Expert Systems with Applications 36 (2009) 9813–9818

Contents lists available at ScienceDirect

Expert Systems with Applications

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

TSVM-HMM: Transductive SVM based hidden Markov model for automatic image annotation

Yufeng Zhao *, Yao Zhao, Zhenfeng Zhu

Institute of Information Science, Beijing Jiaotong University, Shanfyuancun, Xizhimen Wai, Beijing 10044, China

ARTICLE INFO

Keywords:
Automatic image annotation (AIA)
Hidden Markov model (HMM)
Transductive SVM
Visual feature distribution
Keyword correlation

ABSTRACT

Automatic image annotation (AIA) is an effective technology to improve the performance of image retrieval. In this paper, we propose a novel AIA scheme based on hidden Markov model (HMM). Compared with the previous HMM-based annotation methods, SVM based semi-supervised learning, i.e. transductive SVM (TSVM), is triggered out for remarkably boosting the reliability of HMM with less users' labeling effort involved (denoted by TSVM-HMM). This guarantees that the proposed TSVM-HMM based annotation scheme integrates the discriminative classification with the generative model to mutually complete their advantages. In addition, not only the relevance model between the visual content of images and the textual keywords but also the property of keyword correlation is exploited in the proposed AIA scheme. Particularly, to establish an enhanced correlation network among keywords, both co-occurrence based and WordNet based correlation techniques are well fused and are able to be helpful for benefiting from each other. The final experimental results reveal that the better annotation performance can be achieved at less labeled training images.

1. Introduction

In content-based image retrieval (CBIR) research, it is well-known that using low-level features solely to find similar images is usually insufficient due to the semantic gap. One solution for this problem is to manually annotate each image with keywords, which assign multiple semantic contents for each image. However, manual annotation is tedious and difficult, especially for the large image database. Therefore, automatic image annotation (AIA) has become a focus and received massive investigation on it while still keeps the advantages of semantically annotating. The purpose of AIA is to mine the relevance model between the visual content of images and the textual keywords. As a milestone in AIA, Duygulu, Barnard, de Freitas, and Forsyth (2002) propose translation model (TM) to treat AIA as a process of translation from a set of blob tokens, obtained by clustering image regions, to a set of keywords. By importing the relevance language models into AIA, TM is promoted by cross-media relevance model (CMRM) (Jeon, Lavrenko, & Manmantha, 2003). Subsequently, CMRM is improved through continuous-space relevance model (CRM) (Manmatha, Lavrenko, & Jeon, 2003) and multiple-Bernoulli relevance model (MBRM) (Feng, Manmatha, & Lavrenko, 2004). Another way of building relevance model is to introduce the latent variables to associate images with keywords, such as LSA model (Chen, Tsai, & Chan, 2008), LPS model (Kumar, Torr, & Zisserman, 2005) and hidden concept model (Zhang, Zhang, Li, Ma, & Zhang, 2005). Conforming to the essence of relevance models, the classification technique is also employed to touch on the case of AIA. Yang and Dong (2006) create a three-level classification scheme for each keyword by fusing one-class, two-class and multi-class SVM classifier.

Being more reliable than Yang and Dong (2006), Goh, Chang, and Li (2005) propose ASVM-MIL to extend the conventional support vector machine (SVM) to construct MIL. Instead of depending on the discriminative classification (Goh et al., 2005; Yang & Dong, 2006), generative model is involved into AIA because of intrinsic advantages (Ghoshal, Ircing, & Khudanpur, 2005; Gustavo, Antoni, Pedro, & Nuno, 2007). As one representative work of generative models, hidden Markov model (HMM) is an available method to resolve AIA (Ghoshal et al., 2005). Recently, semi-supervised learning (SML) is also utilized to focus on AIA for avoiding the image segmentation (Gustavo et al., 2007).

However, although some endeavors have been made for these approaches to enhance the quality of AIA, there are still some disadvantages. First, some of these relevance models not only require a large number of labeled images to guarantee their feasibleness but also have the problem of highly unbalance between the discriminative classification and the generative model, i.e. they are solely concerned in AIA. Second, the correlation among the textual keywords is ignored in these relevance models so that less human's intention are involved. To address on these problems, a novel TSVM-HMM based annotation scheme is proposed in this paper.

* Corresponding author. Tel.: +86 010 51688005; fax: +86 010 51688667.
E-mail address: snowmanzhao@163.com (Y. Zhao).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.050

First, only given several labeled regions for a keyword, SVM based semi-supervised learning, i.e. transductive SVM (TSVM), is a promising way to find out the underlying relevant regions from the unlabeled ones. With these minded relevant regions, GMM can be correctly developed to describe the visual feature distribution of each keyword. Second, co-occurrence based keyword correlation is well combined with WordNet based keyword correlation to build up an enhanced correlation network among the textual keywords. Thus, HMM over all the semantic keywords is able to be reliably constructed with the visual feature distribution of each keyword and the keyword correlation. Compared with the previous annotation methods, the proposed TSVM-HMM based annotation scheme fuses both the discriminative classification and the generative model to mutually complete each other. Furthermore, less users' labeling efforts are required because of the self-learning ability of TSVM. Meanwhile, the better keyword correlation is useful to gain more human's intentions and hence the AIA quality is greatly promoted.

The outline of this paper is as follows: The proposed AIA scheme is briefly reviewed in Section 2. Correspondences of TSVM-HMM based annotation scheme is explained in Section 3. Section 4 presents the experimental results to verify the performance of the proposed method. The paper is then concluded in Section 5.

