Modeling, classifying and annotating weakly annotated images using Bayesian network

Sabine Barrat and Salvatore Tabbone

LORIA - UMR 7503, BP 239, 54506 Vandœuvre-les-Nancy Cedex, France

Abstract

In this paper, we propose a probabilistic graphical model to represent weakly annotated images. We consider an image as weakly annotated if the number of keywords defined for it is less than the maximum number defined in the ground truth. This model is used to classify images and automatically extend existing annotations to new images by taking into account semantic relations between keywords. The proposed method has been evaluated in visual-textual classification and automatic annotation of images. The visual-textual classification is performed by using both visual and textual information. The experimental results, obtained from a database of more than 30000 images, show an improvement by 50.5% in terms of recognition rate against only visual information classification. Taking into account semantic relations between keywords improves the recognition rate by 10.5%. Moreover, the proposed model can be used to extend existing annotations to weakly annotated images, by computing distributions of missing keywords. Semantic relations improve the mean rate of good annotations by 6.9%. Finally, the proposed method is competitive with a state-of-art model.

Key words: probabilistic graphical models, Bayesian networks, image classification, image annotation, semantic similarity, Wordnet

1. Introduction

The rapid growth of Internet and multimedia information has shown a need in the development of multimedia information retrieval techniques, especially the image retrieval. We can distinguish two main trends. The first one, called "text-based image retrieval", consists in applying text-retrieval

techniques from fully annotated images. The text describes high-level concepts but this technique presents some drawbacks: it requires a tedious work of annotation. Moreover annotations could be ambiguous because two users can use different keywords to describe an image. Consequently some approaches [19, 1, 23] have proposed to use Wordnet [10] in order to reduce these potential ambiguities. The second approach, called "content-based image retrieval" [32] is a younger field. These methods rely on visual features (color, texture or shape) computed automatically, and retrieve images using a similarity measure. However the obtained performances are not really acceptable, except in the case of well-focused corpus.
In order to improve the recognition, a solution consists in combining visual and semantic information. Some researchers have already explored this possibility [2, 3, 15, 20, 36, 7, 29].
Automatic image annotation [33, 28] can be used in image retrieval systems to organize and locate images of interest from a database, or to perform visual-textual classification. This method can be seen as a type of multi-class image classification with a very large number of classes, as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques attempting to automatically apply annotations to new images. Many works have been proposed in this sense. We can cite, without being exhaustive, classification-based methods [13, 39], probabilistic modeling-based methods [4, 12, 6], annotation refinement [38, 31] and discriminative methods [17, 14]. For example, the paper in [38] proposes to segment images into visual tokens described by color, texture and shape features. A clustering algorithm is applied to group similar visual tokens and relevant features, are selected in each cluster. A weight is assigned to each feature in each cluster, according to how relevant the feature is to the cluster. Finally links are determined between keywords and blob-tokens. To annotate an image automatically, the distance between visual features of a given image and the visual features of all centroids of blob-tokens is computed. Another way to annotate images consists in classifying images in semantic categories [37]. This kind of method presents the drawback to require that each keyword corresponds to a class. In fact, a same word cannot annotate images of different classes. Concerning the probabilistic modeling-based methods, they consist in learning associations between images and keywords. The first outstanding work in this sense, proposed by Mori et al. [27] in 1999, is a co-occurrence model. This model consists of the count of co-occurrences of keywords and graphical

features from a training set. The counts are used to predict the keywords for other images. This model has the drawback to require discrete features or a pre-discretization step. This work has been improved, in 2002, by Duygulu et al. in [9] by the introduction of a statistical model of translation. In this approach, images are segmented into regions classified in function of theirs graphical features. A relation between region classes and keywords is then learnt, using a method based on EM algorithm. This process similar with learning a lexicon from an aligned bitext accepts continuous features but requires a manual annotation of regions. Jeon et al. [16] introduced the Cross-Media Relevance Model (CMRM) which uses keywords shared by some images to annotate new images. In fact, like in the approach [9], images are supposed to be described by a little vocabulary associated to classes of image regions. By using a training set of annotated images, the joint probability distribution of region classes and keywords is learnt. This method has then been improved by the Continuous-space Relevance Model [22]. In this approach each image is divided into regions and each region is described by a vector of continuous features. From a training set of annotated images, a probabilistic model of features and keywords is learnt, in order to predict the probability to generate a keyword knowing the features of image regions. This model has the same drawback as the models [9, 16]: it requires a manual annotation of regions of some images, which is costly for the user. In [40], EM algorithm and Bayes rule are used to connect each feature to keywords: each image is annotated by the keyword which has the highest probability given the visual features of the test image. This probability is obtained thanks to semantic concepts and Bayes rule. Jin et al. [18] propose a language model to annotate images. This language model is used to estimate the probability of a keyword set given an image. The set with the highest probability is assigned to the image. Due to a preset threshold, some images cannot be annotated and the user has to manually annotate them.

