Pattern Recognition 48 (2015) 174–188

Contents lists available at ScienceDirect

Pattern Recognition

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

Automatic image annotation using semi-supervised generative modeling

S. Hamid Amiri, Mansour Jamzad *

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form
11 June 2014
Accepted 9 July 2014
Available online 19 July 2014

Keywords:
Image annotation
Semi-supervised learning
Generative modeling
Gamma distribution

ABSTRACT

Image annotation approaches need an annotated dataset to learn a model for the relation between images and words. Unfortunately, preparing a labeled dataset is highly time consuming and expensive. In this work, we describe the development of an annotation system in semi-supervised learning framework which by incorporating unlabeled images into training phase reduces the system demand to labeled images. Our approach 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 generative models. Performance evaluation of the proposed method on a standard dataset reveals that using unlabeled images will result in considerable improvement in accuracy of the annotation systems when a limited number of labeled images for each semantic class are available.

1. Introduction

The growth of digital image datasets and photo sharing communities on the Internet makes it necessary to provide a proper mechanism for searching the images in large collections. Content-based image retrieval (CBIR) focuses on the problem of searching images based on their content. The first generation of CBIR was based on query-by-example (QBE) paradigm [1], in which the CBIR system retrieves most visually similar images to a query image given by the user. The semantic gap between low level features and high level concepts is a fundamental problem in QBE systems. To bridge the semantic gap, many systems use the relevance feedback approach [2,3] to incorporate user knowledge into the retrieval process. Besides these query-based search strategies, the new generation focuses on development of an automatic system that semantically describes the content of an image. In this approach, a set of semantic labels are assigned to each image to describe its content. Then, a system is developed to train a model for the relation between visual features and tags of images. This formulation, called automatic image annotation or linguistic indexing, allows the system to assign words to every new test image. Furthermore, the retrieval procedure could be performed based on input texts provided by the user. To bridge the semantic gap in an annotation system, some methods utilized active learning [4] to exploit the user knowledge in developing the annotation system and thus to reduce the required supervision. In what follows, we briefly overview main features of the prior studies.

1.1. Related works

There has been a great effort to design an annotation system using statistical learning. According to Ref. [5], one strategy for statistical annotation is unsupervised labeling [6,7] which estimates the joint density of visual features and words by an unsupervised learning algorithm. These methods introduce a hidden variable and assume that features and words are independent given the value of hidden variable. Another formulation for statistical annotation is supervised multi-class labeling (SML) [5,8] that estimates a conditional distribution for each semantic class to determine probability of a feature vector given the semantic label.

Regardless of learning strategy, training dataset plays a major role in developing annotation systems. Manual assignment of too many words to a large number of images in the dataset is highly time consuming and labor intensive. On the other hand, a classifier may have poor generalization when some semantic labels have a few images. It seems essential to develop an annotation system that depends on a small number of labeled images in the training phase.

Although it is difficult to prepare an annotated dataset, one could easily obtain unlabeled images in large quantity (e.g., using photo sharing communities on the Internet like Flicker and ImageNet). This large amount of pictures motivates image annotation systems to increase their generalization by incorporating unlabeled images into

* Corresponding author. Tel.: +98 21 6616 6618; fax: +98 21 6601 9246.
E-mail addresses: s_amiri@ce.sharif.edu (S. Hamid Amiri), jamzad@sharif.edu (M. Jamzad).

http://dx.doi.org/10.1016/j.patcog.2014.07.012
0031-3203/

the training phase. To reach this aim, semi-supervised learning (SSL) [9] is emerged in machine learning community. Semi-supervised generative models [10,11] and graph-based methods [12] are two main classes of the SSL that have recently received lots of interests, especially in image annotation [13–20].

A large number of semi-supervised annotations utilize graph-based learning to infer tags of unlabeled images. The main challenging issue in these methods is graph construction. Liu et al. [13] proposed a method called nearest spanning chain (NSC) which constructs the learning graph using chain-wise statistical information instead of the traditional pair-wise similarities. Besides constructing multiple NSCs that is computationally expensive, they do not show how the number of labeled images affects the annotation results. The authors in Ref. [14] focused on graph construction in the presence of noisy annotations. To this end, by solving an optimization problem, a sparse graph is constructed and the training process is modified to handle noisy annotations. A disadvantage of this approach is that edges of the sparse graph are considered to be a subset of edges in a kNN graph. The method in Ref. [15] constructs a bi-relational graph that comprises both visual features and semantic labels of images. By propagating the words of annotated images over bi-relational graph, annotations of unlabeled images are extracted.

