J. Vis. Commun. Image R. 21 (2010) 798–805

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Fusing semantic aspects for image annotation and retrieval

Zhixin Li [a,b,*], Zhiping Shi [a,c], Xi Liu [a], Zhiqing Li [a], Zhongzhi Shi [a]

[a] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China
[c] Information Engineering College, Capital Normal University, Beijing 100048, China

ARTICLE INFO

Article history:
Received 19 October 2009
Available online 30 June 2010

Keywords:
Automatic image annotation
Latent aspect model
PLSA
Adaptive asymmetric learning
Image retrieval
Semantic gap
Visual feature
Textual word

ABSTRACT

In this paper, we present an approach based on probabilistic latent semantic analysis (PLSA) to achieve the task of automatic image annotation and retrieval. In order to model training data precisely, each image is represented as a bag of visual words. Then a probabilistic framework is designed to capture semantic aspects from visual and textual modalities, respectively. Furthermore, an adaptive asymmetric learning algorithm is proposed to fuse these aspects. For each image document, the aspect distributions of different modalities are fused by multiplying different weights, which are determined by the visual representations of images. Consequently, the probabilistic framework can predict semantic annotation precisely for unseen images because it associates visual and textual modalities properly. We compare our approach with several state-of-the-art approaches on a standard Corel dataset. The experimental results show that our approach performs more effectively and accurately.

1. Introduction

With the development of digital imaging and data storage, searching and indexing large image databases efficiently and effectively has become a challenging problem. In order to solve the problem, there exist two distinct approaches in the literature. One solution is to annotate each image manually with keywords or captions and then search images using a conventional text search engine. This technique uses text to capture semantic content of images and allows query by text. However, expensive labor makes this solution difficult to be extended to large image databases. The other solution is to query by visual example. Under this paradigm, various low-level visual features are extracted from each image in the database and image retrieval is formulated as searching for the best database match to the feature vector extracted from the query image. Although this process is accomplished quickly and automatically, the retrieval results are usually semantically irrelevant to the query example due to the notorious *semantic gap* [1]. As a result, automatic image annotation has emerged as a striking and crucial problem for semantic image retrieval [2].

As a latent aspect model, PLSA has been applied in many research areas of computer vision, such as object recognition and scene classification. Furthermore, many approaches based on PLSA successfully achieve the task of automatic image annotation [3–5]. However, most of these approaches learn the latent space from either two modalities (visual features or textual words) equivalently or one modality only. In this paper, we present an extended model to fuse aspects learned from both visual and textual modalities asymmetrically. In addition, an adaptive learning approach is proposed to fit the model. From the theoretical perspective, the proposed probabilistic framework is similar to PLSA-WORDS [3]. The proposed learning approach, however, is quite different from theirs. First, when constructing latent space, PLSA-WORDS uses aspects of one PLSA model to learn the semantic information from textual modality, while our approach employs two sets of aspects to learn the semantic information from visual and textual modalities respectively. More important, the learning process of PLSA-WORDS is relatively static. In contrast, our approach learns semantic information in an adaptive mode. That is, it fuses two sets of aspects with different weights which are determined by the visual representations of images in training data set.

The main contributions of this work are the following. Firstly, we present a probabilistic framework based on PLSA to achieve the task of automatic image annotation and retrieval. The framework constructs the latent space by using two PLSA models to capture semantic aspects from visual and textual modalities respectively. Secondly, an adaptive asymmetric learning algorithm is proposed to fuse the

[*] Corresponding author at: Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. Fax: +86 10 82610254.

E-mail addresses: lizx@ics.ict.ac.cn (Z. Li), shizp@ics.ict.ac.cn (Z. Shi), liux@ics.ict.ac.cn (X. Liu), lizq@ics.ict.ac.cn (Z. Li), shizz@ics.ict.ac.cn (Z. Shi).

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.06.004

aspects of these two models. For each image document, the aspect distributions of different modalities are fused by multiplying different weights. Finally, we test the performance of our approach using a standard Corel dataset which consists of 5000 images. In comparison with several state-of-the-art approaches, our approach achieve higher annotation accuracy and superior retrieval effect.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 presents PLSA model and its principles. Section 4 describes image representation and proposes the adaptive asymmetric approach to learn the correlation between visual features and textual words. Furthermore, this section gives the training, annotating and retrieving algorithms. Experimental results are reported and analyzed in Section 5. Finally, the overall conclusions of this work are presented in Section 6.

2. Related work

Various approaches have been proposed for semantic image annotation and retrieval. The state-of-the-art techniques can be roughly categorized into two different schools of thought.