In addition, there have been a number of papers in visual and textual information combination. In [2], Barnard et al. segment the images into regions. Each region is represented by a set of visual properties and a set of keywords. Then, the images are clustered by hierarchically modeling their distributions of words and image feature descriptors. Grosky et al. [15] use Latent Semantic Indexing (LSI) and word weighting schemes to reduce the dimensionality. Feature vectors of visual feature descriptors and category label bits are concatenated in order to retrieve images. Benitez et al. [3] extract knowledge from annotated image collections by clustering the images based on visual

feature descriptors and text feature descriptors. Perceptual relationships, based on descriptor similarity and statistics between clusters, are discovered. Magalhaes et al. [25] use information theory to develop a model that supports heterogeneous types of documents (text documents, image documents, or documents with both text and images). Also, to take into account the subjectivity of human perception and bridge the gap between the high-level concepts and the low-level features, relevance feedback has been proposed to enhance the retrieval performance [20]. Finally, more sophisticated graphical models, such as the approach described in [5], based on Multinomial Dirichlet Mixture models or Gaussian-multinomial mixture model (GM-Mixture), Latent Dirichlet Allocator (LDA) and Correspondence LDA (CLDA), have also been applied to the image annotation problem [4]. In the same way, the model [26], uses non-parametric methods to estimate probabilities within an inference network and can be used to image retrieval and annotation. For more details on semantic information extraction from multimedia content, we refer the reader to the survey in [24]. These probabilistic methods are named "generative". They try to construct a model of the system which has generated the observed data, and provide decision rules from this modeling. We distinguish generative methods from discriminative ones. Discriminative approaches directly provide decision rules, without taking into account the features of the system which has generated the data. Such methods can be used in automatic image annotation. For example, the method in [14] employs a confidence-based dynamic ensemble (CDE), using one-class, two-class, and multiclass SVMs to annotate images for supporting keyword retrieval of images.

The contribution of this paper is to propose a scheme for image classification optimization, using a joint visual-text clustering approach and automatically extending image annotations. The proposed approach is derived from the probabilistic graphical model theory. More precisely, the model presented here is dedicated for both tasks of weakly-annotated image classification and annotation. In fact the classification methods before mentioned are efficient but they require that all images, or image blobs are annotated. Moreover most existing annotation models are not able to classify images. We introduce a method to deal with missing data in the context of text annotated images as defined in [4, 20]. The proposed model does not require that all images be annotated: when an image is weakly annotated, the missing keywords are considered as missing values. Besides our model can automatically extend existing annotations to weakly-annotated images, without User inter-

vention. The uncertainty around the association between a set of keywords and an image is tackled by a joint probability distribution over the dictionary of keywords and the visual features extracted from our collection of images. The model [4] is the most related to our approach, because it enables to classify images based on visual and textual features and to automatically annotate new images. However our model is less restrictive for the user. In fact our classifier does not need that all images should be annotated. Moreover, the model [4] assumes that the keywords are independent given its parents. On the contrary our model has the advantage to take into account the possible semantic relations between keywords. In fact, semantic relations, as defined in Wordnet, are represented by edges in our Bayesian network. We will show that these semantic relations improve the recognition rate as well the mean rate of good annotations.
The rest of this paper is organized as follows. Section 2 describes the probabilistic model of weakly-annotated image representation and how to use it to classify and to extend existing annotations to images. Experimental results on a database of more than 30000 images are given in section 3. Also, a comparison with the GM-Mixture model [4] is provided. Finally, a conclusion and future works are given to our work (section 4).

2. Representation and classification of weakly-annotated images

Our work is focused on weakly-annotated image modeling and classification. Now visual descriptors often provide vectors of continuous values, and the associated keywords often correspond to discrete variables. So we have chosen to construct a Bayesian classifier which allows discrete and continuous variable combination and to manage missing values.

Let f_j be a query image characterized by a set of features F. F is composed of:

- m visual features, denoted $v_1, ..., v_m$,
- n possible keywords, denoted $KW\ 1, ..., KW\ n$.

The chosen visual features are issued from one color descriptor (a color histogram) [34] and one shape descriptor based on the Fourier/Radon transform [35]. We are interested in the probability distributions of these features and their conditional dependence relations. Let us consider the visual features as continuous random variables and theirs associated keywords as discrete

variables. This model is too large to be represented as a unique joint probability distribution, therefore it is required to introduce some sparse and structural *a priori* knowledge. The probabilistic graphical models, and especially Bayesian networks, are a good way to solve this kind of problem. In fact within Bayesian networks the joint probability distribution is replaced by a sparse representation only among the variables directly influencing one another. Interactions among indirectly-related variables are then computed by propagating inference through a graph of these direct connections. Consequently the Bayesian networks are a simple way to represent a joint probability distribution over a set of random variables, to visualize the conditional properties and to compute complex operations like probability learning and inference, with graphical manipulations. Then, a Bayesian network seems to be appropriate to represent and classify images with associated keywords.

2.1. A sparse image/keyword representation

Formally, a Bayesian Network for a set of random variables V is a pair $B =< G, \Theta >$. The first component, G, is a directed acyclic graph whose vertices correspond to random variables $V_1, ..., V_n$, and whose edges represent direct dependencies between variables. The graph G encodes independence assumptions: each variable V_i is independent of its non descendants given its parents in G. The second component of the pair, Θ, represents the set of parameters that quantifies the network. It contains a parameter $\theta_{v_i|Pa(v_i)} = P_B(v_i|Pa(v_i))$ for each possible value v_i of V_i, and $Pa(v_i)$ of $Pa(V_i)$, where $Pa(V_i)$ denotes the set of parents of V_i in G. That is, the Bayesian network, in its initial state, contains the initial *a priori* probabilities of each node of the network: $P_B(v_i|Pa(v_i))$. Thanks to the conditional independence assumption of each variable given its parents, the joint probability distribution $P_B(V_1, ..., V_n)$ can be reduced to this formula:

$$P_B(V_1, ..., V_n) = \prod_{i=1}^{n} P_B(V_i|Pa(V_i)) = \prod_{i=1}^{n} \theta_{v_i|Pa(v_i)}$$