To further refine the annotation results, there are some works that combine graph-based learning with other techniques. Tang et al. [16] proposed a method to combine graph-based learning with multiple instance learning [17] for image annotation. After constructing two graphs based on multiple and single instance representations, the graphs are integrated into a unified graph for the learning process. This method utilizes a simple weighted-sum rule for integration. It suffers from early fusion problems in multi-modal representation [18]. Shao et al. [19] presented a framework to combine graph-based learning with a probabilistic model for learning latent topics of images. In this framework, a hidden variable is associated to each image and probabilities of hidden variables are considered to reside on a manifold. This approach suffers from the limitations of unsupervised labeling discussed in Ref. [5]. Zhu et al. [20] proposed a technique which utilizes graph-based learning to refine the candidate annotations obtained from the progressive relevance-based method [21]. Since the candidate annotations are obtained by propagating words from labeled images to unlabeled ones, noisy annotations will highly degrade the performance of this approach.

In addition to the above limitations, there are two main problems that should be considered in graph-based image annotation. First, these methods are transductive and can only predict labels for specific unlabeled samples observed in the training phase. To annotate a new test image, we must add the image to the unlabeled set and run the training phase again. The second problem is related to the required memory and time complexity of graph-based approaches. The primary bottleneck of these approaches comes from the complexity of handling the adjacency matrix of graph which super-linearly grows by increasing the number of images.

It is shown that annotation could be performed in a short time when generative models are utilized in SML formulation [8]. Additionally, generative models are inductive and can predict the label of every sample in the feature space. Thus, in this work, we focus on the semi-supervised generative models for image annotation. We follow PSU protocol [5] for developing the system which is previously utilized in a supervised system called ALIPR [8].

1.2. Overview of ALIPR

ALIPR utilizes Corel60k [7] dataset whose images are organized based on PSU protocol [5]. This protocol assumes that images are divided into distinct categories or “concepts”. Each concept contains a set of words describing the entire images in that concept even though some of these words do not occur in each individual image. It is possible that two different concepts could share some words in their descriptions. Corel60k is comprised of 599 concepts with about 100 images in each concept. This dataset also includes total of 417 distinct words.

With the above structure, ALIPR constructs a generative model for each concept as follows:

(1) Extracting the color and texture signatures of images.
(2) Partitioning images of a concept using a new clustering algorithm named D2-clustering and extracting a prototype for each cluster.
(3) Computing the distance between each prototype and the signatures assigned to it.
(4) Fitting Gamma distribution into each cluster based on the distances in that cluster.
(5) Combining the distribution components to construct a mixture model for each concept.

The advantage of ALIPR over the prior statistical systems such as [5–7] was its higher speed in annotation procedure.

1.3. Contributions of this paper

The current study targets to develop a semi-supervised annotation system to include unlabeled images in concept modeling. To this end, it covers the following contributions:

- We propose a new feature extraction strategy that obtains color and texture signatures in less time than ALIPR. Besides, the new features provide more discriminative property.
- We embedded spectral clustering in the prototype extraction to overcome limitations of ALIPR. Indeed, D2-clustering in ALIPR is a divisive hierarchical clustering and has two limitations. First, since a set of linear optimization problems with a large number of variables must be solved in each level of clustering, the prototype extraction is computationally expensive. Second, the division process utilizes a greedy strategy which could not lead to well-structured clusters.
- To use unlabeled images in the training phase, the formulation of prototype extraction is modified by assigning a weight to each signature that indicates the membership degree of the signature to one concept.
- Given an initial model for each concept, we incorporate unlabeled images into the training phase to improve models through modifying parameters of the clusters. Thus, a new parameter estimation method is presented which utilizes the unlabeled images.

The last two contributions provide semi-supervised annotation and are considered as main contributions of the proposed solution. On the other hand, the first two contributions are not the main contributions but are mandatory for achieving good efficiency in the annotation system.[1]

In our approach, we assume that training dataset follows PSU protocol. However, in our experiments, we will also discuss how our approach can be applied to datasets that are not organized based on PSU protocol. In fact, the PSU protocol assumption in the training phase does not weaken the applicability of our approach to other datasets.

[1] An early version of supervised framework [24] explains the feature extraction and clustering phases.