method making the inference in several steps by using several layers. Example of annotation are shown on fig 12 and 3.

	Layer 1.	Layer 2
PSL Vinet's measure	83,27 %	86,14%
MPM Vinet's measure	84,26%	68,27%
PSL IAWER	9,79 %	12,44%
MPM IAWER	8,91%	29,14%

Table 3: Annotation evaluation method using Vinet's measure (a good score is close from 100%), and using IAWER (a good score is close from 0%)

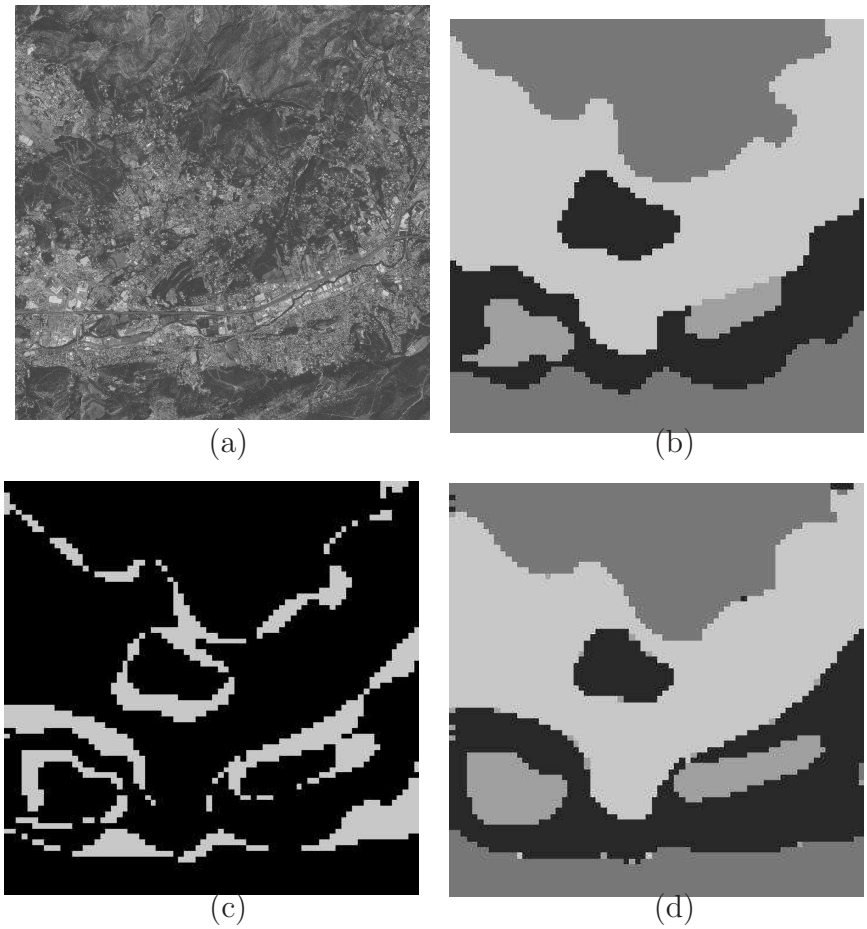
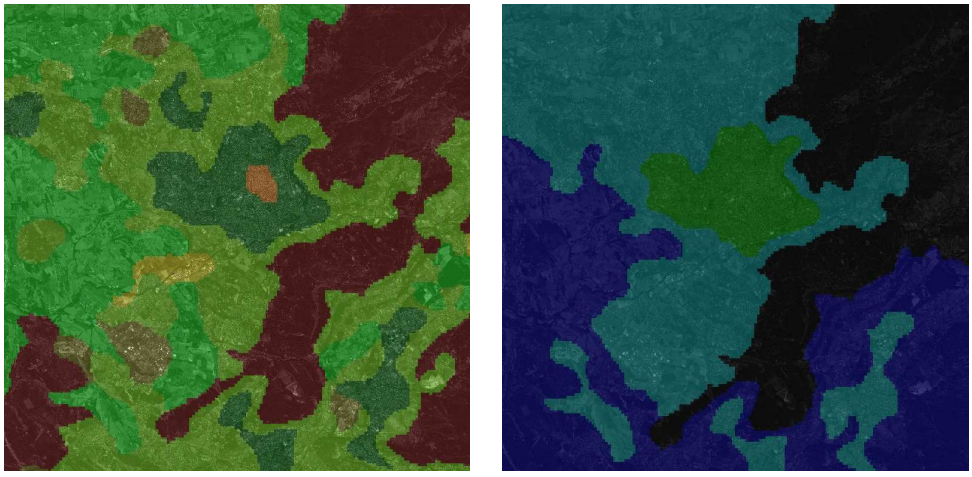
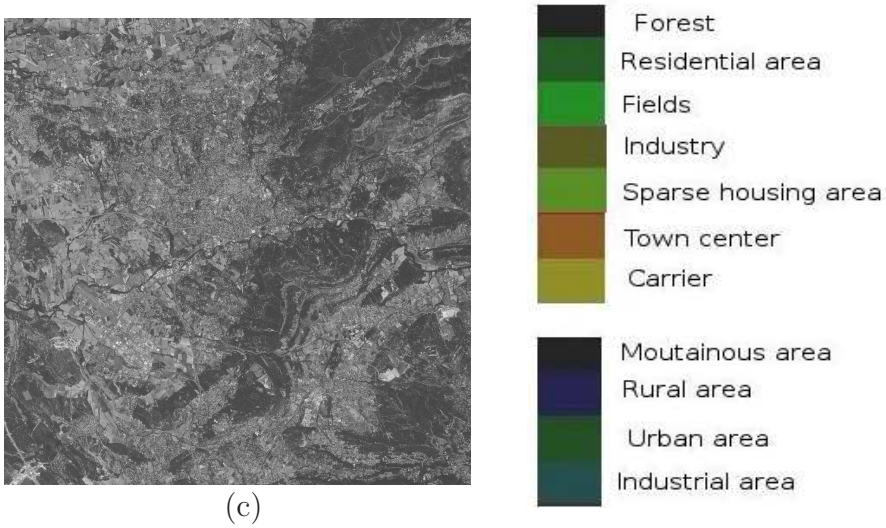