But our aim is to assign a new image designed by a particular instance $f = \{f_1, ..., f_n\}$ of the feature vector $F = \{F_1, ..., F_n\}$ to a class c_i among k classes. In order to perform this goal, we induce a Bayesian network NB (for Naive Bayes), that encodes a distribution $P_{NB}(F_1, ..., F_n, C)$, from a given training set (composed of labeled data). Then the resulting model can

be used to classify the new instance I. In fact, let $f_1, ..., f_n$ be the features extracted of I. The Bayes rule is applied to compute the probability of c_i given the particular instance f. Then the classifier based on NB returns the label c_i, $i \in \{1, ..., k\}$, that maximizes the posterior probability $P = P_{NB}(c_i|f_1, ..., f_n)$, where:

$$P = \frac{P_{NB}(f_1, ..., f_n|c_i) \times P_{NB}(c_i)}{P_{NB}(f_1, ..., f_n)}$$

and $P_{NB}(f_1, ..., f_n) = \sum_{j=1}^{k} P_{NB}(f_1, ..., f_n|c_j) \times P_{NB}(c_j)$

This simple Bayesian classifier, called Naive Bayes, can be represented by the structure in Figure 1, where:

- C refers to the class variable,
- $F_1, ..., F_n$ are the feature variables.

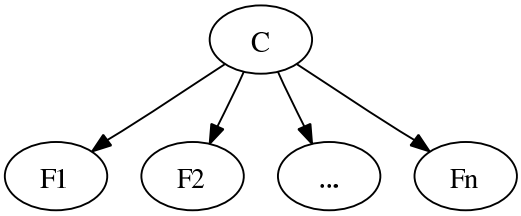


Figure 1: Naive Bayes

2.2. A Gaussian-Mixtures and Bernoulli mixture model

The naive Bayes is a simple and efficient model, but it requires discrete variables. However, we have to manage continuous variables (corresponding to the visual features) and discrete variables (corresponding to the keywords). Therefore a Bayesian classifier, which involves both types of variables, is proposed. We present a hierarchical probabilistic model of multiple-type data (images and associated keywords) in order to classify large databases of annotated images. A Gaussian-Mixtures and Bernoulli Mixture model is proposed. In fact, the observation of some peaks on the different histograms of the feature variables, has led us to consider that the visual features can be

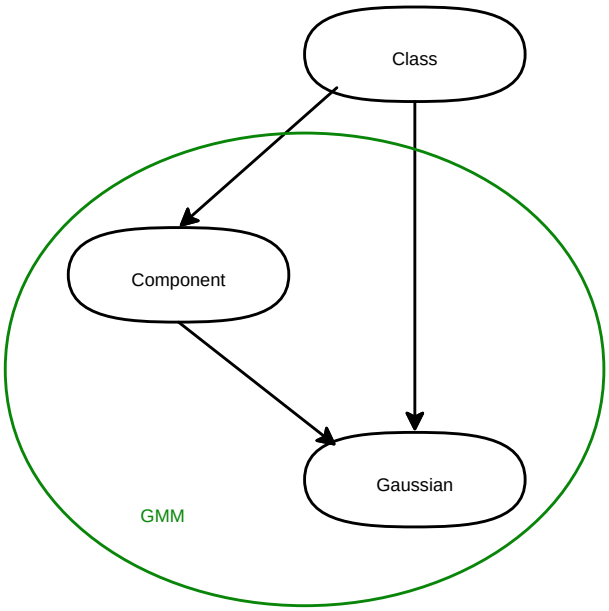


Figure 2: A Probabilistic graphical model as GMMs

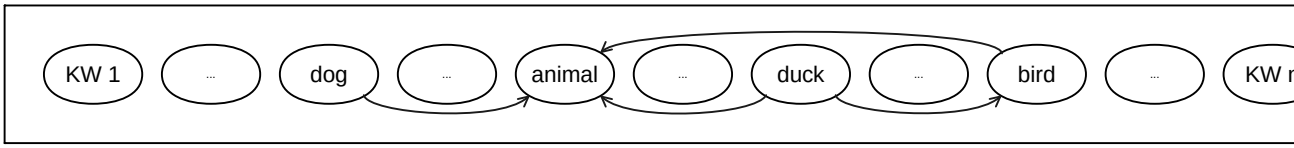


Figure 3: dependences between keywords

estimated by mixtures of Gaussian densities. The discrete variables corresponding to the words of the vocabulary have a Bernoulli distribution: in fact, for a given image, each keyword variable can take two states: "true" when the word annotates the given image, or "false", when the word does not belong to the given image annotation. A Bernoulli distribution has been prefered to a multinomial distribution because Bernoulli distribution enables to represents dependences between keyword variables. It is not the case with Multinomial distribution, because Multinomial distribution considers each keyword of annotation as a discrete variable, and associates a probability to each keyword of the vocabulary. In our case, each keyword of the vocabulary is considerated as a Bernoulli discrete variable which associates a probability of presence and a probability of absence (equal to 1− the probability of pres-

