Pattern Recognition 48 (2015) 498–508

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning descriptive visual representation for image classification and annotation

Zhiwu Lu [a,*], Liwei Wang [b]

[a] School of Information, Renmin University of China, Beijing 100872, China
[b] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 7 September 2013
Received in revised form
20 July 2014
Accepted 11 August 2014
Available online 20 August 2014

Keywords:
Image classification
Image annotation
Visual representation
Matrix factorization

ABSTRACT

This paper presents a novel semantic regularized matrix factorization method for learning descriptive visual bag-of-words (BOW) representation. Although very influential in image classification, the traditional visual BOW representation has one distinct drawback. That is, for efficiency purposes, this visual representation is often generated by directly clustering the low-level visual feature vectors extracted from local keypoints or regions, without considering the high-level semantics of images. In other words, it still suffers from the semantic gap and may lead to significant performance degradation in more challenging tasks, e.g., image classification over social collections with large intra-class variations. To learn descriptive visual BOW representation for such image classification task, we develop a semantic regularized matrix factorization method by adding Laplacian regularization defined with the tags (easy to access) of social images into matrix factorization. Moreover, given that image annotation only provides the tags of training images in advance (while the tags of all social images are available), we can readily apply the proposed method to image annotation by first running a round of image annotation to predict the tags (maybe incorrect) of test images and thus obtaining the tags of all images. Experimental results show the promising performance of the proposed method.

1. Introduction

Inspired by the success of bag-of-words (BOW) in text information retrieval, we can similarly represent an image as a histogram of visual words through quantizing the local keypoints or regions within the image into visual words, which is known as visual BOW in the areas of pattern recognition and image analysis. As an intermediate representation, it can help to reduce the semantic gap between the low-level visual features and the high-level semantics of images to some extent. Hence, in the literature, many efforts have been made to apply the visual BOW representation to image classification (one typical task in pattern recognition and image analysis). In fact, the visual BOW representation has been shown to give rise to encouraging results in image classification [1–5].

However, as reported in previous work [6–9], the traditional visual BOW representation has one distinct drawback as follows. That is, for efficiency purposes, the visual vocabulary is commonly constructed for visual BOW generation by directly clustering the low-level visual feature vectors extracted from local keypoints or regions within images, without considering the high-level semantics of images. In other words, the traditional visual BOW representation still suffers from the so-called problem of semantic gap and thus may lead to significant performance degradation in more challenging tasks such as image classification over social collections (e.g. Flickr). Here, it is worth noting that the social images are shared by users in a completely unconstrained way and thus are more difficult to classify with larger intra-class variations (see examples in Fig. 1). In this paper, our main motivation is to propose a new method for learning descriptive visual BOW representation to overcome the aforementioned drawback associated with the traditional visual BOW representation.

Considering that matrix factorization has been successfully applied to image representation [10,11], we develop a semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags (easy to access) of social images. The basic idea is to formulate the problem of learning descriptive visual BOW representation as low-rank matrix factorization (see Fig. 2). We further define Laplacian regularization [12–14] with the tags of images (unlike [11] that directly makes use of the class labels of images) and add this term into the objective function of matrix factorization. Due to the special definition of Laplacian regularization, our new SRMF problem can be solved efficiently based on the label propagation technique proposed in [13]. Although a Laplacian regularized matrix

* Corresponding author. Tel./fax: +86 10 62514562.
E-mail addresses: zhiwu.lu@gmail.com (Z. Lu), wanglw@cis.pku.edu.cn (L. Wang).

http://dx.doi.org/10.1016/j.patcog.2014.08.008
0031-3203/

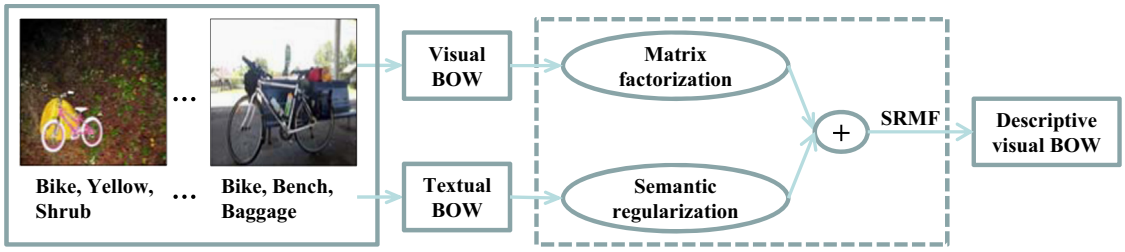

Fig. 1. Illustration of our semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags of images.

