Pattern Recognition 61 (2017) 169–184

Contents lists available at ScienceDirect

Pattern Recognition

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

A multi-expert based framework for automatic image annotation ☆

Abbas Bahrololoum, Hossein Nezamabadi-pour *

Department of Electrical Engineering, Shahid Bahonar University of Kerman, P.O. Box 76169-133, Kerman, Iran

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form
25 June 2016
Accepted 22 July 2016
Available online 25 July 2016

Keywords:
Automatic image annotation
Feature space
Concept space
Prototype
Semantic gap
Fusion

ABSTRACT

Automatic image annotation (AIA) for a wide-range collection of image data is a difficult challenging topic and has attracted the interest of many researchers in the last decade. To achieve the goal of AIA, a multi-expert based framework is presented in this paper which is based on the combination of results obtained from feature space and concept space. Considering a real-world image dataset, a large storage is required; therefore, the idea of generating prototypes in both feature and concept spaces is used. The prototypes are generated in learning phase using a clustering technique. The input unlabeled images are assigned to the nearest prototypes in both feature and concept spaces, and primary labels are obtained from the nearest prototypes. Eventually, these labels are fused and final labels for a target image are chosen. Since all feature types do not describe a concept label equally, some prototypes are more effective to represent a concept and bridge the semantic gap, so a metaheuristic algorithm is employed to search for the best subset of feature types and best criterion of fusion. To evaluate the performance of the proposed framework, an example of its implementation is presented. A comparative experimental study with several state-of-the-art methods is reported on two standard databases of about 20k images. The obtained results confirm the effectiveness of the proposed framework in the field of automatic image annotation.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The considerable development of the digital acquisition, computer hardware, storage techniques and Internet technology makes millions of images accessible to people. One widely adopted solution for accessing and retrieving digital images in addition to video is to annotate the content with semantically meaningful labels. Two types of annotation approaches are available: manual and automatic [1]. Manual image annotation is time-consuming, laborious and expensive task; to address this, many researches have focused on automatic image annotation.

The goal of automatic image annotation goes to assign a collection of keywords (annotation) from a given dictionary to a target image (previously unseen). That is, the input is the target (untagged) image and the output is a collection of keywords that describe the target image in the best possible way [2]. In other words, the automatic system semantically describes the content of an image. To do this, a set of semantic labels is assigned to each image to describe its content [3]. Then, a system is developed to provide a model for the relation between visual features and tags of images.

Automatic image annotation has been reviewed extensively for several years. The image annotation is just an extremely challenging task. The same object can be captured from different angles, distances or under different luminance conditions. This is subjective and sometimes it is difficult to automatically describe image content by keywords [1]. Additionally, an object of the real world with the same “name” may have different visual content (e.g., shape, color). The semantic gap between low level features and high level concepts (i.e. the interpretation of the images in the way that humans do) is a fundamental problem in a content-based image retrieval (CBIR) system.

To bridge the semantic gap, some systems use the relevance feedback technique [4–8] to incorporate user knowledge into the retrieval process. Some approaches attempt to reduce annotation errors by making use of word relations [9]. Other approaches make use of external resources such as auxiliary texts of web images, WorldNet and ontology, Google distance, click-through data, and Wikipedia articles [10]. Topic based approaches model joint distributions of visual features and words [11]. On the other hand, multiple instance learning (MIL) approaches [12] focus on solving the problem of weakly labeling in image annotation that is the absence of correspondence between labels and regions in images. Multiple feature spaces [13] are also selected to improve the performance of CBIR systems. Recently, studies [14] on

☆Authors' Google scholar Homepage: https://scholar.google.com/citations?user=GopXT0MAAAAJ&hl=en; https://scholar.google.com/citations?user=OJQ70wEAAAAJ&hl=en.

* Corresponding author.
E-mail addresses: a.bahrololoum@uk.ac.ir (A. Bahrololoum), nezam@uk.ac.ir (H. Nezamabadi-pour).

http://dx.doi.org/10.1016/j.patcog.2016.07.034
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

jointly modeling scene classification and image annotation have been used.

However, in image annotation problem, images are often described by multiple feature space (multiview features). Different views such as color, texture and shape features, describe different attributes of an image [15–17]. Each view describes a property of the image, and the weaknesses of a view can be reduced by the strengths of others.

Multiview learning algorithms can be grouped into three categories: a) co-training, b) multiple kernel learning and c) subspace learning. Co-training style algorithms usually train separate learners on distinct views, which are then forced to be consistent across views. It is assumed that the features obtained from the different views are sufficient and they are conditionally independent of one another to train a classifier. Multiple kernel learning algorithms calculate separate kernels on each view which are combined with a kernel-based method. Subspace learning-based approaches aim to obtain an appropriate subspace to explore the complementary properties of different views by assuming that input views are generated from a latent view [16].