ence) of each keyword as annotation. By this way, the Bernoulli distribution does not fix a limit to the number of keywords per image. On the contrary, Multinomial distributions fix a limit because the number of keyword variables is defined to construct the model.
Now let F be the training set composed of m instances $f_{1_i}, ..., f_{m_i}, \forall i \in \{1, ..., n\}$, where n is the dimension of the signatures provided by the concatenation of the feature vectors issued from the computation of all the descriptors on each image on the training set. Each instance $f_j, \forall j \in \{1, ..., m\}$ is characterized by n continuous variables. A supervised classification is considered and F instances are divided into k classes $c_1, ..., c_k$. Let $G_1, ..., G_g$ be g groups whose each has a Gaussian density with a mean $\mu_l, \forall l \in \{1, ..., g\}$ and a covariance matrix $\sum_l$. Besides, let $\pi_1, ..., \pi_g$ be the proportions of the different groups, $\theta_l = (\mu_l, \sum_l)$ be the parameter of each Gaussian and $\Phi = (\pi_1, \pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ the global mixture parameter. The probability density of F conditionally to the class $c_i, \forall i \in \{1, ..., k\}$ can be defined by

$$P(f, \Phi) = \sum_{l=1}^{g} \pi_l p(f, \theta_l)$$

where $p(f, \theta_l)$ is the multivariate Gaussian defined by the parameter θ_l.

Then, we have one Gaussian Mixture Model per class. This problem can be represented by the probabilistic graphical model in Figure 2, where:

- The "Class" node is a discrete node, which can take k values corresponding to the pre-defined classes $c_1, ..., c_k$.
- The "Component" node is a discrete node which corresponds to the components (i.e. the groups $G_1, ..., G_g$) of the mixtures. This variable can take g values, i.e. the number of Gaussians used to compute the mixtures. It is an hidden variable which represents the weight of each group (i.e. the $\pi_l, \forall l \in \{1, ..., g\}$).
- The "Gaussian" node is a continuous variable which represents each Gaussian $G_l, \forall \in \{l = 1, ..., g\}$ with its own parameter $(\theta_l = (\mu_l, \sum_l))$. It corresponds to the set of feature vectors in each class.
- Finally the edges represent the effect of the class on each Gaussian parameter and its associated weight. The green circle does not belong to the graphical model : it is just a way to show the relation between

the proposed probabilistic graphical model and GMMs: we have one GMM (encircled in green), composed of Gaussians and their associated weight, per class.

Now the model can be completed by the discrete variables, denoted $KW\ 1, ..., KW\ n$, where n is the size of the vocabulary, and KW_i represents each keyword of the vocabulary. Dirichlet priors [30], have been used for the probability estimation of the variables $KW\ 1, ..., KW\ n$. That is we introduce additional pseudo counts at every instance in order to ensure that they are all "virtually" represented in the training set. Therefore every instance, even if it is not represented in the training set, will have a not null probability. Like the continuous variables corresponding to the visual features, the discrete variables corresponding to the keywords are included in the graphical model by connecting them to the class variable. Finally our model can be depicted by the Figure 4. The hidden variable α shows that a Dirichlet prior is used. The box around the variable KW denotes n repetitions of KW, for each keyword of the vocabulary. n is the size of the vocabulary. The edges representing semantic relations between keywords are not drawn in the box, to keep more clarity. Figure 3 represents more precisely the keyword variables and theirs potential dependences. The n nodes correspond to the n keywords of the vocabulary: $KW\ 1, ..., KW\ n$. Only some keyword dependences are represented. For example "bird" and "animal" have a semantic relation, which is represented by a directed edge from the node "bird" to the node "animal". In the same way an edge is observed between the nodes "duck" and "animal" and the nodes "duck" and "bird". In general, an edge is added between two keywords of the same synset (as defined in Wordnet [10]). Each edge between two keywords is directed from the most specific keyword (hypernym) to the most general keyword (hyponym). Concerning the edges between two keywords where a keyword is a part of the other, the edge is directed from the keyword which is the part of the other (meronym), to the keyword representing the "whole" (holonym). In this way we represent Wordnet's ontologies. Some dependences of our database are given in section 3 (Table 2).
The structure of this model has been established at the hand. No learning algorithm of structure has been used. In the same way, semantic relations have been established at the hand, by taking each couple of keywords of the vocabulary and searching in Wordnet a possible semantic relation between the keywords of each couple.

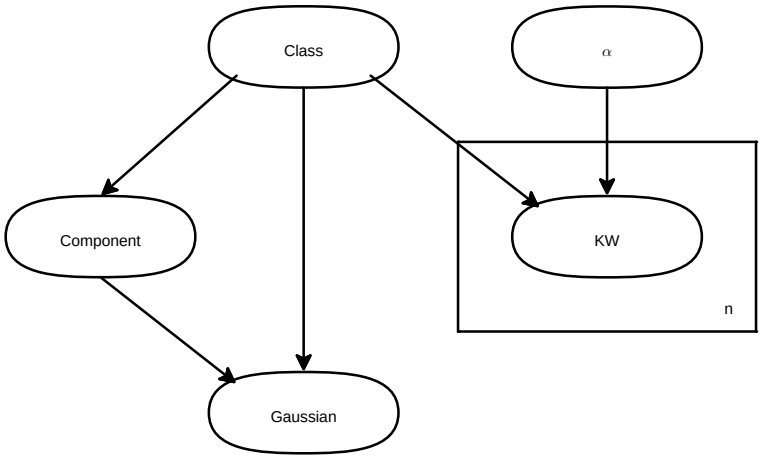


Figure 4: The Gaussian-Mixtures and Bernoulli mixture model

This Bayesian classifier (4) means that each image and its keywords are assumed to have been generated conditional on the same class. Therefore the resulting Bernoulli and Gaussian mixture parameters should correspond: concretely if an image, represented by visual descriptors, has an high probability under a certain class, then its keywords should have an high probability under the same class. We keep in mind that the notions of "keywords" and "classes" have not the same sense. In fact, a database is classified in several classes. And each image of the database can be annotated by several keywords. The value of an image keyword does not determine the class value of the image.

