J. Vis. Commun. Image R. 21 (2010) 355–363

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Modeling, classifying and annotating weakly annotated images using Bayesian network

Sabine Barrat *, Salvatore Tabbone

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

ARTICLE INFO

Article history:
Received 31 March 2009
Accepted 18 February 2010
Available online 26 February 2010

Keywords:
Probabilistic graphical models
Bayesian networks
Image classification
Image annotation
Semantic similarity
Wordnet
Visual features
Bayesian classifier

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 30,000 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.

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 pre-

* Corresponding author.
E-mail address: barrat@loria.fr (S. Barrat).

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.02.010

dict 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 intervention. 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 *30,000* 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 $v_1, \ldots, v_m$,
- n possible keywords, denoted $KW1, \ldots, KWn$.

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