Figure 11: (a) Initial 6000 × 6000 SPOT5 image of "Marseille" @CNES (b) Ground truth mask of layer 1 (c) Misannotated pixels (d) Annotation by PSL method. Classes are: industrial area, sparse housing area, residential area and mountain,



(a) (b)



(c)

Figure 12: (a) Annotation with the first layer's labels (b) Annotation with the second layer's labels (c) Initial 6000 × 6000 SPOT5 image of "Aix en Provence" @CNES

7 Conclusion

In this work, we use semantic networks to take into account the semantical relationships between the concepts and to tackle the problem with high generality. We took advantage of a correspondance between the structure of the semantic networks and the structure of the probabilistical models. By considering the semantic relationships of hyonymy/hyperonymy and meronymy/holonymy, the concepts can be handled by the system considering their level of generality or complexity. This makes possible a good estimation of the density of high level concepts by expressing this density with the densities of concepts lower in the hierarchy. Moreover, the use of the semantic network requires no expert knowledge, as the semantic network is built automatically through an algorithm of model selection which infers the paradigmatic relationships from a weakly training set. Experiments prove the reliability of the construction of the semantic network, and the efficiency and richness of the annotation of semantic labeling and retrieval.

Acknowledgment

This work has benefitted from fruitful discussions with M. Datcu (DLR), Véronique Prinet (INRIA/Liama), Régis Behmo (Centrale Paris/Liama), O. Cappé (CNRS/LTCI). We thank CNES for providing the SPOT database.

References

[1] Shih-Fu Chang. Ana B. Benitez, John R. Smith. Medianet: A multimedia information network for knowledge representation. In *SPIE Conference on Internet Multimedia Management Systems*, volume 4210, 2000.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei, and M.I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[3] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

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

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

[7] Herbert Daschiel and Mihai Datcu. Image information mining system evaluation using information-theoretic measures. *EURASIP J. Appl. Signal Process.*, 2005(1):2153–2163, 2005.

[8] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, London, UK, 2002. Springer-Verlag.

[9] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *J. Intell. Inf. Syst.*, 3(3-4):231–262, 1994.

