Pattern Recognition 79 (2018) 242–259

Contents lists available at ScienceDirect

Pattern Recognition

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

A survey and analysis on automatic image annotation

Qimin Cheng[a,*], Qian Zhang[a], Peng Fu[b], Conghuan Tu[a], Sen Li[a]

[a] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[b] Department of Earth and Environmental Systems, Indiana State University, Terre Haute, IN 47809, USA

ARTICLE INFO

Article history:
Received 19 April 2017
Revised 31 January 2018
Accepted 11 February 2018
Available online 13 February 2018

Keywords:
Automatic image annotation
Generative model
Nearest-neighbor model
Discriminative model
Tag-completion
Deep learning

ABSTRACT

In recent years, image annotation has attracted extensive attention due to the explosive growth of image data. With the capability of describing images at the semantic level, image annotation has many applications not only in image analysis and understanding but also in some relative disciplines, such as urban management and biomedical engineering. Because of the inherent weaknesses of manual image annotation, Automatic Image Annotation (AIA) has been raised since the late 1990s. In this paper, a deep review of state-of-the-art AIA methods is presented by synthesizing 138 literatures published during the past two decades. We classify AIA methods into five categories: 1) Generative model-based image annotation, 2) Nearest neighbor-based image annotation, 3) Discriminative model-based image annotation, and 4) Tag completion-based image annotation, 5) Deep Learning-based image annotation. Comparisons of the five types of AIA methods are made on the basis of the underlying idea, main contribution, model framework, computational complexity, computation time, and annotation accuracy. We also give an overview of five publicly available image datasets and four standard evaluation metrics commonly used as benchmarks for evaluating AIA methods. Then the performance of some typical or well-behaved models is assessed based on benchmark dataset and standard evaluation metrics. Finally, we share our viewpoints on the open issues and challenges in AIA as well as research trends in the future.

1. Introduction

The big data era is characterized by the huge amount of image data available. Traditional image annotation techniques, labeling image contents at the semantic level manually, are not applicable in the big data era. The main disadvantages of manual image annotation are intuitionistic. First, it is unpractical to annotate the mass image data totally through manual ways. Second, the subjectivity of manual annotation will lead to ambiguity over image contents. In other words, different persons may have totally different understandings of the very same image because of differences in the educational background, thinking mode, and even life experience.

Given deficiencies of traditional manual image annotation, research on Automatic Image Annotation (AIA) technology has become a tendency. Inspired by the word co-occurrence model proposed by Mori et al. [1] in 1999, more and more scholars have turned to conduct studies on annotating images by weak-supervision or totally automatic ways. These achievements have boosted the development of AIA to a great extent during the past two decades. AIA methods are concerned with models/algorithms to label images by their semantic contents or to explore the similarity between image features and semantic contents with high efficiency and low subjectivity. Relevant labels are predicted for untagged images from a label vocabulary through the weak-supervision way or totally automatically. The key of the AIA is to narrow the semantic gap between low-level visual features and high-level semantic labels, i.e., to learn high-level semantic labels from low-level visual features by exploring the image-image correlation, image-label correlation, and label-label correlation. In addition to its applications in image understanding and analysis, such as image retrieval [2–5], scalable mobile image retrieval [6], face recognition [7], facial landmark annotation [8], and photo tourism [9], AIA is also used in urban management, biomedical engineering, social media services and tourism industry, to name a few. As an interdisciplinary discipline, AIA integrates achievements from data mining, semantic analysis, Natural Language Processing (NLP),Automatic Deep Understanding (ADU) of documents, document analysis and recognition, multimedia systems, machine learning, and even biology and statistics.

During the past two decades, considerable efforts have been made to develop various AIA methods [2–4,10–26]. The learning-based annotation techniques/algorithms include the TM model [10], CMRM model [2], CRM model [11], MBRM model

* Corresponding author.
E-mail address: chengqm@hust.edu.cn (Q. Cheng).

https://doi.org/10.1016/j.patcog.2018.02.017
0031-3203/

[3], Plsa-based model [27], Markov Random Fields(MRF)-based model [4,28], classification model [14,29–32], graph-based semi-supervised learning methods [33–41], ML-LOC [18], and deep learning-based methods [13,25,26,42–49].The retrieval-based annotation techniques/algorithms include the baseline model [50], UDML model [51]), PDML model [52], and 2PKNN model [16]. More recently, some researches perform AIA through automatically filling in the missing tags as well as correcting noisy tags for given images [20,24,53–55].

At present, a general classification and deep review of AIA methods is still lacking. Despite some surveys of AIA methods, the foci were placed on CBIR [56–59], feature extraction and semantic learning/annotation [60–62], statistical approaches [63], image segmentation [64], face recognition [65], Natural Language Processing (NLP) [66] and relevance feedback [67,68]. In this paper, a comprehensively comparative review of AIA methods is presented by synthesizing 138 literatures published during the past two decades. Specifically, this review covers papers published in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), International Journal of Computer Vision (IJCV), Pattern Recognition (PR), Journal of Machine Learning Research (JMLR), ACM Transactions on Graphics (TOG), IEEE Transaction on Multimedia (TMM), and IEEE Transaction on Image Processing (TIP), and papers published in conferences such as AAAI Conference on Artificial Intelligence (AAAI), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and International Conference on Computer Vision (ICCV).