2.3. Parameter learning and inference

Our major problem deals with missing values. Indeed, the color features for gray-level images, and especially some keywords for a large subset of images, are missing. Concerning the visual features, the missing values are clearly homogeneously distributed (because they correspond to gray-level images). But the missing values are randomly distributed for the variables $KW\ i$, $\forall i \in \{1, ..., n\}$. We have used EM algorithm because, with this algorithm, we can learn the Gaussian mixture parameters and tackle the problem of missing values.
The general purpose of this algorithm, detailed in [8], consists in computing, in an iterative way, the likelihood maximum when the instances can be viewed as incomplete data: each iteration consists of an Expectation compu-

tation step before a Maximization step. This algorithm has been chosen for its simplicity and generality.

An inference algorithm is also necessary to classify new images. Indeed, the inference process consists in computing posterior probability distributions of one or several other subsets of nodes. In the case of classification, the class node is inferred. According to our Bayesian network topology, the inference process propagates the values from the image feature level represented by the "Gaussian" node, trough the "Component" and Keyword nodes, until the "Class" node level. A message passing algorithm [21] is applied to the network. In this technique, each node is associated to a processor, which can send some messages to its neighbors, in an asynchronous way, until it reaches stability.

Thus a query image f_j, characterized by its visual features $v_{j_1}, ..., v_{j_m}$ and its possible keywords $KW\ 1_j, ..., KW\ k_j$ is considered as an "evidence" represented by:

$$P(f_j) = P(v_{j_1}, ..., v_{j_m}, KW\ 1_j, ..., KW\ n_j) = 1$$

when the network is evaluated. Thanks to the inference algorithm, the probabilities of each node are updated in function of this evidence. After the belief propagation, we know, $\forall i \in \{1, ..., k\}$, the posterior probability $P(c_i|f_j) = P(c_i|v_{j_1}, ..., v_{j_m}, KW\ 1_j, ..., KW\ n_j)$. The query f_j is assigned to the class c_i which maximizes this probability.

2.4. Annotation extension of images

Given an image without keyword, or a weakly annotated image [1], the proposed Bayesian model described before can be used to compute a distribution over words conditionally to the image and its possible existing keywords. In fact, for a query image f_j annotated by a set of $k, \forall k \in \{0, ..., n\}$ keywords, denoted EKW (for Existing KeyWords) where n is the size of the vocabulary, the inference algorithm enables to compute the posterior probability $P(KW_{i_j}|f_j, EKW)\ \forall KW_{i_j} \notin EKW$. This distribution represents a prediction of the missing keywords for that image.

[1] we recall that we consider an image as weakly annotated if the number of keywords defined for it is less than the maximum defined in the ground truth

For example, let us consider Table 1 which presents 4 images with possible keywords and the keywords obtained after automatic annotation extension with (column 3) or without (column 2) considering potential semantic relations between keywords. The first image annotation, composed of 3 keywords at the beginning, has been extended by one wrong keyword. In fact, the good missing keyword is "shrubs". This mistake is probably due to the large number of database images annotated by these 4 keywords "bear", "black", "water" and "grass", which generates a hight joint probability of this keyword set. Considering the second image, its annotation has not been extended without taking into account semantic relations between keywords. It is due to the preset threshold used to select keywords. In fact, a keyword is selected as annotation if the probability of this keyword as annotation is strictly greater than a preset threshold, defined to 0.5 in our experiments. The annotation has not been extended because no probability has exceeded the threshold. On the contrary, by taking into account semantic relations between keywords, the second image has been annotated by a correct keyword "water", thanks to the existing semantic relation between the keywords "river" and "water" which increases the probability of the keyword "water" given the keyword "river". Finally, the third and the last images belong to the same class "penguin". The third penguin image is fully annotated. On the contrary, the fourth image has two missing keywords. Its annotation has been extended by a wrong third keyword "iceberg" (with or without semantic relations between keywords). In fact, this third keyword should have been "snow". This mistake is due to the high color similarity between the third and the fourth image. Finally, thanks to the relations between keywords, the fourth image annotation has been extended with a correct fourth keyword "bird". In fact, the existing semantic relation between the keywords "penguin" and "bird" (see Table 2) increases the probability of the keyword "bird" given the keyword "penguin".

3. Experimental results

In this section, we present an evaluation of our model on more than 30000 images from the Corel image libraries and kindly provided by Vasconcelos and al. [6]. These images are split up into 306 classes. Knowing a keyword of an image does not determine the class value of this image. In fact, a same keyword can appear in the annotation of images of different classes. For example, 4 images of different classes with the keyword "duck" in common,

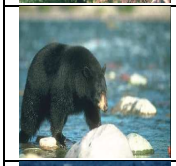
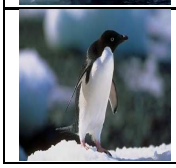
image	initial keywords	after annotation extension without SR	after annotation extension with SR
	bear black water	bear black water grass	bear black water grass
	bear black river	bear black river	bear black river water
	penguin iceberg water bird	penguin iceberg water bird	penguin iceberg water bird
	penguin water	penguin water iceberg	penguin water iceberg bird

Table 1: Examples of images and possible keywords before and after annotation extension with and without taking into account semantic relations

are given in Table 3. Moreover, some classes are overlapped, i.e. some images of the database are in two different classes.