2. TSVM-HMM based annotation scheme

For further discussion, some necessary definitions are first listed. Let $Tr = \{I_1, I_2, \ldots, I_N\}$ be the set of training images with annotations, where N is the total number of images. For a given image I, let $I^{reg} = \{r_1, \ldots, r_T\}$ be its region set and $I^{kw} = \{w_1, \ldots, w_M\}$ be its keyword set, where T denotes the number of regions and M is the number of keywords. The vocabulary V is generated by collecting all the keywords of N labeled training images.

Generally, it is not provided that the region r_t in the region set I^{reg} corresponds to one keyword w_j in keyword set I^{kw}. Thus, AIA can be modeled as a hidden Markov process and formulated to be the joint likelihood (Ghoshal et al., 2005):

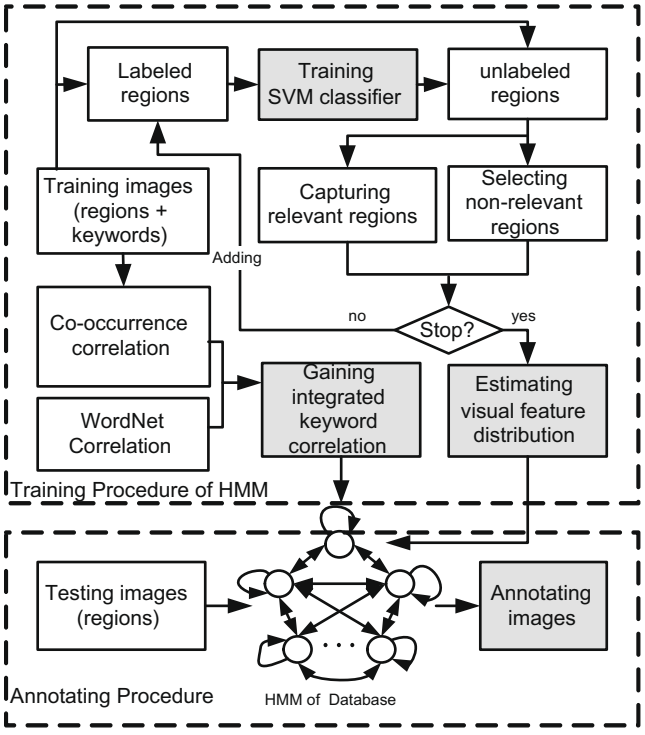

Fig. 1. Proposed TSVM-HMM based annotation scheme.

$$f(r_1^T, I^{kw}|w_0) = \sum_{w_1^T \in I^{kw}} \prod_{t=1}^{T} f(r_t|w_t) p(w_t|w_{t-1}) \quad (1)$$

where r_1^T is the region sequence $\langle r_1, \ldots, r_T \rangle$ of image I and w_1^T is its corresponding keyword sequence $\langle w_1, \ldots, w_M \rangle$. The region sequence is ordered according to the size of region area, and the region with the largest area is arranged on the first. In Eq. (1), the performance of HMM based annotation scheme will be directly affected by the emission density f and the transition probability$p$, which represent the visual feature distribution associated with each keyword and the keyword correlation, respectively. Rather than the complex maximum likelihood algorithm in (Ghoshal et al., 2005), we propose that TSVM is triggered out for exploring more relevant regions from the unlabeled ones with less users' labor efforts. Then, the visual feature distribution f of each keyword modeled as GMM is directly estimated via these relevant regions. Moreover, different from the aforementioned relevance models, the keyword correlation p is also applied to exploiting more human's intentions by fusing both co-occurrence based and WordNet based keyword correlation. Therefore, both the generative model integrated with discriminative classification and the keyword correlation ensures to create a more reliable HMM. With the obtained HMM, the test images is annotated by the Balm–Welch algorithm. Fig. 1 illustrates the framework of TSVN-HMM based annotation method.

3. Correspondences of TSVM-HMM based annotation scheme

The proposed TSVM-HMM based annotation scheme is composed of four correspondences. First, given several relevant regions for each keyword, more relevant regions are mined by TSVM. Second, the visual feature distribution$f$of each keyword is well characterized through GMM, which is modeled with the expanded relevant regions. Third, the enhanced keyword correlation is captured via combining the co-occurrence based keyword correlation with the WordNet based keyword correlation. Fourth, the Balm–Welch algorithm contributes to annotating the test images by the reliable HMM.

3.1. TSVM for mining relevant regions

Given a keyword w, several labeled regions are taken as the relevant examples and the initial non-relevant examples are randomly sampled from the remaining regions. A two-class SVM classifier is trained firstly. Then, based on the learnt SVM classifier, the most confident relevant regions and the most non-relevant ones are added into the relevant and non-relevant training set, respectively. With the expanded training set, SVM classifier will be re-trained until the maximum time of iteration is reached. Finally, an expanded set of labeled regions (including pseudo labeled

Table 1
Pseudo-code of TSVM for mining relevant regions.

1. Input: R_L^0 a set of labeled regions for the keyword w
   R_U^0 a set of unlabeled regions for the keyword w
   S a SVM classifier
   m, n, K controlling parameters
2. For $k = 1$ to K
   (a) Learning a SVM classifier S from R_L^k
   (b) Using S to classify regions in R_U^k
   (c) Selecting m most confidently predicted regions from R_U^k which are labeled relevant examples
   (d) Selecting n most confidently predicted examples from R_U^k which are labeled as non-relevant
   (e) Adding $m + n$ regions with their corresponding labels into R_L^k
   (f) Removing these $m + n$ regions from R_U^k
3. Output: R_L^K an expanded set of labeled regions