We focus on a more general classification of the AIA methods by five categories: generative model-based AIA methods, nearest neighbor model-based AIA methods, discriminative model-based AIA methods, tag completion-based AIA methods, and deep learning-based AIA methods. Those AIA methods are analyzed and compared based on the underlying idea, main contribution, model framework, computational complexity, and annotation accuracy (in Section 2). Then this paper reviews five publicly available image datasets and four standard evaluation metrics adopted by AIA methods (in Section 3). This paper also assesses some typical or well-behaved models based on the benchmark dataset and the standard evaluation metrics (in Section 4). We also discuss some challenges, open issues, and promising directions in AIA (in Section 5).

2. Annotation methods

There are various classification schemes for AIA techniques, such as probability and non-probability methods, learning-based and retrieval-based methods, supervised, semi-supervised and unsupervised methods. In this paper we classify these methods into five categories: 1) generative model-based AIA methods, which are dedicated to maximizing generative likelihood of image features and labels; 2) nearest neighbor model-based AIA methods, which assume that images with similar features have a great probability to share similar labels; 3) discriminative model-based AIA methods, which view the annotation task as a multi-label classification problem; 4) tag completion-based AIA methods, which can not only predict labels by automatically filling in the missing labels but also can correct noisy tags for given images; 5) deep learning-based AIA methods, which use deep learning algorithms to derive robust visual features or exhaustive side information for AIA, especially for large-scale AIA.

The aforementioned five categories of AIA methods can be further classified into several sub-categories according to their underlying ideas. Fig. 1 provides a taxonomy, as well as some hot topics of AIA methods by covering 138 literatures.

In Fig. 1, generative model-based AIA methods can be mainly divided into three classes including the relevance model, topic model and hidden Markov model (HMM). As for nearest neighbor model-based AIA methods, three key issues, i.e., distance metric learning (DML), class-imbalance, and weak-labeling, are receiving more attention. With regards to the discriminative model-based AIA, research efforts have been mainly devoted to developing the graph-based semi-supervised learning methods. The advantage of the graph-based methods is that the label correlation can be easily incorporated into the graph in the propagation process. As such, the way to describe the label correlation plays an important role in developing AIA methods. For the tag completion-based AIA methods, they can be further divided into the matrix completion, linear sparse reconstructions, subspace clustering, and low-rank matrix factorization. With respect to deep learning-based AIA methods, great progress has been made in two facets for annotation, i.e., derivation of robust visual features, and exhaustive utilization of side information.

2.1. Generative model-based AIA methods

The generative model-based AIA methods are quite popular, and great achievements have been made in the early 21st century. The generative models are dedicated to maximizing the generative likelihood of image features and labels. For an untagged image, the generative model-based AIA techniques provide the probability of an image label by computing a joint probabilistic model of image features and words from training datasets. The generative models used for AIA mainly consist of the relevance model, topic model, and Markov random field model.

2.1.1. The relevance model

The relevance model-based AIA methods are generally implemented in three steps: define the joint distributions over image features and labels; compute the posterior probability of each label for the unlabeled images (usually the visual feature); to annotate a new image by choosing a label of the highest probability. Various relevance models have been developed for image annotation, including the translation model (TM) [10], across media relevance model (CMRM) [2], continuous space relevance model (CRM) [11], and multiple Bernoulli relevance models (MBRM) [3].

The TM creates a one-to-one match between a blob and a word [10]. In this model, regions are firstly clustered from training images and represented by the index of the closest centroid of the cluster (blob). Next, each blob is associated with a word in the vocabulary, similar to the process of learning a lexicon, by maximizing the joint probability through the EM algorithm, which is computationally expensive and time-consuming.

The CMRM also uses the blob generated from image features to describe an image [2]. It computes the joint distribution between keywords and the entire image rather than specific blobs that are used in TM since the blob vocabulary may give rise to many errors. In the CMRM, an image I is represented by a set of blobs $\{b_1, b_2 \dots b_n\}$, and the conditional probability of image I belonging to a class w is approximated as (1):

$$p(w|I) = p(w|b_1, b_2 \dots b_n) \tag{1}$$

The training set derived from annotated images is used to estimate the joint probability for the word w and the blobs $\{b_1, b_2 \dots b_n\}$. The joint probability distribution can be computed over the image j in the training set T as (2):

$$p(w, b_1, b_2 \dots b_m) = \sum_{j \in T} p(j) p(w, b_1, b_2 \dots b_m | j) \tag{2}$$

Once the image j is known, the prior probability $p(j)$ is constant for the entire training set. By assuming words w and blobs $\{b_1, b_2 \dots b_m\}$ are independent, a word model and a blob model are