Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Automatic image annotation based on Gaussian mixture model considering cross-modal correlations

Dongping Tian [a,*], Zhongzhi Shi [b]

[a] Institute of Computer Software, Baoji University of Arts and Sciences, Baoji, Shaanxi 721007, PR China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 21 August 2016
Revised 8 December 2016
Accepted 11 January 2017
Available online 12 January 2017

Keywords:
Automatic image annotation
Gaussian mixture model
RPEM
Semantic correlation
Max-bisection
Image retrieval

ABSTRACT

Automatic image annotation has been an active topic of research in the field of computer vision and pattern recognition for decades. In this paper, we present a new method for automatic image annotation based on Gaussian mixture model (GMM) considering cross-modal correlations. To be specific, we first employ GMM fitted by the rival penalized expectation-maximization (RPEM) algorithm to estimate the posterior probabilities of each annotation keyword. Next, a label similarity graph is constructed by a weighted linear combination of label similarity and visual similarity by seamlessly integrating the information from both image low level visual features and high level semantic concepts together, which can effectively avoid the phenomenon that different images with the same candidate annotations would obtain the same refinement results. Followed by the rank-two relaxation heuristics over the built label similarity graph is applied to further mine the correlation of the candidate annotations so as to capture the refining annotation results, which plays a crucial role in the semantic based image retrieval. The main contributions of this work can be summarized as follows: (1) Exploiting GMM that is trained by the RPEM algorithm to capture the initial semantic annotations of images. (2) The label similarity graph is constructed by a weighted linear combination of label similarity and visual similarity of images associated with the corresponding labels. (3) Refining the candidate set of annotations generated by the GMM through solving the max-bisection based on the rank-two relaxation algorithm over the weighted label graph. Compared to the current competitive model SGMM-RW, we can achieve significant improvements of 4% and 5% in precision, 6% and 9% in recall on the Corel5k and Mirflickr25k, respectively.

1. Introduction

With the prevalence of digital imaging devices such as webcams, phone cameras and digital cameras, the number of accessible images is growing at an exponential speed. As a result, how to make the best use of these resources becomes an emerging and important problem. Content-based image retrieval has been studied and explored over the last few decades, which heavily depends on the low-level features to find images relevant to the query concept and is represented by the query examples provided by the user. However, its performance is far from satisfactory due to the well-known gap between visual features and semantic concepts, i.e., images of dissimilar semantic content may share some common low-level features while images of similar semantic content may be scattered in the feature space. In general, people prefer to query images by semantic keywords rather than their low-level features, such as color, texture and shape, etc. However, manual annotation is time-consuming and labor-intensive, especially it cannot scale well when the volume of data is large and cannot keep up with the rapid growth of available image data anymore, which results in automatic image annotation (AIA) emerging as a striking and crucial problem in the area of computer vision. From the literature, it can be clearly observed that the state-of-the-art research on AIA can be roughly divided into two types. The first one poses image annotation as a supervised classification problem (referred to as discriminative model) like SML [5] and ALIP [25], which treats each semantic concept as an independent class and constructs different classifiers for different concepts. To be more specific, such kind of approaches predicts the annotations of a new image by calculating similarity at the visual level and propagating the corresponding words subsequently. In contrast, the second category treats the words and visual tokens in each image as equivalent features in different modalities (referred to as generative model) [15,29]. Image annotation is thereafter formalized by modeling

* Corresponding author at: Institute of Computer Software, Baoji University of Arts and Sciences, No. 44, Baoguang Road, Weibin District, Baoji, Shaanxi 721007, PR China.

E-mail addresses: tdp211@163.com, tiandp@ics.ict.ac.cn (D. Tian).

http://dx.doi.org/10.1016/j.jvcir.2017.01.015
1047-3203/© 2017 Elsevier Inc. All rights reserved.

the joint distribution of visual and textual features on the training data and predicting the missing textual features for a new image. By comparison, the former method is relatively direct and natural to be understood. However, its performance is limited with the increase of the number of the semantic concepts and explosive multimedia data on the web. On the other hand, the latter often requires large-scale parameters to be estimated and the accuracy is strongly affected by the quantity and quality of the training data available.

The remainder of this paper is organized as follows. Section 2 summarizes some related work from two respects of latent aspect models and GMM in the field of image annotation and retrieval. In Section 3, the proposed automatic image annotation framework is elaborated, including GMM and its parameter estimation, construction of the weighted label similarity graph and max-bisection based on the rank-two relaxation heuristics for precise image annotation, respectively. Section 4 reports and analyzes experimental results on two benchmark image datasets. Finally, we end this paper with some concluding remarks and future work in Section 5.