The first one defines auto-annotation as a traditional supervised classification problem, which treats each word (or semantic category) as an independent class and creates a different class model for every word (or semantic category). This approach separates the textual components from the visual components, computing similarity at the visual level. Then it annotates a new image by propagating the corresponding class words. A representative work is automatic linguistic indexing of pictures (ALIP) proposed by Li and Wang [6]. ALIP uses two-dimensional multiresolution hidden Markov models (2D MHMMs) to capture spatial dependencies of visual features of given semantic categories. Besides, the content-based soft annotation (CBSA) system proposed by Chang et al. [7] is based on binary classifiers (BPMs and SVMs) trained for each word and it indexes a new image with the output of each classifier. Caneiro et al. [8] propose supervised multiclass labeling (SML), which employs optimal principle of minimum probability of error and treats annotation as a multiclass classification problem where each of the semantic concepts of interest defines an image class. At annotation stage, these classes all directly compete for the image to annotate. Therefore, this approach no longer suffers a sequence of independent binary tests.

The second perspective takes a different stand and treats images and texts as equivalent data. It attempts to discover the correlation between visual features and textual words on an unsupervised basis, by estimating the joint distribution of features and words and posing annotation as statistical inference in a graphical model. Mori et al. [9] propose co-occurrence model which collects the co-occurrence counts between words and features and uses them to predict annotated words for unseen images. Duygulu et al. [10] improve the co-occurrence model by utilizing machine translation models, in which the words and blobs are considered as two equivalent languages. After training, the translation model can translate blobs into words, that is, it can attach words to a new image region. Barnard et al. [11] discuss several models to represent the joint distribution of words and blobs. Once the joint distribution has been learned, the annotation problem is converted into a likelihood problem relating blobs to words. However, the performance of these models is strongly affected by the quality of image segmentation. Similarly, Blei et al. [12] employ correspondence latent Dirichlet allocation (LDA) model [13] to build a language-based correspondence between words and images. The model can be viewed in terms of a generative process that first generates the region descriptions and subsequently generates the caption words. Afterwards, Monay et al. [3] propose a new way of modeling multi-modal co-occurrences. This approach constrains the definition of latent space to ensure its consistency in semantic terms (words), while retaining the ability to jointly model visual information. In addition to this, Jeon et al. [14] propose cross-media relevance models (CMRM) to annotate image, assuming that the blobs and words are mutually independent given a specific image. Lavrenko et al. [15] propose similar continuous-space relevance model (CRM), in which the word probabilities are estimated using multinomial distribution and the blob feature probabilities using a non-parametric kernel density estimate. Compared with CMRM, CRM directly models continuous feature, therefore it does not rely on clustering and consequently does not suffer from the granularity issues. Feng et al. [16] propose multiple Bernouli relevance model (MBRM), in which a multiple Bernoulli distribution is used to generate words instead of the multinomial one as in CRM.

3. PLSA model

Although the LDA model [13] has been shown to improve over PLSA [17] in terms of perplexity in text collections, we still choose PLSA to construct our model for two main reasons. First, PLSA allows for an exact EM algorithm. This makes the intended modifications of learning procedure easier. Second, PLSA has been shown to perform well on image classification tasks [18,19], using the aspect mixture proportions to learn the classifiers.

PLSA [17] is a statistical latent aspect model for co-occurrence data which associates an unobserved class variable with each observation. The model can be fitted to a training set through an Expectation–Maximization (EM) based iterative algorithm.

3.1. The aspect model

PLSA model introduces a hidden variable z_k ($k \in 1,\ldots,K$) in the generative process of each element x_j ($j \in 1,\ldots,M$) in a document d_i ($i \in 1,\ldots,N$). Given this unobservable variable (latent aspect) z_k, each occurrence x_j is independent of the document it belongs to, which corresponds to the following joint probability: $P(d_i,z_k,x_j) = P(d_i)P(z_k|d_i)P(x_j|z_k)$. The joint probability of the observed variables is obtained by marginalizing over the latent aspect z_k,

$$P(d_i,x_j) = P(d_i)\sum_{k=1}^{K} P(z_k|d_i)P(x_j|z_k). \tag{1}$$

A representation of the aspect model in terms of a graphical model is depicted in Fig. 1(a). Since the cardinality of the latent aspects is typically smaller than the number of documents (and elements) in the collection, $K \ll min\{N,M\}$, it acts as a bottleneck variable in predicting words.

The model (1) expresses each document as a convex combination of K aspect vectors. This amounts to matrix decomposition as shown in Fig. 1(b). Essentially, each document is modeled as a mixture of aspects — the histogram for a particular document

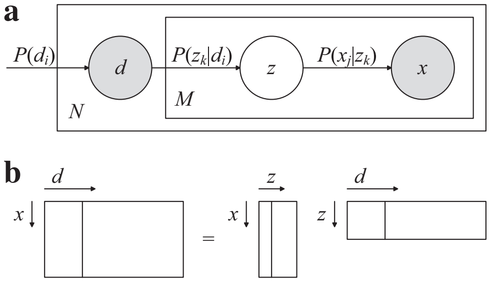

Fig. 1. (a) Graphical model representation of PLSA. (b) Matrix decomposition of conditional distribution.