Finally, all images have not the same number of keywords. 72% of the images of the database are annotated by 4 keywords, 23% by 3 keywords, 4% by 2 keywords and 0.5% by 1 keyword (i.e. 99.5% of the images are annotated by at least 1 keyword), using a vocabulary set of 1036 keywords. Therefore, in this database, the images annotated by less than 4 keywords are considered as weakly annotated. In the same way, all classes have not the same number of annotated images. Moreover, the database does not contains only animals, but other types of images. For example it contains, without being exhaustive, images of landscapes, particular places, peoples of different countries, sport scenes, every day objects, foods, ...

Figure 5 shows a screen shot of our application graphics interface. This view enables to select training and test images. Selected images are outlined in black before validation. Keyword annotations of a given image can be

obtained by moving the mouse cursor over this image.



Figure 5: Image selection in the class "arabian horses"

Once training and test sets are validated, training images are outlined in blue and tests images are outlined in green. For example, Figure 6 shows a subset of the training and test images of the class "arabian horses". This selection is repeated for each class of interest.

First of all, some dependences between keywords have been established from the vocabulary. We define the dependence relation between two keywords of the same synset (semantic group), as defined in Wordnet [10]. Wordnet is a large lexical database of English language, where the words (nouns, verbs, adjectives and adverbs) are grouped into sets of cognitive synonyms (denoted synsets), each expressing a distinct concept. That is two keywords



Figure 6: Subsets of training and test images in the class "arabian horses"

having a semantic relation would be grouped in the same synset. These semantic relations are represented by dependences in our model, i.e. by links in the Bayesian network. Some of these dependences are given in the table 2. The first column contains the keywords which are the source of the dependence with the corresponding keyword in the second column. The last column gives the type of semantic relation between the two corresponding keywords.

We have evaluated our method by performing 6 cross validations whose each proportion of the training set is 25%, 35%, 50%, 65%, 75% and 90% of the database, the remaining respectively 75%, 65%, 50%, 35% and 10% are used for the test. In each case the tests are repeated 10 times in order that each image would be used for the training and the test. For each training set size, the recognition rate is obtained by taking the mean recognition rate of 10 tests. For each test, the recognition rate corresponds to the ratio between the number of good classified images and the number of images in the testing set. In all the tests, our Gaussian-Mixtures and Bernoulli mixture model (denoted GM-B) has been performed with mixtures of 2 Gaussians and diagonal covariance matrices.

Let us consider Table 4. Our Gaussian-Mixtures and Bernoulli mixture model has been used to combine different types of information. The notation "C + S" means that the color and shape descriptors ("C" for Color, "S" for Shape) have been combined and "C + S + KW" adds textual information (KW for keywords). The recognition rates confirm that combining visual

with semantic features performs always better than any of them alone.

Table 5 shows the recognition rates obtained with our GM-B model, by taking into account semantic relations between keywords (column "with SR", SR for semantic relations), or not (column "without"). These results show that semantic relations between keywords improve the recognition by 10.5%. Moreover, Table 5 shows the effectiveness of our approach (GM-B model) compared to the Gaussian-multinomial mixture model (GM-Mixture) [4]. The GM-Mixture model has been used without image segmentation, as in our approach: the color and shape descriptors have been computed on the whole images and the keywords are associated to the whole images too. Moreover, as a supervised classification problem is considered in this paper, the discrete variable z used in [4] to represent a joint clustering of an image and its caption, is not hidden for the images of the training set. Actually, this discrete variable corresponds to our class variable and the number of clusters is known (it is our number of classes). The results have been obtained by using the visual features and theirs possible associated keywords. It appears that with the semantic relations between keywords, our GM-B model has a better mean recognition rate than the GM-Mixture model. Moreover, for each training set size, a Student t-test for paired samples [11] has been used to compare the mean recognition rates (over the 10 tests of the cross validation) of the GM-Mixture and the GM-B model. Whatever the size of training set, the t-value (see standard deviations of difference and t-values in Table 6)) shows that the mean recognition rate obtained by our GM-B model, with semantic relations, is statistically different than that of GM-Mixture, with a risk smaller than 1% and a freedom degree of 9.

Now, let us consider the annotation extension problem. At least a keyword annotation per image is needed to compare the annotations after automatic annotation extension to the ground truth annotations. Therefore, 99.5% of the database images, annotated by at least 1 keyword, have been selected as ground truth. Like for the classification evaluation, 6 cross validations have been performed. The tests are repeated 10 times in order that each image would be used for the training and the test. For each test, the test images have been automatically annotated by 4 keywords. For each training set size, the rate of good annotations is obtained by taking the mean rate of the 10 tests. For each test, the rate of good annotations corresponds to the ratio between the number of annotations obtained automatically which corresponds to the ground truth and the number of keywords obtained automatically. The threshold used for annotation has been fixed at 0.5. That is

to say, for a given image, a keyword is selected as annotation if his probability to annotate this image, knowing the visual features and possible existing keywords of this image, is strictly greater than 0.5. Table 7 compares the rate of good annotations obtained by taking into account semantic relations between keywords. We can observe that semantic relations between keywords improve the rate of good annotations by 6.9%. Table 7 compares also the rates of good annotations obtained by the GM-Mixture model and our GM-B model. We can see that our model is better than the GM-Mixture model, even if we do not take into account semantic relations between keywords. Finally, using the same test as in Table 6, we can remark that the mean rate of good annotations obtained by our GM-B model, with semantic relations, is statistically different than that of GM-Mixture (see t-values in Table 8).

