Neurocomputing 168 (2015) 853–860

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-view based multi-label propagation for image annotation

Zhanying He[a], Chun Chen[a,*], Jiajun Bu[a], Ping Li[a], Deng Cai[b]

[a] *Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China*
[b] *State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310058, China*

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form
25 March 2015
Accepted 8 May 2015
Communicated by M. Wang
Available online 19 May 2015

Keywords:
Image annotation
Multi-view
Multi-label

ABSTRACT

Multi-view learning and multi-label propagation are two common approaches to address the problem of image annotation. Traditional multi-view methods disregard the consistencies among different views while existing algorithms toward multi-label propagation ignore the underlying mutual correlations among different labels. In this paper, we present a novel image annotation algorithm by exploring the heterogeneities from both the view level and the label level. For a single label, its propagation from one view should agree with the propagation from another view. Similarly, for a single view, the propagations of related labels should be similar. We call the proposed approach as Multi-view based Multi-label Propagation for image annotation (MMP). MMP handles the consistencies among different views by requiring them to generate the same annotation result, and captures the correlations among different labels by imposing the similarity constraints. By taking full advantage of the dual-heterogeneity from views and labels, MMP is able to propagate the labels better than state of the art. Furthermore, we introduce an iterative algorithm to solve the optimization problem. Extensive experiments on real image data have shown that the proposed framework has effective image annotation performance.

1. Introduction

In recent years, with the exponential increasing of digital cameras, people are overwhelmed by a huge number of accessible images, whereby most of them are unlabeled. To effectively manage, access and retrieve this multimedia data, a widely adopted solution is to associate textual annotations to the semantic content of images. With the annotations, an image retrieval problem can be converted into a text retrieval problem, which enjoys both efficient computation and high retrieval accuracy [1]. Since manual annotation is usually time consuming and tedious, semi-supervised multi-label propagation lends itself as an effective technique, through which users only need to label a small number of images, and other unlabeled image can work together with these labeled image for learning and inference [2].

In general, one important step of an automated image annotation task is to extract visual features for image representation [3]. However, we can obtain heterogenous features (multiple views) from images. Different kinds of features describe various aspects of image's visual characteristics and have different discriminative power for image understanding [4]. Numerous studies have been devoted to the multi-view based image annotation problem, but disregard the consistencies among different views. Although some sparsity-based approaches have been studied on the selection of heterogenous image features, they only combine several types of features into a single "big" view [5,4].

The following step of an automated image annotation task is to associate each unlabeled image with a number of different given labels. But most existing work in the line of multi-label propagation suffer (or partially suffer) from the disadvantage that they consider each label independently when handling the multi-label propagation problem [2,6].

To the best of our knowledge, there does not exist an effective annotation method to fully explore both the view heterogeneity and the label heterogeneity simultaneously. Therefore, we present a novel multi-view based framework for multi-label propagation (MMP) to bridge multi-view learning and multi-label propagation together. The central idea is that: (1) the label propagation from one view should agree with the propagation from another view; (2) the propagations of related labels should be similar. Specially, for each view, MMP models the relationships between the images and the features by constructing a bipartite graph [7,8], and models the manifold structure among images by using graph Laplacian [9,10]. Thus, both the image-feature relationships and the geometrical structure are captured by minimizing the fitting error. Given a label, different views should generate the same annotation result so that the consistencies among different views are handled. And we believe that the information of related labels can help to improve the annotation performance. So we impose

[*] Corresponding author.
E-mail addresses: hezhanying@zju.edu.cn (Z. He), chenc@zju.edu.cn (C. Chen), bjj@zju.edu.cn (J. Bu), lpcs@zju.edu.cn (P. Li), dengcai@cad.zju.edu.cn (D. Cai).

http://dx.doi.org/10.1016/j.neucom.2015.05.039
0925-2312/© 2015 Elsevier B.V. All rights reserved.

the similarity constraints between related labels to capture the correlations among different labels. Combining the overall consistencies among views and the similarity of related labels, MMP solves the complex problem with the heterogeneities from both the view level and the label level. Furthermore, we introduce an iterative algorithm to solve the optimization problem.

It is worthwhile to highlight the following contributions of our proposed MMP algorithm in this paper.

- We propose a novel image annotation method which fully explores both the view heterogeneity and the label heterogeneity simultaneously. The proposed algorithm handles the two types of heterogeneities by requiring that: (1) the label propagation from one view should agree with the propagation from another view; (2) the propagations of related labels should be similar.
- Though this study is mainly motivated by the previous work in [11], we concern about the real application problem of image annotation rather than the mathematical framework itself. Besides, we introduce image-feature, inter-image and inter-label relationships to improve the performance of the proposed Multi-view Based Multi-label Propagation algorithm.
- We introduce an iterative algorithm to solve the optimization problem and calculate its computational cost. We implement the effective experiments on a real image data set and discuss tuning process of all parameters.

