J. Vis. Commun. Image R. 34 (2016) 167–175

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Image distance metric learning based on neighborhood sets for automatic image annotation ☆

Cong Jin [a,*], Shu-Wei Jin [b]

[a] School of Computer, Central China Normal University, Wuhan 430079, PR China
[b] Département de Physique, École Normale Supérieure, 24, rue Lhomond, 75231 Paris Cedex 5, France

ARTICLE INFO

Article history:
Received 24 May 2015
Accepted 28 October 2015
Available online 19 November 2015

Keywords:
Automatic image annotation
Improve performance
Image distance metric learning
Neighborhood sets
Algorithm performance
Visual similarity
Semantic similarity
Probability density ratio

ABSTRACT

Since there is semantic gap between low-level visual features and high-level image semantic, the performance of many existing content-based image annotation algorithms is not satisfactory. In order to bridge the gap and improve the image annotation performance, a novel automatic image annotation (AIA) approach using neighborhood set (NS) based on image distance metric learning (IDML) algorithm is proposed in this paper. According to IDML, we can easily obtain the neighborhood set of each image since obtained image distance can effectively measure the distance between images for AIA task. By introducing NS, the proposed AIA approach can predict all possible labels of the image without caption. The experimental results confirm that the introduction of NS based on IDML can improve the efficiency of AIA approaches and achieve better annotation performance than the existing AIA approaches.

1. Introduction

In the past decade, with the rapid development of Internet and the popularization of digital cameras and mobile phones, more and more digital images are hosted and shared on online image sites. Systems managing and analyzing images on image sites heavily depend on textual annotations of images. Many existing image search engines rely on associated text in the web page such as Yahoo!, Google and so on. Automatic image annotation (AIA) has become a challenging task for users quickly and efficiently to retrieval the interest image resources.

As an important part of image retrieval, the accuracy of annotated image semantic directly affects the performance of image retrieval system. In the past two decades, there have been three main image retrieval technologies [1]: text-based image retrieval (TBIR), content-based image retrieval (CBIR) [2] and semantic-based image retrieval (SBIR) [3]. For TBIR, the images are generally annotated manually, however, due to the cost of manual annotation is so high that it is difficult to realize image retrieval for large image databases. The manual annotation will not only limit the number of retrieved images, but also the efficiency of the retrieval system is low. For CBIR, it computes relevance only based on the similarity of low-level visual features such as colors, textures and shapes [4,5]. In fact, people prefer retrieving images according to high-level semantic content. However, there is a gap between low-level visual features and high-level semantic contents, therefore the performances of many existing content-based image annotation algorithms were not so satisfactory [6–8].

Due to there exists difficulties in the TBIR and CBIR, SBIR has been focused. For SBIR, the images with semantic labels can then be retrieved accordance to the labels similarity using TBIR. AIA approach aims to automatically generate labels to describe the content of a given image. Currently, the most common AIA approaches have two types [9,10], i.e., classification-based and probabilistic modeling-based approaches.

In first type, AIA can be viewed as a classification problem [11], which can be solved by a classifier. For annotating an image without caption, first, represent image into a low-level visual features vector. Then, classify the image into some categories. Finally, propagate the semantic of the corresponding category to given image. So, the unlabeled image may be automatically annotated.

In second type, probabilistic model [12] attempts to infer the joint probabilities between images and semantic concepts. Images given class can be regarded as instances of stochastic process that characterizes the class. Then, statistical models, such as Markov, Gaussian, and Bayes and so on, are trained and images are classified based on probability computation.

Although these two types are the most common annotation approaches, there are still some disadvantages. For example,

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail address: jincong@mail.ccnu.edu.cn (C. Jin).

http://dx.doi.org/10.1016/j.jvcir.2015.10.017
1047-3203/

classification-based annotation approach heavily relies on visual similarity for judging semantic similarity. In fact, it is well-known that semantic similarity does not equal to visual similarity [13,14]. In addition, in many papers, the distance between the images is measured according to some traditional methods, e.g., Euclidean distance, Mahalanobis distance, Hamming distance, Cosine distance, Histogram distance and so on [15,16]. Although these traditional distances are simple and convenient, it cannot accurately measure the similarity between two images in many cases.