4. Conclusion and future works

We have proposed a method for modeling, classifying and annotating weakly annotated images. This method has the advantage to take into account semantic relations between annotations. In fact, experimental results of joint visual-textual classification have demonstrated that semantic relation representation improves the recognition rate. Moreover, our approach has especially shown good performances in automatic annotation extension. Finally, the evaluation has shown that our model is competitive with a state-of-art model.
Further works will be devoted to capture the user's preference by considering a relevance feedback process. More precisely, the user's preference can be represented by the network parameter update (i.e. the probabilities of each variable in function of the new classified instance) during the inference process.

References

[1] Aslandogan, Y. A., Thier, C., Yu, C. T., Zou, J., Rishe, N., 1997. Using semantic contents and wordnet in image retrieval. SIGIR Forum 31 (SI), 286–295.

[2] Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., Jordan, M. I., 2003. Matching words and pictures. Journal of Machine Learning Research 3 (6), 1107–1135.

[3] Benitez, A., Shih-Fu, C., 2002. Perceptual knowledge construction from annotated image collections. in ICME '02 1, 189–192.

[4] Blei, D., Jordan, M., 2003. Modeling annotated data. In: SIGIR '03. pp. 127–134.

[5] Bouguila, N., Ziou, D., 2007. Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization. Journal of Visual Communication and Image Representation 18 (4), 295–309.

[6] Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N., 2007. Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 29 (3), 394–410.

[7] Chang, C., Wang, H., Li, C., 2009. Semantic analysis of real-world images using support vector machine. Expert Systems with Applications 36 (7), 10560–10569.

[8] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), 1–38.

[9] Duygulu, P., Barnard, K., de Freitas, J. F. G., Forsyth, D. A., 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV. pp. 97–112.

[10] Fellbaum, C. (Ed.), 1998. WordNet - An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

[11] Feller, W., 1968. An Introduction to Probability Theory and Its Applications. Vol. 1. Wiley.

[12] Feng, S. L., Manmatha, R., Lavrenko, V., 2004. Multiple bernoulli relevance models for image and video annotation. CVPR '04 2, 1002–1009.

[13] Gao, Y., Fan, J., Xue, X., Jain, R., 2006. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In: ACM MULTIMEDIA '06. pp. 901–910.

[14] Goh, K., Chang, E., Li, B., 2005. Using one-class and two-class svms for multiclass image annotation. IEEE Transactions on Knowledge and Data Engineering 17 (10), 1333–1346.

[15] Grosky, W. I., Zhao, R., 2001. Negotiating the semantic gap: From feature maps to semantic landscapes. In: SOFSEM '01. pp. 33–52.

[16] Jeon, J., Lavrenko, V., Manmatha, R., 2003. Automatic image annotation and retrieval using cross-media relevance models. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 119–126.

[17] Jeong, J., Park, K., Lee, O., Lee, D., 2007. Automatic extraction of semantic relationships from images using ontologies and svm classifiers. In: MCAM '07. pp. 184–194.

[18] Jin, R., Chai, J. Y., Si, L., 2004. Effective automatic image annotation via a coherent language model and active learning. In: MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia. pp. 892–899.

[19] Jin, Y., Khan, L., Wang, L., Awad, M., 2005. Image annotations by combining multiple evidence & wordnet. In: MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia. pp. 706–715.

[20] Kherfi, M. L., Brahmi, D., Ziou, D., 2004. Combining visual features with semantics for a more effective image retrieval. In: ICPR '04. Vol. 2. pp. 961–964.

[21] Kim, J. H., Pearl, J., 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In: IJCAI-83. pp. 190–193.

[22] Lavrenko, V., Manmatha, R., Jeon, J., 2003. A model for learning the semantics of pictures. In: Proceedings of the 16th Conference on Advances in Neural Information Processing Systems (NIPS).

[23] Liu, S., Liu, F., Yu, C., Meng, W., 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR

conference on Research and development in information retrieval. pp. 266–272.

[24] Magalhaes, J., Rüger, S., 2006. Semantic-Based Visual Information Retrieval, yu-jin zhang Edition. IDEA group publishing.

[25] Magalhaes, J., Rüger, S., 2007. Information-theoretic semantic multimedia indexing. In: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 619–626.

[26] Metzler, D., Manmatha, R., 2004. An inference network approach to image retrieval. In: CIVR. pp. 42–50.

[27] Mori, Y., Takahashi, H., Oka, R., 2000. Automatic word assignment to images based on dividing image and vector quantization. In: RIAO. pp. 285–293.

[28] Poppe, C., Martens, G., Mannens, E., de Walle, R. V., 2009. Personal content management system: A semantic approach. Journal of Visual Communication and Image Representation 20 (2), 131 – 144, special issue on Emerging Techniques for Multimedia Content Sharing, Search and Understanding.

[29] Rahman, M. M., Bhattacharya, P., Desai, B. C., 2009. A unified image retrieval framework on local visual and semantic concept-based feature spaces. Journal of Visual Communication and Image Representation 20 (7), 450–462.

[30] Robert, C., 1997. A decision-Theoretic Motivation. Springer-Verlag.

