Pattern Recognition 47 (2014) 1361–1374

Contents lists available at ScienceDirect

Pattern Recognition

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

Support vector description of clusters for content-based image annotation

Liang Sun [a], Hongwei Ge [a], Shinichi Yoshida [c], Yanchun Liang [b,*], Guozhen Tan [a]

[a] College of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[b] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[c] School of Information, Kochi University of Technology, Kochi 782-8502, Japan

A R T I C L E I N F O

Article history:
Received 23 November 2012
Received in revised form
7 September 2013
Accepted 13 October 2013
Available online 19 October 2013

Keywords:
Support vector clustering
Kernel method
Image retrieval
Content-based image annotation
Computer vision

A B S T R A C T

Continual progress in the fields of computer vision and machine learning has provided opportunities to develop automatic tools for tagging images; this facilitates searching and retrieving. However, due to the complexity of real-world image systems, effective and efficient image annotation is still a challenging problem. In this paper, we present an annotation technique based on the use of image content and word correlations. Clusters of images with manually tagged words are used as training instances. Images within each cluster are modeled using a kernel method, in which the image vectors are mapped to a higher-dimensional space and the vectors identified as support vectors are used to describe the cluster. To measure the extent of the association between an image and a model described by support vectors, the distance from the image to the model is computed. A closer distance indicates a stronger association. Moreover, word-to-word correlations are also considered in the annotation framework. To tag an image, the system predicts the annotation words by using the distances from the image to the models and the word-to-word correlations in a unified probabilistic framework. Simulated experiments were conducted on three benchmark image data sets. The results demonstrate the performance of the proposed technique, and compare it to the performance of other recently reported techniques.

1. Introduction

The number of image archives on the Internet is growing rapidly with the proliferation of user-contributed images. Thus, searching for images that match a user query presents a significant challenge. Popular search engines, such as Google and Yahoo!, rely mainly on textual descriptions contained in the filenames or keywords. Many search engines do not consider the content of images and are able to search only the annotations. On the other hand, in real-world image systems, many images are created without direct annotations of semantic content, which limits the ability to search by text-based engines. This creates the need for content-based image retrieval (CBIR) [1–3]. Research on CBIR has attracted the attention of researchers in various fields, including computer vision and machine learning. In a CBIR system, instead of using text descriptions, searches are based on low-level features such as color, texture, and shape. Some aspects of CBIR have been shown to be successful, and some progress has been made [4–9]. However, it still suffers from the "semantic gap" problem, which arises because the low-level features of an image are not sufficient to encapsulate the high-level semantics [1,10].

One natural way to mitigate the semantic gap problem is to assign tags to images. Appropriate tagging can help to increase the efficiency of retrieval. However, manual tagging is tedious and labor intensive [11–13], so there has been a surge of interest in developing automatic or semi-automatic tagging based on the low-level visual contents of an image [11–27]. These methods are referred to as content-based image annotation (CBIA). In greater detail, the methods presented in [13–17] are based on multiple classifiers. They partition the images into different classes, and then they assign to each class a distinct topic of interest and a set of descriptive words. The system treats the annotation of an untagged image as a classification problem, and it selects the relevant annotation based on the classification results. The methods presented in [18–26] are probabilistic modeling methods, which are also referred to as generative modeling methods. They use statistical tools to determine the correlations between images and annotations so that they can compute the joint probability that an untagged image is labeled with a particular word. In the annotating process, the relevant words are selected by a graph-based technique [23,24] or by a data fusion and aggregation technique [18–22,25,26]. More recently, the development of image platforms on the Internet, e.g., Flickr [28], Alipr [19], and PhotoStuff [29], has enabled users to

* Corresponding author. Tel.: +86 43185153829; fax: +86 43185168752.
E-mail addresses: ycliang@jlu.edu.cn, liangyanchun@gmail.com (Y. Liang).

0031-3203/$ - see front matter
http://dx.doi.org/10.1016/j.patcog.2013.10.015

annotate images and give feedback on annotations. This provides opportunities to develop automatic annotating methods that use the information provided by users and the existing search results. The methods presented in [11,13], and [27] fall into this category.

Many of the above methods require substantial machine learning techniques to fill the gap between the low-level visual contents of an image and the high-level semantics. Among machine learning techniques, support vector clustering (SVC) [30–32], is a recently developed algorithm that was inspired by the support vector machine (SVM) [33]. The SVC maps data points from the original space to a higher-dimensional space by means of a Gaussian kernel and then seeks the smallest sphere that encloses most of the data points. The sphere is mapped back to the original space as several arbitrarily shaped contours, each of which encloses a subset of the data points. SVC has many advantages over other algorithms, including its ability to determine, without prior knowledge, the topological structure of a system, to delineate the boundaries of irregularly shaped clusters, and to deal with outliers by employing a soft-margin constant [34,35]. In real-world systems, the same descriptions may apply to a wide variety of images. For example, if an image is tagged with “historical building”, then it might be a picture taken in the desert, near the beach, or in the city. There is not a clear way in which these images should be organized. Since SVC is able to delineate irregular cluster boundaries, it can be used to develop unified models to describe unorganized images.