[10] F.Kummert, H.Niemann, R.Prechtel, and G.Sagerer. Control and explanation in a signal understanding environment. *Signal Processing*, 932(1-2):111–145, 1993.

[11] 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.

[12] Alexander Gammerman and Vladimir Vovk. Kolmogorov complexity: Sources, theory and applications. *The Computer Journal*, 42(4):252–255, 1999.

[13] James Hafner, Harpreet S. Sawhney, Will Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(7):729–736, 1995.

[14] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1778–1785, Washington, DC, USA, 2005. IEEE Computer Society.

[15] Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1992.

[16] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom. Mind the gap: another look at the problem of the semantic gap in image retrieval. In Edward Y. Chang, Alan Hanjalic, and Nicu Sebe, editors, *Multimedia Content Analysis, Management, and Retrieval 2006*, volume 6073. SPIE, 2006.

[17] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[18] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1998.

[19] M.J. Hunt. Figures of merit for assessing connected-word recognisers. In *ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, volume 2, pages 127–131, Noordwijkerhout, The Netherlands, 1989.

[20] A. Zisserman J. Sivic. Video google: a text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 471–478. Springer Berlin / Heidelberg, 2003.

[21] 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.

[22] J.Buckner, M.Pahl, and O.Stahlhut. Semantic interpretation of remote sensing data. *Int. Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(3):62–66, 2002.

[23] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, New York, NY, USA, 2003. ACM.

[24] Feng Kang, Rong Jin, and Joyce Y. Chai. Regularizing translation models for better automatic image annotation. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 350–359, New York, NY, USA, 2004. ACM.

[25] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, New York, NY, USA, 2002. ACM.

[26] Elliott Lieb. *The Stability of Matter*. Springer-Verlag, 1991.

[27] C. Liedtke, J. E., O. Grau, S. Growe, and R. Tonjes. Aida: A system for the knowledge based interpretation of remote sensing data. In *3rd Int.Airborne Remote Sensing Conference and Exhibition,*, volume 2, pages 313–320, Copenhagen, Denmark, 1997.

[28] D. Lowd and P. Domingos. Naive bayes models for probability estimation. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 529–536, New York, NY, USA, 2005. ACM.

[29] J. Lyons. *Language and Linguistics: an introduction*. Cambridge Univ Press, 1981.

[30] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 341–349, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[31] Pascale Masson and Wojciech Pieczynski. Sem algorithm and unsupervised segmentation of satellite images. *IEEE Transactions on GRS*, 31:61, 1993.

[32] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, New York, NY, USA, 2003. ACM.

[33] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 348–351, New York, NY, USA, 2004. ACM.

[34] 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.

[35] Milind R. Naphade, Igor Kozintsev, Thomas S. Huang, and Kannan Ramchandran. A factor graph framework for semantic indexing and retrieval in video. In *CBAIVL '00: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00)*, page 35, Washington, DC, USA, 2000. IEEE Computer Society.

[36] S. Newsam, L. Wang, S. Baghavaty, 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, 2004.

[37] H. Niemann, G. F. Sagerer, S. Schröder, and F. Kummert. Ernest: A semantic network system for pattern understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(9):883–905, 1990.

[38] 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, 2000.

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

[40] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1989.

[41] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[42] R.Tnjes, S.Growe, J.Bckner, and C-E.Liedtke. Knowledge based interpretation of remote sensing images using semantic nets. *Signal Processing*, 65(7):811–821, 1999.

[43] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. Continued in following volume.

[44] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.

[45] J.R. Smith and S. Chang. Querying by color regions using visualseek content-based visual query system. *Intelligent multimedia information retrieval*, pages 23–41, 1997.

[46] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan. Exploiting ontologies for automatic image annotation. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 552–558, New York, NY, USA, 2005. ACM.

[47] 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.

[48] 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.

[49] L. Vinet. *Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques*. PhD thesis, Université de Paris IX Dauphine, 1991.

[50] Fangshi Wang, De Xu, Hongli Xu, and Weixin Wu. Construction of semantic network for videos. In *ICICIC '06: Proceedings of the First International Conference on Innovative Computing, Information and Control*, pages 217–220, Washington, DC, USA, 2006. IEEE Computer Society.

[51] 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.

[52] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2(4):259–362, 2006.

Dépôt légal : 2008 – 3ème trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI