Visual Features with Semantic Combination Using Bayesian Network for a More Effective Image Retrieval

Sabine Barrat, Salvatore Tabbone

▶ **To cite this version:**

Visual Features with Semantic Combination Using Bayesian Network for a More Effective Image Retrieval

Sabine Barrat and Salvatore Tabbone
LORIA-UMR 7503, University of Nancy 2,
BP 239, 54506 Vandœuvre-lès-Nancy, France
{barrat,tabbone}@loria.fr

Abstract

In many vision problems, instead of having fully annotated training data, it is easier to obtain just a subset of data with annotations, because it is less restrictive for the user. For this reason, in this paper, we consider especially the problem of weakly-annotated image retrieval, where just a small subset of the database is annotated with keywords. We present and evaluate a new method which improves the effectiveness of content-based image retrieval, by integrating semantic concepts extracted from text. Our model is inspired from the probabilistic graphical model theory: we propose a hierarchical mixture model which enables to handle missing values and to capture the user's preference by also considering a relevance feedback process. Results of visual-textual retrieval associated to a relevance feedback process, reported on a database of images collected from the Web, partially and manually annotated, show an improvement of about 44.5% in terms of recognition rate against content-based retrieval.

1 Introduction

We can distinguish two main techniques in image retrieval. The first one, called "text-based image retrieval", consists in applying text-retrieval techniques from fully annotated images. The second approach, called "content-based image retrieval" is a more young field and use a similarity measure (similarity of color, texture or shape) between a query image and an image of the used corpus. In order to improve the retrieval, a solution consists in combining visual and semantic informations. Some researchers have already explored this possibility [1]. In this perspective, the contribution of this paper is to propose a scheme for image retrieval optimization, by using a joint visual-text cluster-

ing approach and integrating the user in the loop of the retrieval. The proposed approach is derived from the probabilistic graphical model theory. We introduce a Bayesian network to deal with missing data in the context of text annotated images as defined in [2]. 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 numerical features extracted from our collection of grey-level and color images. Now Bayesian networks are a simple way to represent a joint probability distribution over a set of random variables, to visualize the conditional properties and they enable to compute complex operations like probability learning and inference, with graphical manipulations. Finally this method naturally allows the introduction of relevance feedback by putting the stress on user's responses, which can be reflected by the probabilities, during the learning procedure. Then a Bayesian network seems to be appropriate to represent and retrieve images by integrating the user's preference.

2 Bayesian network for image retrieval with relevance feedback

We present a hierarchical probabilistic model of multiple-type data (images and associated keywords) in order to retrieve in large annotated image databases. In first some definitions of the variables used to represent the distinct elements are needed. Let us consider the user's preference as two images that can be pointed out during the retrieval. The first, an image considered close from the query, will constitute a positive example for the relevance feedback process. And the other one, an image considered far from the query, will represent a negative example for the relevance feedback process. These two images will be represented by the same visual features than the query image. The visual features are considered as continuous variables, and the possible

associated keywords as discrete variables. Now 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 a mixture of Gaussian densities.

Let F be an image set composed of m instances $f_{1_i}, ..., f_{m_i}, \forall i \in \{1, ..., n\}$, where n is the dimension of the concatenated signatures provided by the descriptors on each image. Each instance $f_j, \forall j \in \{1, ..., m\}$ is then characterized by n continuous variables. 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 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. Now let us consider three latent discrete variables, each used to represent a virtual belonging class for the query image, the positive example and the negative example respectively. Then now, we have one Gaussian Mixture Model per class, for each latent class variable. This problem can be represented by the probabilistic graphical model in Figure 1, where:

- The "Class" node is a latent discrete node

- 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. The green circle is just used 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.

Thus the proposed model includes 3 instances of this graphical model: one for the positive examples (denoted GMMs 1), one for the negative examples (GMMs 2) and one for the query (GMMs 3). These submodel are encircled in green in the complete proposed model Figure 2 where:
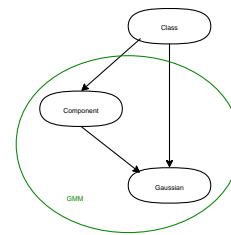

Figure 1. A Probabilistic graphical model as GMMs

- the nodes CL-PI, CL-NI and CL-QI correspond respectively to the latent "Class" variable of the positive example (PI), the negative example (NI) and the query (QI) images,

- the nodes C-M1, C-M2 and C-M3 correspond respectively to the latent "Component" variable of the Gaussian Mixture Models associated to the positive examples, the negative examples and the query images,

- the nodes G-M1, G-M2 and G-M3 correspond respectively to the continuous "Gaussian" variable of the Gaussian Mixtures Models associated to the positive examples, the negative examples and the query images.

Now the model can be completed by the discrete variables corresponding to the possible keywords associated to a query. These discrete variables, denoted $KW 1, ..., KW n$, are assumed to be distributed as a multinomial distribution over the vocabulary of keywords. Dirichlet priors [6], have been used for the probability estimation of the keyword variables. 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 query image visual features, the discrete variables corresponding to the keywords associated to the query images are included in the graphical model by connecting them to the latent query image class variable. At last, a latent discrete variable is needeed to represent the virtual belonging class of relevant images. The class of relevant images is assumed to have some dependencies with the other latent class variables. Then the variable representative of the relevant image class, denoted RI in Figure 2, is at the root of the graph and linked with the three other latent class variables. The latent variable "α" shows that a Dirichlet prior is used. The box around the variable KW denotes n repetitions

of KW, for each keyword. The subgraph $Query$, encircled in red, means that each image and its keywords are assumed to have been generated conditional on the same virtual class. Therefore the resulting multinomial and Gaussian mixture parameters should correspond: concretely if a query image, represented by visual descriptors, has an high probability under a virtual class, then its keywords should have an high probability under the same virtual class.
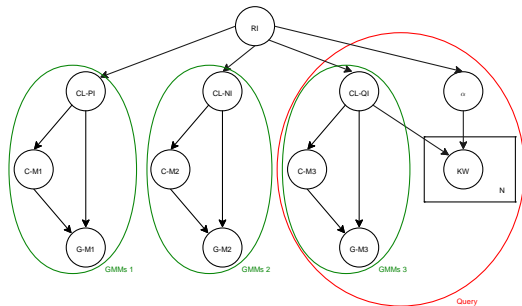

Figure 2. The proposed model

3   Parameter learning and inference

The EM algorithm has been used to learn the Gaussian mixture parameters. Our major problem deals with missing values. Indeed only some data are fully observed. It's the case of all visual features for color images or just shape features for grey-level images. On the contrary the color features for grey-level images, all visual features for relevance feedback images and especially some keywords for a large subset of images, are missing. Concerning the visual features, the missing values are clearly homogeneously distributed. The missing values are randomly distributed for the variables G-M1, G-M2 and KW i, $\forall i \in \{1, ..., n\}$. This kind of problem can be tackled with the EM algorithm too. The general purpose of this algorithm, explained in detail in [3] consists in computing, in an iterative way, the likelihood maximum when the instances can be viewed as incomplete data. An inference algorithm is also necessary to retrieve new images. Indeed, the inference process consists in computing posterior probability distributions of one or several other subsets of nodes. In the case of retrieval, the node of relevant image class, RI, is inferred. According to our Bayesian network topology, the inference process propagates the values from the image feature level represented by the "Gaussian" nodes G-M1, G-M2 and G-M3 through the "Component" nodes C-M1, C-M2, C-M3 and the Keyword nodes KW $i, \forall i \in \{1, ..., n\}$, then by the "Class" node level composed of CL-PI, CL-NI and CL-QI until

the root node RI. A message passing algorithm [5] is applied to the network. Thus a query image f_j, its possible keywords KW i and the two possible relevance feedback examples f_k and f_l, will be considered as an "evidence" represented by a joint probability equal to 1, when the network will be evaluated. Thanks to the inference algorithm, the probabilities of each node are updated in function of this evidence. After the belief propagation, we know the posterior probability $P(RI|f_j, f_k, f_l, KW)$. The highest probability determines the cluster of relevant images. Finally, the algorithm of retrieval can be decomposed in 3 steps. In first the class, denoted C, of relevant images is infered from the query image, the positive example and the negative example. Then the inference of the RI node is repeated on the rest of the database images, in order to obtain a clustering. At last, the k most relevant images corresponding to the query are the ones with the k highest probabilities of belonging to the cluster C.