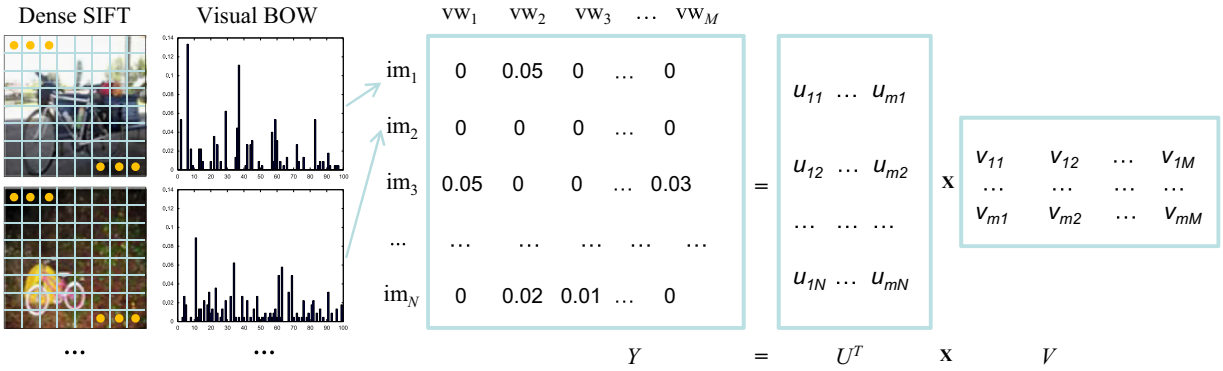

Fig. 2. Illustration of learning descriptive visual BOW representation from a low-rank matrix factorization viewpoint ($m \ll \min(N, M)$). Each column denote a visual word $\mathrm{vw}_j (j = 1, \ldots, M)$, while each row denotes an image $\mathrm{im}_i (i = 1, \ldots, N)$.

factorization method has also been proposed in [10], our SRMF method has two distinct differences as follows: (1) we define Laplacian regularization in this paper mainly to guarantee a good approximation to the original visual BOW representation, while this term is defined in [10] mainly to find a good dimension reduction; (2) we do not consider the nonnegative constraints and thus a sound initialization can be derived from eigenvalue decomposition, while for [10] only a random initialization (which may severely affect the performance) can be provided without any prior knowledge of the original visual BOW representation. Similarly, our SRMF method is also different from [11] in these two aspects.

It should be noted that the tags of images used for learning descriptive visual BOW representation are very easy to access (although noisy) for social image collections (e.g. Flickr). In contrast, the class labels of images used for learning image representation based on nonnegative matrix factorization in [11] are commonly very expensive to obtain in practice. Similarly, it is also very expensive to obtain the constraints with respect to local keypoints, although this kind of high-level semantics has been successfully used for visual vocabulary optimization in [6]. In addition, other than many previous approaches [7–9] to visual vocabulary optimization that have ignored the high-level semantics of images, our SRMF method can explicitly utilize the tags of images (i.e. semantics) to define Laplacian regularization for learning descriptive visual BOW representation.

In summary, we propose a novel semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags of images, which can effectively reduce the semantic gap associated with the traditional visual BOW representation. As illustrated in Fig. 1, the proposed SRMF method consists of two important components for learning descriptive visual BOW representation: matrix factorization over visual BOW representation and semantic regularization with textual BOW representation (derived from the tags of images). Here, it is worth noting that our SRMF method can run very efficiently even on large image datasets. Moreover, due to problem formulation from a matrix factorization viewpoint, our SRMF method can readily deal with the out-of-sample extension issue when a new image is coming.

Although originally developed for image classification over social collections, our SRMF method can be further extended to image annotation [15–17]. For this challenging task, the main difference is that only the tags of training images are provided in advance, while the tags of all images are available for image classification over social collections. To apply our SRMF method to image annotation, we need to first run a round of image annotation to predict the tags (maybe incorrect) of test images. Once we have obtained the tags of all images, we can perform the same SMRF method to learn descriptive visual BOW representation for image annotation. Although the predicted tags of test images are also used as inputs by the traditional image annotation refinement [18–20], we actually focus on visual representation refinement (by SMRF) for image annotation in this paper.

Finally, to emphasize our main contributions, we summarize the following distinct advantages of our SRMF method for learning descriptive visual representation:

- This is the first attempt to develop a semantic regularization matrix factorization method to learn descriptive visual representation for image classification and annotation.
- Our SRMF method is shown to achieve promising results in both image classification and annotation tasks. More notably, when the global visual features are also considered, we can obtain the best results so far in the literature, to the best of our knowledge.
- Our SRMF method can readily deal with the out-of-sample extension issue, i.e., we can learn descriptive visual representation for the newly coming images very efficiently.
- Besides image classification and annotation, our SRMF method can be applied to many other challenging tasks in pattern recognition and image analysis.

Upon our short conference version [21], the present work has made the following additional contributions: (1) out-of-sample