In this paper, we propose a novel AIA approach, named NSIDML, which is characterized by learning image distance metric (IDM) based on existing knowledge of the samples and predicting all possible labels based on learned IDM using neighborhood set [17] (NS). In proposed NSIDML, NS is used to reduce the bias between visual similarity and semantic similarity. In proposed image distance metric learning (IDML) algorithm, all training samples are used for better utilizing existing resources and obtaining a more efficient AIA approach. The main contribution of this work is as follows:

(1) In this work, knowledge of sample set without caption, in the training set, is sufficiently considered, but not to limit knowledge of the training samples with caption, which ensures that the existing resources can be sufficiently utilized.
(2) For image without caption, the number of its labels is not predetermined. In other words, the number of labels is completely determined by image content.
(3) In the image annotation process, the proposed AIA approach is almost no human interaction. In other words, it can automatically implement and reduce the impact of human subjectivity.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 reviews the preliminary knowledge including the neighborhood of image and block diagram of AIA approach. Section 4 describes the proposed IDML algorithm. Section 5 introduces the proposed AIA approach NSIDML based on IDML. Section 6 presents the experimental and comparison results. Finally, conclusions are given in Section 7.

2. Related work

In this section, we will introduce related work of image annotation approach; in particular, the classification-based and probabilistic model-based annotation approaches.

2.1. Classification-based image annotation

Classification-based approach for image semantic annotating in large image dataset has been researched [18–22]. Cusano et al. [18] described an innovative image annotation tool for classifying image regions in one of seven classes. The annotation is performed by a classification system based on a multi-class support vector machine. Li et al. [19] proposed a hierarchical generative model that classifies the overall scene, recognizes and segments each object component, as well as annotates the image with a list of tags. Sun et al. [20] proposed 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. Chen et al. [21] proposed an adaptive recognition model (ARM) for image annotation. The ARM consists of an adaptive classification network (CFN) and a nonlinear correlation network (CLN). Image annotation is carried out by an ARM. The exploitation of a controllable synthetic dataset helps to systematically study the function of keyword correlation and effectively analyze the performance of the ARM. Guillaumin et al. [22] proposed a discriminatively trained nearest neighbor model, called TagProp. Tags of test images are predicted using a weighted nearest-neighbor model. The authors also introduce a word specific sigmoidal modulation of the weighted neighbor tag predictions to boost the recall of rare words.

2.2. Probabilistic model-based image annotation

One of the earliest examples of a probabilistic approach to image annotation is the co-occurrence model proposed by Mori et al. [23]. The authors segmented images using a simple regular grid and a probabilistic generative model was learnt based on the co-occurrence statistics of vocabulary keywords and the clusters derived from segmented image regions. Wang et al. [24] developed a probabilistic model for jointly modeling the image. The proposed model treated the class label as a global description of the image, and treated annotation terms as local descriptions of parts of the image. Its underlying probabilistic assumptions naturally integrate these two sources of information. An approximate inference and estimation algorithms based on variational methods can be derived. Li et al. [25] proposed a multi-correlation probabilistic matrix factorization (MPMF) algorithm for the correlation estimation. Different from the traditional solutions which treat the image-word correlation, image similarity and word relation independently or sequentially, in the proposed MPMF, these three elements are integrated together simultaneously and seamlessly. Carneiro et al. [26] proposed a probabilistic formulation for semantic image annotation and retrieval. In particular, images are represented as bags of localized feature vectors, a mixture density estimated for each image, and the mixtures associated with all images annotated with a common semantic label pooled into a density estimate for the corresponding semantic class. Zhou et al. [27] proposed a hybrid probabilistic model (HPM) which integrates low-level image features and high-level user provided labels to automatically annotate images. The authors developed an L_1 norm kernel method to estimate the correlations between image features and semantic concepts.

In above these approaches, a lot of image annotation approaches are based on image segmentation [18–20,23], which is an important method to extract the labels of the image, and these labels can well describe content of the image. However, there are some disadvantages for image segmentation, such as segmentation results fragile and erroneous etc. In addition, many measure of the sample similarity is obtained according to some traditional distances [18–20,24,25,27], and however these traditional distances are not always appropriate.

3. Preliminary

Set theory [28] is the branch of mathematical logic that studies sets, which is collections of objects. The language of set theory can be used in the definitions of nearly all objects. In this paper, NS is applied to solve difficult of AIA task.

For further discussion, some necessary notations and definitions are first introduced. Let $Tr = \{I_1, I_2, \ldots, I_{N1}\}$ be the set of training images and $Te = \{I_1, I_2, \ldots, I_{N2}\}$ the set of testing images without caption, $N1 + N2 = n$, and an image is represented as a M-dimensions vector $I = \{x^1, x^2, \ldots, x^M\}$, $I \in Tr \cup Te$. Let $L = \{l_1, l_2, \ldots, l_m\}$ be the set of possible annotated labels, and each image $I \in Tr$ is associated with a subset $Y \subseteq L$. Y may be represented as an m-dimensional vector, i.e., $Y = (y^1, y^2, \ldots, y^m)$, which $y^j = 1$ only if image I has label