J. Vis. Commun. Image R. 55 (2018) 816–828

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Leveraging multi-modal fusion for graph-based image annotation ☆

S. Hamid Amiri [a,*], Mansour Jamzad [b]

[a] Department of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran
[b] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 9 November 2017
Revised 12 June 2018
Accepted 16 August 2018
Available online 22 August 2018

Keywords:
Image annotation
Tag
Manifold
Multi-modal representation
Graph-based learning
Supergraph

ABSTRACT

Considering each of the visual features as one modality in image annotation task, efficient fusion of different modalities is essential in graph-based learning. Traditional graph-based methods consider one node for each image and combine its visual features into a single descriptor before constructing the graph. In this paper, we propose an approach that constructs a subgraph for each modality in such a way that edges of subgraph are determined using a search-based approach that handles class-imbalance challenge in the annotation datasets. Multiple subgraphs are then connected to each other to have a supergraph. This follows by introducing a learning framework to infer the tags of unannotated images on the supergraph. The proposed approach takes advantages of graph-based semi-supervised learning and multi-modal representation simultaneously. We evaluate the performance of the proposed approach on different datasets. The results reveal that the proposed approach improves the accuracy of annotation systems.

1. Introduction

Due to the proliferation of digital image databases and photo-sharing websites, there is an essential need for proper searching techniques in image collections. To organize large image datasets, a popular approach is to assign some tags to images, describing the contents of images. These tags provide an effective way to access large scale image datasets. This strategy also allows the users to perform image retrieval based on the input texts.

To automatically assign semantic labels to images, machine vision community are developing annotation systems. An annotation system could assist the users in labeling, retrieval and categorization of images. Generally, these systems use machine learning techniques to model the relations between visual features of images and semantic labels. The success of such systems depends on the availability of a fully-labeled dataset with large number of images for each tag. However, preparing such dataset is time-consuming and labor-intensive. To alleviate this problem, recent annotation systems utilize semi-supervised learning [1] to incorporate unannotated images into the training phase and thus reduce the demand of systems on fully-labeled datasets.

Among semi-supervised learning techniques, graph-based approaches have received superior attentions in the recent years. These approaches rely on the manifold assumption [2] which means that labels of input samples must smoothly change with respect to the manifold structure. In the graph-based learning, this structure is approximated by a graph with nodes corresponding to data samples including labeled and unlabeled ones. Also, an edge in graph represents the similarity of two nodes. Based on the manifold assumption, two nodes must have similar labels if they are connected to each other by an edge with a large weight. Learning on the similarity graph is performed by propagating labels of the labeled samples to unlabeled ones. Popular algorithms for label propagation are Gaussian random field, harmonic functions approach [3], local and global consistency method [4], Laplacian regularization [5] and linear neighborhood propagation [6].

Conventional approaches for graph-based annotation [7–10] consider one node for every image on a data graph. Given the similarity between every pair of images, the weights of edges on the graph are determined directly from similarities (e.g., kNN and full strategy) or after postprocessing on the similarities (e.g., sparse representation [8] or nearest spanning chain [7]). With this strategy, various visual features extracted from an input image are mapped to one node in the data graph. Considering each of the visual features such as color, texture and shape as an individual

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: s.hamidamiri@sru.ac.ir (S.H. Amiri), jamzad@sharif.edu (M. Jamzad).

https://doi.org/10.1016/j.jvcir.2018.08.012
1047-3203/

modality, a multi-modal representation is provided for the images. In the above graph-based annotation strategies, multiple visual features of images are fused at *early or feature level*, without taking the advantages of multi-modal representation of images [11].

While it is straightforward to combine visual features at the feature level, the integrated descriptor does not maintain the structures of individual modalities. This issue plays an important role in graph-based learning techniques which are based on the manifold assumption. More precisely, visual features are different in their natures and the feature vectors of each visual modality reside on a manifold whose structure could be significantly different from that of other modalities. In feature level fusion, all these manifolds are replaced with a single manifold without preserving the structures of individual manifolds.

In this paper, we propose an approach that efficiently integrates different modalities of images (i.e. feature vectors) on the learning graph. This approach constructs a supergraph that is formed of individual subgraphs for each modality to separately approximate the structure of manifold associated to corresponding modality. The subgraphs are connected to each other by some edges to form a *supergraph* that structurally combines multiple manifolds of different modalities.