2. Related work

In recent years, a huge amount of automatic image annotation methods have been proposed in the literature. The early notable work includes the translation model [12] which treats AIA as a process of translation from a set of blob tokens to a set of words. The cross-media relevance model (CMRM) [21] that assumes the blobs and words are mutually independent given a specific image. Subsequently CMRM is improved through continuous space relevance model (CRM) [35], multiple Bernoulli relevance model (MBRM) [15] and dual cross-media relevance model (DCMRM) [28], etc. It is to be noted that the latent aspect models, such as probabilistic latent semantic analysis (PLSA) [16], latent Dirichlet allocation (LDA) [2] and correlated topic model (CTM) [3], has been an active topic of research in computer vision due to its potentially large impact on both image understanding and web image search. In the context of PLSA, Monay et al. present a series of PLSA models for automatic image annotation [36–38], among which PLSA-MIXED [36] learns a standard PLSA based on a concatenated representation of the textual and visual features, while PLSA-WORDS or PLSA-FEATURES [37,38] allows modeling of an image as a mixture of latent aspects that is defined either by its textual captions or by its visual features for which the conditional distributions are estimated from one of the two modalities only. Subsequent work [27] brings forward PLSA-FUSION for AIA, which utilizes two linked PLSA models to learn the mixture of the aspects in an adaptive mode from both visual and textual modalities, respectively. It sounds rational in theory due to this model allows for the visual and textual modalities simultaneously when constructing the latent space. More recent work [46] exploits PLSA to formulate the potential functions of conditional random field for AIA. In order to extract effective features to fully reflect the intrinsic content of images, a multi-feature PLSA (MF-PLSA) is put forward to circumvent this problem by combining low-level visual features for image region annotation in that it handles data from two different visual feature domains [67]. In [17], Hong et al. present a multiple-instance learning method with discriminative feature mapping and feature selection for AIA, which can explore both the positive and negative concept correlations as well as select the effective features from a large and diverse set of low-level features for each concept. Especially in literature [45], a unified two-stage refining image annotation framework is proposed by integrating PLSA with random walk. Conducted experiments reveal that this approach outperforms the baselines regarding their effectiveness and efficiency in the task of automatic image annotation and retrieval. Besides, Dong et al. [11] model the dependences of adjacent shearlet subbands using linear regression for texture classification and retrieval. Xu et al. [60] come up with a regularized LDA for tag refinement that facilitates the topic modeling by using both the statistics of tags and visual affinities of images in the corpus.

Gaussian mixture model, as another kind of supervised learning method, has been extensively applied in machine learning and pattern recognition communities. As the representative work of GMM for image annotation, Yang et al. [66] formulate AIA as a supervised multi-class labeling problem by applying color and texture features to form two separate vectors, for which two independent GMMs are estimated from the training set as the class densities by means of the EM algorithm in conjunction with a denoising technique. In [55], an effective visual vocabulary is constructed by using hierarchical GMM instead of the traditional clustering methods. Meanwhile, PLSA is utilized to explore semantic aspects of visual concepts and discover the topic clusters among documents and visual words so that each image can be projected on to a lower dimensional topic space for more efficient annotation. In addition, the conventional GMM is adapted to a global one estimated by all patches from training images along with an image-specific GMM obtained by adapting the mean vectors of the global one [56]. Afterwards GMM is embedded into the max-min posterior pseudo-probabilities for AIA [57], in which the concept-specific visual vocabularies are generated by assuming that the localized features of images with a specific concept satisfy the distribution of Gaussian mixture model. It is generally believed that the spatial relationships between objects are more useful and stable than the position information. Based on this argument, the recent work [32] utilizes GMM for region-based image annotation by taking full account of region-based color and coordinate of matching. To be specific, this method first partitions image into disjoint, connected regions with color features and x-y coordinate while a training dataset is modeled through the GMM to have a stable annotation results in the later phase. Note that in our previous work [44], a semi-supervised learning is developed based on GMM and random walk as well as the transductive support vector machine for automatic image annotation. Alternatively, a specific GMM is constructed for image clustering and retrieval in [41]. In particular, each cluster of data, modeled as a GMM into an input space, is interpreted as a hyperplane in a high dimensional mapping space where the underlying coefficients are found by solving a quadratic programming problem. In [33], GMM is exploited to work on color histograms built with weights delivered by the bilateral filter scheme, which enables the retrieval system not only to consider the global distribution of the color image pixels but also to take into account their spatial arrangement. In recent work [42], a multilayer PLSA is formulated for image retrieval, which can effectively eliminate the noisiest words generated in the process of the vocabulary construction. In the meanwhile, the edge context descriptor is extracted by GMM and a spatial weighting scheme is constructed based on GMM to reflect the information about the spatial structure of the images. At the same time, a generalized Gaussian mixture model is developed for content based image retrieval [39]. The GMM-cluster forest method is formulated to support multi-features based similarity search in high-dimensional spaces [53]. Beyond this, Gaussian mixture model has also been extensively applied in a variety of computer vision tasks [1,6,10,20,58] such as image classification, object detection and scene understanding. Table 1 presents a comparison of several classic AIA methods based on generative and discriminative models, including their classification, basic characteristics and merits and demerits.

As discussed earlier, most of these approaches can achieve promising performance and motivate us to explore better auto-