4   Experimental results

In this section, we present an evaluation of our model on more than 3000 free images collected from the Web, and kindly provided by Kherfi et al. [4]. These images have been manually classified into 16 classes. The class number has been arbitrarily chosen, in function of the data. Each class contains 230 images. For example, 4 images of the class "horse" are presented in Figure 3.

Figure 3. Examples of horse-class images

65% of the image database have been manually annotated by 1 keyword, 28% by 2 keywords and 6% by 3 keywords, using a vocabulary set of 39 keywords. For example, among the 4 images from the Figure 3, the first image is annotated by the 2 keywords "animal" and "horse". The second is only annotated by one keyword: "animal". The others have not been annotated. The chosen visual features are issued from one color descriptor, a color histogram, and one shape descriptor based on the Fourier/Radon transforms. The experiments have been done by considering successively each database image as the query example. For each query example, retrieval results are return as a ranked list of 30 images. This number has been chosen for reasons of adaptability to the user interface size. The images of the same class as the query example are considered as relevant images. Positive feedback process has been running as

VF	VF + KW	VF + KW + PRF		VF + KW + RF	
		1 iteration	10 iterations	1 iteration	10 iterations
44.3	70.7	79.3	82.4	86.7	91

Table 1. Retrieval average precisions (P in %)

follows: one image from the same class as the query example, randomly chosen, is assigned at the positive example. Concerning negative feedback process, one irrelevant image, randomly chosen too, is assigned as negative example. The relevance feedback process has been tested for 1 and 10 iterations. The average precision, P, used in Table 1, is expressed in %. Let n be the size of the database. And $QI_j, \forall j \in \{1, ..., n\}$, a query image. The $k, \forall k \in \{1, ..., 30\}$ retrieval precisions of the image QI_j are defined by:

$$P_{j_k} = \frac{\text{\# relevant images in the k first images of the list}}{k}$$

and $P = \frac{\sum_{j=1}^{n} \frac{\sum_{k=1}^{30} P_{j_k}}{30}}{n} \times 100$. Let us consider Table 1. The notation "VF" means only visual features have been used. The notation "KW" indicates that textual information has been used. Then the notations "PRF" and "RF" indicate respectively the use of a relevance feedback process with positive examples only, and with positive and negative examples. The results confirm that combining visual with semantic features and a relevance feedback approach improves the retrieval precision. In fact we observe that the combination of visual features and possible keywords (denoted case 2) increases the retrieval precision of 26.4% on average compared to the content-based retrieval (case 1). Moreover we can notice the relevance feedback approach with positive examples only (case 3) improves the retrieval precision of 10.1% on average. The addition of negative examples outperforms the precision of 8%. To sum up an improvement of 44.5% on average has been obtained by using our visual-textual retrieval associated to the relevance feedback approach with positive and negative examples (denoted case 4), compared to the content-based retrieval. Now let us consider Figure 4, representing the average precisions in function of the retrieve list rank k, for the 4 cases before mentioned. The red curve, always lower than the 3 others, shows the robustness of visual-textual retrieval. Then we remark than the red and green curves, obtained without feedback, continuously decrease. On the contrary, the 2 others curbs increase from about 25 images, thanks to the feedback. Finally, the pink curve, always higher than the blue one, shows the efficiency of negative example addition in the relevance feedback process.
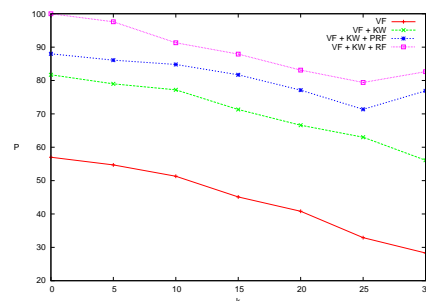

Figure 4. Precision P function of rank k

5 Conclusion and future works

We have proposed an efficient model which enables to combine visual and textual information, to handle missing values and capture the user's preference. We have done our experiments on a partially annotated Web image database. The results show that our visual-textual retrieval associated to a relevance feedback method improves the retrieval precision.

References

[1] A. Benitez and C. Shih-Fu. Perceptual knowledge construction from annotated image collections. *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, 1:189–192, 2002.

[2] D. M. Blei and M. 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 Press.

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

[4] M. L. Kherfi, D. Brahmi, and D. Ziou. Combining visual features with semantics for a more effective image retrieval. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on*, volume 2, pages 961–964. IEEE Computer Society, 2004.

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

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