Modeling, classifying and annotating weakly annotated images using Bayesian network

Sabine Barrat, Salvatore Tabbone

Modeling, classifying and annotating weakly annotated images using Bayesian network

Sabine Barrat and Salvatore Tabbone
LORIA-UMR7503 - University of Nancy 2
BP 239 - 54506 Vandœuvre-les-Nancy Cedex, France
{barrat,tabbone}@loria.fr

Abstract

We propose a probabilistic graphical model to represent weakly annotated images[1]*. 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 classification and automatic annotation of images. The experimental results, obtained from a database of more than* 30000 *images, by combining visual and textual information, show an improvement by* 50.5% *in terms of recognition rate against only visual information classication. Taking into account semantic relations between keywords improves the recognition rate by* 10.5% *and the mean rate of good annotations by* 6.9%*. The proposed method is experimentally competitive with the state-of-art classifiers.*

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 to fully annotated images. The second approach, called "content-based image retrieval" is a more young field. These methods rely on visual features (color, texture or shape) computed automatically, and retrieve images using a similarity measure.
In order to improve the recognition, a solution consists in combining visual and semantic information. Some researchers have already explored this possibility [1, 6]. Automatic image annotation 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 kind of multi-class image classification with a very large number of classes, as large as the vocabulary size. Many works have been proposed in this direction and we can cite, without being exhaustive, classification-based methods [5], probabilistic modeling-based methods [2] and annotation refinement [8].
The contribution of this paper is to propose a scheme for image classification optimization, by using a joint visual-text clustering approach and automatically extending image annotations. The model presented here is dedicated for both tasks: weakly-annotated image classification and annotation. In fact the classification methods before mentioned are efficient but they require that all images, or image regions are annotated. Moreover, most existing annotation models are not able to classify images. The proposed approach is derived from the probabilistic graphical model theory. We introduce a method to deal with missing data in the context of text annotated images as defined in [2, 6]. The uncertainty around the association between a set of keywords and an image is tackled by a joint probability distribution over a dictionary of keywords and the numerical features extracted from our collection of images (grey-level and color). The Gaussian-multinomial Mixture model [2] is the most related to our approach. However our model is less restrictive for the user. In fact, our classifier does not need that all images be annotated. Moreover, our model has the advantage to take into account the possible semantic relations between keywords, contrary to the model [2] which assumes that the keywords are independent given its parents.
Section 2 describes the probabilistic model of weakly-annotated image representation and how to use it to classify and extend existing annotations to images. Section 3 presents the experimental results. Finally, conclusions and future works are given in Section 4.

[1] 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

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 combines discrete and continuous variables and takes into account the problem of missing values. Let f_j be a query image characterized by a set of features F. F is composed of m visual features, denoted $v_1, ..., v_m$ and n possible keywords, denoted $KW\ 1, ..., KW\ n$. The chosen visual features are issued from one color descriptor, a color histogram, and one shape descriptor based on the Fourier/Radon transform. 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 big 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, 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 and associated keywords.

We have to manage continuous variables (corresponding to visual features) and discrete variables (corresponding to 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 weakly 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 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 can't belong to the given image annotation.

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 then characterized by n continuous variables. A supervised classification is considered then 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, $\theta_l = (\mu_l, \sum_l)$ the parameter of each Gaussian and $\Phi = (\pi_1, \pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ the global mixture parameter. Then 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 a probabilistic graphical model (see Figure 1), 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's 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. We have one GMM, 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 [7], 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, some edges are added to represent semantic relations between keywords of the vocabulary. For example, the keywords "dog" and "animal" are clearly dependent. In fact these two keywords belong to the same concept group. This dependence is represented by a directed edge from the node "dog" to the node "animal".
Then our classifier can be depicted by the Figure 1. 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. But, Figure 2 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".