1.1. Related works

Automatic image annotation methods are usually classified into two categories, namely probabilistic modelling-based methods [18,19] and classification-based methods [20–22]. One strategy for statistical annotation is unsupervised labeling which estimates the joint density of visual features and words by implementing an unsupervised learning algorithm on a training image dataset. These methods introduce a hidden variable and assume that features and words are independent of the hidden variable value. Another formulation for statistical annotation is supervised multi-class labeling [20] that estimates a conditional distribution for each semantic class to determine probability. The problem of multi-label classification generalizes the traditional multi-class classification problem, the former allows a set of labels to be associated with an instance whereas the latter allows only one. An image to be annotated can get several labels simultaneously, that makes this problem as a multi-label one [23].

The authors in [24] present a multi-label classification framework for automatic image annotation. The proposed framework comprises an initial clustering phase that breaks the original training set into several disjoint clusters of data. It then trains a multi-label classifier from the data of each cluster. Given a new test instance, the framework first finds the nearest cluster and then applies the corresponding model.

The authors in [25] propose a solution to the problem of large scale concept space learning and mismatch between semantic and visual spaces (semantic gap). To tackle the first issue, they present the use of higher level semantic space with lower dimension by clustering correlated keywords into topics in a local neighborhood. The topics are used as lexis for assigning multiple labels to unlabeled images. To deal with the problem of semantic gap, they propose a way to reduce the bias between visual and semantic spaces by finding optimal margins in both spaces. In particular, the proposed method is an iterative solution that alternately maximizes the sum of margins to reduce the gap between visual and semantic similarities.

In the paper [26], authors present multiview Hessian Regularization (mHR) for image annotation. The proposed method combines multiview features and Hessian regularizations obtained from different views. It is claimed that the method effectively explores the complementary properties of different features from different views and thus boosts the image annotation performance significantly. In [27], the authors propose the multiview Hessian discriminant sparse coding (mHDSC) scheme for image annotation. The method employs Hessian regularization (HR) to encode the local geometry. And, it is applied to multiview features. In addition, mHDSC acts the label information as an additional view of feature to boost the discrimination of the dictionary.

The co-occurrence model proposed by Mori et al. [28] is perhaps one of the first attempts at image auto-annotation. They first divide images into rectangular tiles of the same size, and calculate a feature descriptor of color and texture for each tile. All the descriptors are clustered into a number of groups, each of which is represented by a centroid. On the other hand, each tile inherits the whole set of labels from the original image. Second, for the set of segments, the probability of each keyword is estimated by using a vector quantization of the segment's features. This method has a relatively low annotation performance [1].

Duygulu et al. propose machine translation model (TM) [29], which considers image annotation as a translation problem between two languages: one language is visual vocabulary of image contents; the other is real text. They use normalized cut algorithm to segment images, and then use K-means algorithm to cluster these regions. Image annotation can be regarded as translation processes from visual vocabulary blobs to the semantic keywords. Mapping between blobs and keywords was learned using the expectation–maximization (EM) algorithm. One of the key problems of the model is high computational complexity of the EM algorithm, so it is not suitable for large-scale datasets.

Inspired by the relevance language models for information retrieval and cross-lingual retrieval, several relevance models have been proposed such as continuous relevance model (CRM) which directly uses continuous features of image regions and non-parametric Gaussian kernel to continuously estimate generation probability of visual contents [30], cross-media relevance model (CMRM) which uses joint probability of semantic labels and visual words to annotate images [31], dual cross-media relevance model (DCMRM) which performs image annotation by maximizing the joint probability of images and words [32]. The suggested dual model involves two types of relations, word-to-word and word-to-image relations, both of which are estimated by using search techniques on the web data, and multimodal latent binary embedding (MLBE) [33]. Feng et al. propose the multiple Bernoulli relevance model (MBRM) [34] which utilizes rectangular grids instead of complicated segmentation algorithms to partition images. They applies Bernoulli distribution instead of multinomial distribution to describe the distribution of vocabulary that takes into account image context, i.e., from training images it learns that a class is more associated with some classes and is less associated with some other classes. The authors claimed that this method is more effective for image annotation than the translation model. However, its drawback is that only images consistent with the training images can be annotated with keywords in a limited vocabulary.

Amiri and Jamzad in [3] developed an annotation system for semi-supervised learning framework that constructs a generative model for each semantic class in two main steps. First, based on Gamma distribution a generative model is constructed for each semantic class using labeled images in that class. The second step incorporates the unlabeled images by using a modified EM algorithm to update parameters of the constructed models.

Metzler and Manmatha [35] segmented training images, connected them and their annotations in an inference network, whereby an unseen image is annotated by instantiating the network with its regions and propagating belief through the network to nodes representing the words.

A non-parametric density estimation approach and the technique of kernel smoothing have been proposed by Yavlinsky et al. [36]. They have claimed that the results are comparable with the inference network and CRM. These automatic annotation approaches

Download English Version:

https://daneshyari.com/en/article/533070

Download Persian Version:

https://daneshyari.com/article/533070

Daneshyari.com