In view of the above, in this paper, we present a novel algorithm for CBIA. In this work, images are represented by colored pattern appearance models (CPAM) [36]. The system has two major components, the training process and the tagging process. In the training process, clusters of images with manually tagged words are used as training instances. For each cluster, the feature vectors of the image are mapped to a higher-dimensional space, and the support vectors are used to describe the cluster. Since the mapping process is the same as that in SVC and its objective is to build models that are described by support vectors for image clusters, we call it the support-vector-based method for image annotation (SVIA). In order to tag an image, the method computes the distances between the image and the models that are described by support vectors. A shorter distance indicates a stronger association between the image and the model, and thus indicates a stronger association between the image and the corresponding words of the model. Moreover, the word-to-word correlations contain rich information about the meanings of the images. For example, if an image is tagged with “France”, then it will have a higher probability of being tagged with “Europe”, and if an image is tagged with “indoor”, then it will have a lower probability of being tagged with “grass”. Therefore, the word-to-word correlations are also considered in the annotation framework. When tagging an image, the system predicts the appropriate words by using the distances from the image to the models and the word-to-word correlations in a probabilistic framework.

The remainder of this paper is organized as follows. Section 2 reviews some of the previous work on CBIA; in particular, the probabilistic modeling methods and the SVC algorithm. Section 3 presents the support-vector-based modeling method. Section 4 presents the proposed probabilistic modeling method for the assignment of annotations. Section 5 presents and compares simulated result, followed by concluding remarks in Section 6.

2. Related works

This work is related to probabilistic modeling methods, and the support-vector-based model is related to SVC. In this section, we begin by reviewing the basic concepts and some prevailing methods of the probabilistic modeling approaches to CBIA. Next, we review the basic concepts of SVC and some of its variants.

2.1. Probabilistic modeling approaches to content-based image annotation

2.1.1. Probabilistic models

The probabilistic modeling method computes, for each work, the joint probability that an untagged image is tagged with that word. Given an untagged image I_q, the main objective is to find a group of words w^* in a given vocabulary $\mathcal{W}$ so that the conditional distributions $p(w|I_q)$ are maximized as follows:

$$w^* = \arg\max_{w \subset \mathcal{W}} p(w|I_q). \tag{1}$$

Generally, there are two types of probabilistic models for image annotation, i.e., the two-layer model and the three-layer model. The two-layer model generates words directly from the given image, as shown in Fig. 1(a). By applying Bayes' rule to Eq. (1), we obtain the following formulation:

$$w^* = \arg\max_{w \subset \mathcal{W}} p(w|I_q) = \arg\max_{w \subset \mathcal{W}} \sum_{I_i \in \mathcal{T}} p(w|I_i)p(I_q|I_i)p(I_i), \tag{2}$$

where $\mathcal{T}$ is the training image set, I_i is the ith training image in $\mathcal{T}$, $p(w|I_i)$ is the probability that image I_i is correlated with word w, $p(I_q|I_i)$ is the probability that image I_i is relevant (or similar) to image I_q, and $p(I_i)$ is the prior probability of the training image I_i. The three-layer model, however, introduces a hidden layer of “topics”, and an image is then represented by a mixture of topics. Words are then generated from these topics, as shown in Fig. 1(b). By applying Bayes' rule to Eq. (1), we obtain the following formulation:

$$\begin{aligned} w^* &= \arg\max_{w \subset \mathcal{W}} p(w|I_q) \\ &= \arg\max_{w \subset V} \sum_{I_i \in \mathcal{T}} \left\{ \left(\sum_{t_j \in S} p(w|t_j)p(t_j|I_i) \right) \times p(I_q|I_i)p(I_i) \right\}, \end{aligned} \tag{3}$$

where S is the set of topics, t_j is the jth topic in S, $p(w|t_j)$ is the probability that topic t_j is correlated with word w, $p(t_j|I_i)$ is the probability that image I_i is correlated with t_j, and $p(I_q|I_i)$ and $p(I_i)$ have the same meanings as in Eq. (2). In addition, since the word-to-word correlations contain rich information about the meanings of the images, some of the methods integrate the word-to-word correlation $p(w_i|w_j)$ into the two-layer model or the three-layer model to maintain semantic consistence.

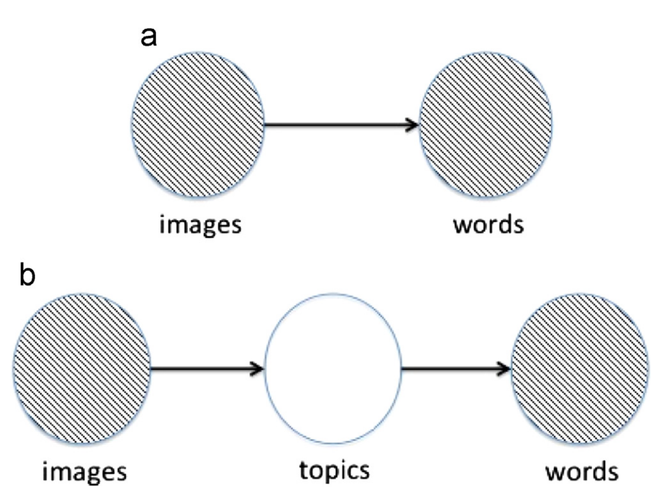

Fig. 1. Schematic diagrams of probabilistic models for image annotation. (a) Two-layer model and (b) three-layer model.