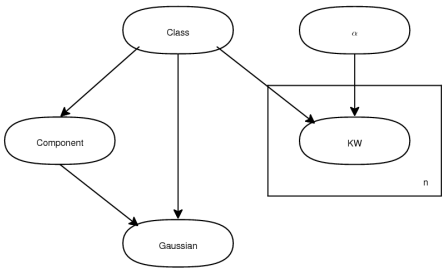

Figure 1. The Gaussian-Mixtures and Bernoulli mixture model

This Bayesian classifier means that each image and its keywords are assumed to have been generated conditional on the same class. Therefore the resulting multinomial 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.
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 $P(f_j) = 1$ when the network is evaluated. 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.

2.1 Annotation extension of images

Given an image without keyword, or a weakly annotated image, 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. For example, let us consider Table 1 which presents 2 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 probably due to the 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 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".

3 Experimental results

We present an evaluation of our model on more than 30000 weakly annotated images from the Corel image libraries and kindly provided by Vasconcelos and al. [3]. These images are split up into 306 classes. For example, 4 images of the class "arabian horses" are given in Figure 3.

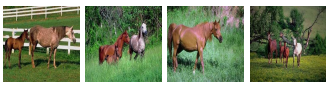

Figure 3. Examples of "arabian horse" class images

72% of the image database is 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.

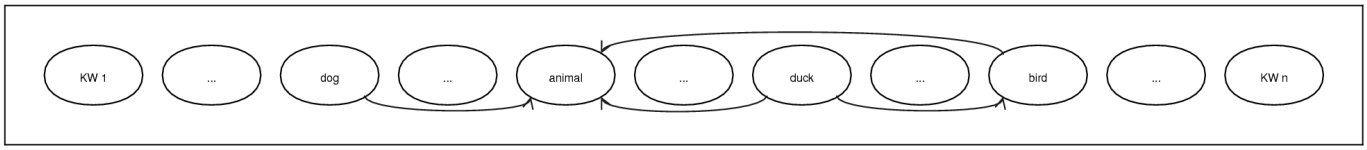

Figure 2. Dependences between keywords

image	initial possible keywords	keywords after annotation extension without taking into account semantic relations	keywords after annotation extension by using semantic relations
	bear black water	bear black water grass	bear black water grass
	bear black river	bear black river	bear black river water

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

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 [4]. 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 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.

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 hold for test set. In each case the tests are repeated 10 times in order that each database instance 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 the 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 training 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 2. Our GM-B 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.

training part	C	S	KW	C + S	C + S + KW
25%	20.6	16.5	48.3	23.6	68.7
35%	22.8	16.8	54.5	24	69.5
50%	23.4	18.4	61.4	24.3	76.2
65%	24.1	19.1	62.4	26	75.4
75%	26	19.9	67.8	26.4	80.4
90%	26	24	69.2	28.8	84

Table 2. Mean recognition rates (in %) of our GM-B model with semantic relations

Table 3 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 taking into account semantic relations between keywords improves the recognition by 10.5%. Moreover, Table 3 shows the effectiveness of our approach (GM-B model) compared to the Gaussian-multinomial mixture model (GM-Mixture) [2]. 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 to represent a joint clustering of an image and its

caption, in the GM-Mixture model, 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.

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 3. Mean recognition rates (in %) of the GM-Mixture model vs. our GM-B model

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. Then 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 database instance 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 4 compares the rate of good annotations obtained by taking into account semantic relations between keywords. We can observe that taking into account semantic relations between keywords improves the rate of good annotations by 6.9%. We can also see that our model is better than the GM-Mixture model, even if we do not take into account semantic relations between keywords.

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 4. Mean rate (in %) of good annotations of the GM-Mixture model vs. our GM-B model

4 Conclusion and future works

We have proposed a method for modeling, classifying and annotating weakly annotated images, which has the advantage to take into account semantic relations between annotations. Experimental results have demonstrated that semantic relation representation improves the recognition rate and the mean rate of good annotations. Moreover the evaluation has shown promising performance improvements with state-of-art classifiers. 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] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. 2003, matching words and pictures. *Journal of Machine Learning Research*, 3(6):1107–1135, 2003.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03*, pages 127–134, 2003.

[3] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29(3):394–410, 2007.

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

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

[6] M. L. Kherfi, D. Brahmi, and D. Ziou. Combining visual features with semantics for a more effective image retrieval. In *ICPR '04*, volume 2, pages 961–964, 2004.

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

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