The rest of this paper is organized as follows: Section 2 briefly introduces the related work about existing image annotation methods. The proposed method is described in Section 3 including the theoretical formula and the optimization algorithm. We set up the experiments and discuss the performance evaluations in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

In general, image annotation methods can be categorized into three types: free text annotation, keyword annotation and annotation based on ontologies [12]. In this study, we focus on the keyword annotation approach which allows users to annotate images with a chosen set of keywords (“labels”) from a controlled or uncontrolled vocabulary [13]. The keyword based image annotation has attracted a lot of attention from researchers in the last decade [14]. It views labels as the central components for sharing retrieval and discovery of the user-generated content.

Since image representation is one of the most important steps in image annotation, much existing work has studied how to exact visual features. Generally, visual content of an image can be represented by either global or local features. Global features take all the pixels of an image into account. Color histogram [15], for example, can be extracted to represent or describe the global color content of images. On the other hand, local features like SBN [16], SIFT [17] and shape context [18] can provide more detailed information of different parts of an image. In order to solve the task of selection of multiple feature types, Cao [5] proposes to learn different metric kernel functions for different features. They formulate the Heterogeneous Feature Machines (HFM) as a sparse logistic regression by the l_1-norm at group level. Also feature selection is studied to identify the most useful dimensions of features [19] and different image representations are explored to improve the performance of image annotation [20,21]. Liu et al. propose partially shared latent factor (PSLF) learning to jointly exploit both consistent and complementary information among multiple views [22]. Integrating image-word correlation, image similarity and word relation together, a multi-correlation probabilistic matrix factorization (MPMF) algorithm is proposed for the correlation estimation in image annotation [23]. Even in the literature of video annotation, researchers are aiming to simultaneously tackle different difficulties, such as insufficiency of training data and the curse of dimensionality, in a unified scheme. Wang et al. represent various crucial factors by different graphs and simultaneously deal with them by learning with multiple graphs [24]. Moreover, Multi-label Boosting by the selection of heterogeneous features with structural Grouping Sparsity (MtBGS) is implemented in [4] to induce a structural sparse selection model to identify subgroups of homogenous features for predicting a certain label. But they combine multiple features into a single “big” view and disregard the consistencies among different views. So we propose a new approach on the idea that the label propagation from one view should agree with the propagation from another view.

Several algorithms have exploited the relationships among different labels to improve the annotation performance [25]. For example, Liu et al. utilize constrained nonnegative matrix factorization (CNMF) to optimize the consistencies between image similarity and label similarity [26]. Qi et al. propose a unified Correlative Multi-Label (CML) framework to simultaneously classify labels and model correlations between them [27]. Chen et al. [28] formulate this problem as a sylvester equation, which is similar to the work in [29]. Tang et al. use KNN to explore the relationships among noisily tagged web images [30]. Gong et al. implement CCA to explore in a three-view embedding space and presented both unsupervised and supervised training ways [31]. Unfortunately, none of them consider utilizing the multi-view characteristics during the multi-label propagation process. Motivated by this, we propose the MMP framework to bridge multi-view learning and multi-label propagation together.

Notation: Small letters (e.g. x) denote scalars. Lowercase bold letters (e.g. $\mathbf{x}$) denote column vectors and $\|\cdot\|$ denotes the vector l_2-norm. Uppercase letters (e.g. X) denote matrices or graphs. The matrix trace is denoted by $\mathrm{Tr}(\cdot)$ and the Forbenius norm of a matrix is denoted by $\|\cdot\|_F$. Script uppercase letters (e.g. $\mathcal{X}$) denote ordinary sets and $|\mathcal{X}|$ is the size of the set. Blackboard bold capital letters (*e.g.* $\mathbb{R}$) denote number sets.

3. The proposed framework

Suppose we have z labels and v views. Each view denotes a type of feature (e.g., color histogram or SIFT). For the jth view, there are d_j features, namely, the feature dimension is d_j. Suppose we have n images. We use $\mathbf{x}_j^s$ to denote the sth image in the jth view. Then we construct a $n \times d_j$ non-negative matrix $X_j = [\mathbf{x}_j^1, \mathbf{x}_j^2, \ldots, \mathbf{x}_j^n]^T$ whose rows are images and columns are features.

To be specific, for the ith label, we define $\mathbf{g}_i(s) > 0$ to indicate that the sth image is positive with the ith label and vice versa. Similarly, for the ith label and the jth view, we define $\mathbf{f}_{ij}(k) > 0$ to indicate that the kth feature of the jth view is positive with the ith label. Under the semi-supervised learning, suppose we have known $m_i (m_i \ll n)$ labeled images which are positive or negative with the ith label. Then our goal is to leverage the label information from all the labels to help annotate the remaining unlabeled images with each label, as well as to use the consistencies among different views and the correlations across related labels to improve the performance.

3.1. Modeling the image-feature relationships

Before we start to model the relationships between the images and features, we normalize X_j to obtain

$$X_j^N = (D_j^N)^{-1/2} X_j (D_j^F)^{-1/2} \tag{1}$$