[31] Rui, X., Li, M., Li, Z., Ma, W.-Y., Yu, N., 2007. Bipartite graph reinforcement model for web image annotation. In: ACM MULTIMEDIA '07. pp. 585–594.

[32] Rui, Y., Huang, T. S., Chang, S.-F., 1999. Image retrieval: Current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation 10 (1), 39 – 62.

[33] Saber, E., Tekalp, A. M., 1997. Region-based shape matching for automatic image annotation and query-by-example. Journal of Visual Communication and Image Representation 8 (1), 3 – 20.

[34] Swain, M. J., Ballard, D. H., 1991. Color indexing. Int. J. Comput. Vision 7 (1), 11–32.

[35] Tabbone, S., Wendling, L., aug 2002. Technical symbols recognition using the two-dimensional radon transform. In: ICPR' 02. Vol. 2. pp. 200–203.

[36] Tollari, S., Mulhem, P., Ferecatu, M., Glotin, H., Detyniecki, M., Gallinari, P., Sahbi, H., Zhao, Z.-Q., 2008. A comparative study of diversity methods for hybrid text and image retrieval approaches. In: Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum.

[37] Torralba, A., Oliva, A., 2003. Statistics of natural image categories. In: Network: Computation in Neural Systems. pp. 391–412.

[38] Wang, C., Jing, F., Zhang, L., Zhang, H.-J., 2006. Image annotation refinement using random walk with restarts. In: ACM MULTIMEDIA '06. pp. 647–650.

[39] Yang, C., Dong, M., Hua, J., 2006. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: CVPR '06. pp. 2057–2063.

[40] Zhang, R., Zhang, Z. M., Li, M., Ma, W.-Y., Zhang, H.-J., 2005. A probabilistic semantic model for image annotation and multi-modal image retrieva. In: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV). pp. 846–851.

source of dependence	destination of dependence	semantic relation type
autumn	season	direct hypernym
bird	animal	inherited hypernym
buffalo	animal	inherited hypernym
beetle	animal	inherited hypernym
butterfly	animal	inherited hypernym
beach	sand	substance meronym
cow	animal	inherited hypernym
chicken	animal	inherited hypernym
chicken	bird	inherited hypernym
deer	animal	inherited hypernym
dog	animal	inherited hypernym
duck	animal	inherited hypernym
duck	bird	inherited hypernym
elephant	animal	inherited hypernym
flower	nature	inherited hypernym
goose	animal	inherited hypernym
goose	bird	inherited hypernym
horse	animal	inherited hypernym
leopard	animal	inherited hypernym
lion	animal	inherited hypernym
leaf	nature	inherited hypernym
monkey	animal	inherited hypernym
owl	animal	inherited hypernym
owl	bird	inherited hypernym
penguin	animal	inherited hypernym
penguin	bird	inherited hypernym
pigeon	animal	inherited hypernym
pigeon	bird	inherited hypernym
sheep	animal	inherited hypernym
seal	animal	inherited hypernym
swan	animal	inherited hypernym
swan	bird	inherited hypernym
springtime	season	direct hypernym
summer	season	direct hypernym
waterfall	water	inherited hypernym
dinosaur	animal	inherited hypernym

Table 2: Examples of dependence relations between keywords and types of dependences

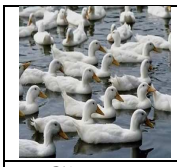
	water duck reflection flock		bird duck water close-up
Class : Hong Kong		Class : african birds	
	bird duck mallard baby		duck food cuisine meal
Class : waterfowl		Class : cuisine	

Table 3: Examples of images, with their class and their possible keywords

training part	C	S	KW	C + S	C + S + KW
25%	20.6	16.5	48.3	23.6	58.5
35%	22.8	16.8	54.5	24	59
50%	23.4	18.4	61.4	24.3	64.2
65%	24.1	19.1	62.4	26	65.6
75%	26	19.9	67.8	26.4	69.8
90%	26	24	69.2	28.8	76

Table 4: Mean recognition rates (in %) of our GM-B model without semantic relations - joint visual-textual classification vs. visual classification

Training part	GM-Mixture	GM-B	
		without	with SR
25%	61	58.5	68.7
35%	62.4	59	69.5
50%	67.2	64.2	76.2
65%	67.7	65.6	75.4
75%	72.2	69.8	80.4
90%	78.6	76	86

Table 5: Mean recognition rates (in %) of the GM-Mixture model vs. our GM-B model - joint visual-textual classification

Training part	standard deviation of difference	t-value
25%	0.33	22.96
35%	0.2	35.5
50%	0.095	94.3
65%	0.098	78.99
75%	0.16	50.86
90%	0.11	64.06

Table 6: Student's test for comparison of mean recognition rates of the GM-Mixture model vs. our GM-B model with SR in joint visual-textual classification

Training part	GM-Mixture	GM-B	
		without	with SR
25%	40	52	71
35%	56.2	72.6	78.9
50%	60	72.8	79.6
65%	61.7	77.1	79.7
75%	66	78.9	82.3
90%	68.7	79	82.4

Table 7: Mean rate (in %) of good annotations of the GM-Mixture model vs. our GM-B model - automatic annotation extension

Training part	standard deviation of difference	t-value
25%	0.58	53.32
35%	0.14	155.85
50%	0.14	138.47
65%	0.23	81.26
75%	0.2	77.71
90%	0.13	100.67

Table 8: Student's test for comparison of mean rates of good annotations of the GM-Mixture model vs. our GM-B model with SR in automatic annotation extension