From multi-modal representation viewpoint, learning on the supergraph provides fusion at the *decision level*. In this type of fusion, multiple decisions (labels) are obtained for input samples through multiple classifiers (subgraphs) which are trained on different modalities (visual features). Then, these local decisions are fused into a single decision for final classification. As discussed in [11], decision level fusion has considerable advantages over feature level fusion.

Recently, we study multi-modal representation for supergraph of prototypes [12] where the supergraph is constructed from prototypes extracted from the feature vectors of images in each modality. The difference of this work is that we construct the supergraph directly from feature vectors of images in such a way that challenges of image annotation problem (e.g. class-imbalance, weak-labeling and etc. [13]) are considered in formation of subgraphs. Thus, in spite of previous graph-based image annotation methods, our graph construction approach is well customized for annotation problem. More specifically, we leverage the methodology used in the state-of-the-art search-based methods [13] to extract nearest images for a given image in each modality. Then, edges of each subgraph will be specified using the customized nearest neighbors approach. In this way, we achieve more efficient subgraph structure which has a great effect on the performance of label propagation step.

Besides the above contribution, we also study the benefits of label propagation on the supergraph theoretically. In this line, we will prove that learning process will maintain labels smoothness constraint on each individual subgraph based on its structure. In addition to this property, we will prove that label propagation on supergraph will consider the correlation among labels on different subgraphs. In other words, while each classifier (subgraph) maintains smoothness constraint, it correlates closely with other subgraphs in the label propagation process. This properties indicate that supergraph is an efficient integrated classifier.

The rest of the paper is organized as follows. Section 2 reviews the related work on image annotation. In Section 3, we define the notations used in this paper and present an overview of the proposed approach. Section 4 presents the methodologies for combining different modalities of images on the learning graph. Section 5 discusses the learning process. Section 6 describes the assignment of labels to unlabeled images. Section 7 presents experimental results and Section 8 concludes the paper.

2. Related work

In this section, we briefly review some of the existing methods for image annotation and discuss their pros and cons. According to [14], automatic image annotation methods can be divided into three main category. The first category uses generative models [15,16] to define a joint distribution over visual features and semantic labels. The methods in the second category are based on discriminative models [17,18] that treat each semantic label as an independent class and train a classifier for each class to determine the presence/absence of the label in a test image. The third category includes search-based approaches [19,14,13, 20–22] that extract the labels of each test image from its visually similar images in the training dataset. These approaches are based on this assumption that similar images share common labels and in most of them, a large number of visual descriptors are extracted from images to guarantee that relevant images only transfer their labels to a test image. Furthermore, the approaches in all categories assume that a complete list of relevant tags is available for each image. However, this assumption does not hold in most of annotation datasets. To alleviate this problem, semi-supervised learning methods are used to leverage unlabeled images, in addition to the labeled ones, for discovering the semantic labels of images.

During the past decade, there have been some efforts to use graph-based semi-supervised learning for image annotation. Since the structure of learning graph plays an important role in graph-based learning, many researchers have focused on developing efficient methods for construction of the learning graph from images. Tang et al. [8] proposed an approach for graph construction in the presence of noisy annotations. More specifically, a sparse graph is constructed in this method where each node of graph corresponds to feature vectors of an image. To specify the edges of graph, sparse representation [23] was utilized to reconstruct each sample (feature vector) from its kNN samples. Since each sample is obtained by concatenating multiple features of an image, sparse coefficients may not properly represent similarities of images. Wang et al. [9] constructed a bi-relational graph that is comprised of two subgraphs for visual features and semantic labels. For each label, a semantic group is formed on the graph using images assigned to it. To infer the labels of unlabeled images, an algorithm was suggested to propagate the tags of annotated images over semantic groups. Tang et al. [10] presented a methodology in which two graphs are constructed using multiple and single instance representations of images. The two graphs are then integrated into a unified graph for learning process. A disadvantage of this approach is that it uses a simple weighted-sum rule for integration.

The above approaches for graph-based annotation consider a single node for each image in the learning graph. Since each image is described by various visual features with different natures, the edges of learning graph may connect dissimilar images to each other and therefore label propagation on the graph will result into irrelevant tags for the unlabeled images. In this paper, we address this problem by constructing a specific graph for each type of visual features and integrating them into a supergraph. The details of proposed approach will be explained in the following.

3. Preliminaries and notations

In this section, we introduce the notations used in the rest of paper and overview the proposed approach. In the following, the terms label and tag are used interchangeably.

We assume that there are n images which are annotated with K different labels. Suppose that $I = \{I_1, \ldots, I_n\}$ is the set of images such that feature vectors of I_i are represented as