Neurocomputing 173 (2016) 373–384

Contents lists available at ScienceDirect

Neurocomputing

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

Image classification by visual bag-of-words refinement and reduction

Zhiwu Lu [a,*], Liwei Wang [b], Ji-Rong Wen [a]

[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 1 June 2014
Received in revised form
23 September 2014
Accepted 26 January 2015
Available online 3 September 2015

Keywords:
Image classification
Visual BOW refinement
Visual BOW reduction
Graph-based method
Semantic spectral clustering

ABSTRACT

This paper presents a new framework for visual bag-of-words (BOW) refinement and reduction to overcome the drawbacks associated with the visual BOW model which has been widely used for image classification. Although very influential in the literature, the traditional visual BOW model has two distinct drawbacks. Firstly, for efficiency purposes, the visual vocabulary is commonly constructed by directly clustering the low-level visual feature vectors extracted from local keypoints, without considering the high-level semantics of images. That is, the visual BOW model still suffers from the semantic gap, and thus may lead to significant performance degradation in more challenging tasks (e.g. social image classification). Secondly, typically thousands of visual words are generated to obtain better performance on a relatively large image dataset. Due to such large vocabulary size, the subsequent image classification may take sheer amount of time. To overcome the first drawback, we develop a graph-based method for visual BOW refinement by exploiting the tags (easy to access although noisy) of social images. More notably, for efficient image classification, we further reduce the refined visual BOW model to a much smaller size through semantic spectral clustering. Extensive experimental results show the promising performance of the proposed framework for visual BOW refinement and reduction.

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 within the image into visual words, which is known as visual BOW in the areas of image analysis and computer vision. As an intermediate representation, the visual BOW model can help to reduce the semantic gap between the low-level visual features and the high-level semantics of images to some extent. Hence, many efforts have been made to apply the visual BOW model to image classification. In fact, the visual BOW model has been shown to give rise to encouraging results in image classification [1–4]. In the following, we refer to the visual words as mid-level features to distinguish them from the low-level visual features and high-level semantics of images.

However, as reported in previous work [5–9], the traditional visual BOW model has two distinct drawbacks. Firstly, for efficiency purposes, the visual vocabulary is commonly constructed by directly clustering the low-level visual feature vectors extracted from local keypoints within images, without considering the high-level semantics of images. That is, although the visual BOW model can help to reduce the semantic gap to some extent, it still suffers from the problem of semantic gap and thus may lead to significant performance degradation in more challenging tasks (e.g., social image classification with larger intra-class variations). Secondly, typically thousands of mid-level features are generated to obtain better performance on a relatively large image dataset. Due to such large vocabulary size, the subsequent image classification may take sheer amount of time. This means that visual BOW reduction becomes crucial for the efficient use of the visual BOW model in social image classification. In this paper, our main motivation is to simultaneously overcome these two drawbacks by proposing a new framework for visual BOW refinement and reduction, which will be elaborated in the following. It should be noted that these two drawbacks are usually considered separately in the literature [6–9,5,10,11]. More thorough reviews of previous methods can be found in Section 2.

To overcome the first drawback, we develop a graph-based method to exploit the tags of images for visual BOW refinement. The basic idea is to formulate visual BOW refinement as a multi-class semi-supervised learning (SSL) problem. That is, we can regard each visual word as a "class" and thus take the visual BOW representation as the initial configuration of SSL. In this paper, we focus on solving this problem by the graph-based SSL techniques [12–14]. Considering that graph construction is the key step of graph-based SSL, we construct a new L_1-graph over images with

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

http://dx.doi.org/10.1016/j.neucom.2015.01.098
0925-2312/

structured sparse representation by exploiting both the original visual BOW model and the tags of images, which is different from the traditional L_1-graph [15–17] constructed only with sparse representation [18,19]. Through semi-supervised learning with such new L_1-graph, we can explicitly utilize the tags of images (i.e. high-level semantics) to reduce the semantic gap associated with the visual BOW model to some extent.

Although the semantic information can be exploited for visual BOW refinement using the above graph-based SSL, the vocabulary size of the refined visual BOW model remains unchanged. Hence, given a large initial visual vocabulary, the subsequent image classification may still take sheer amount of time. For efficient image classification, we further reduce the refined visual BOW model to a much smaller size through spectral clustering [20,21] over mid-level features. A reduced set of high-level features is generated by regarding each cluster of mid-level features as a new higher level feature. Moreover, since the tags of images has been incorporated into the refined visual BOW model, we indirectly consider the semantic information in visual BOW reduction by using the refined visual BOW model for spectral clustering. In the following, our method is thus called as semantic spectral clustering. When tested in image classification, the reduced set of high-level features is shown to cause much less time but with little performance degradation.

In summary, we propose a new framework for visual BOW refinement and reduction, and the system overview is illustrated in Fig. 1. In fact, upon our short conference version [22], we have made two extra contributions (also see Fig. 1): visual BOW refinement, and semantic graph construction for visual BOW reduction. Moreover, the advantages of the proposed framework can be summarized as follows: (1) our visual BOW refinement and reduction are both efficient even for large image datasets; (2) when the global visual features are fused for image classification, we can *obtain the best results so far* (to the best of our knowledge) on the PASCAL VOC'07 [23] and MIR FLICKR [24] benchmark datasets, as shown in our later experiments; (3) although only tested in image classification, our visual BOW refinement and reduction can be extended to other tasks (e.g. image annotation).

The remainder of this paper is organized as follows. Section 2 provides an overview of related work. In Section 3, we develop a graph-based method to explicitly utilize the tags of images for visual BOW refinement. In Section 4, the refined visual BOW model is further reduced to a much smaller size through semantic spectral clustering. In Section 5, the refined and reduced visual BOW models are evaluated by directly applying them to image classification. Finally, Section 6 gives the conclusions.

2. Related work

Since visual BOW refinement and reduction, and structured sparse representation are considered in the proposed framework, we will give an overview of these techniques in the following.

2.1. Visual BOW refinement

In this paper, visual BOW refinement refers to adding the semantics of images to the visual BOW model. The main goal of visual BOW refinement is to bridge the semantic gap associated with the traditional visual BOW model. To the best of our knowledge, there exist at least two types of semantics which can be exploited for visual BOW refinement: (1) the constraints with respect to local keypoints, and (2) the tags of images. Derived from prior knowledge (e.g. the wheel and window of a car should occur together), the constraints with respect to local keypoints can be directly used as the clustering conditions for clustering-based visual BOW generation [5]. However, the main disadvantage of this approach is that the constraints are commonly very expensive to obtain in practice. In contrast, the tags of images are much easier to access for social image collections. Hence, in this paper, we focus on exploiting the tags of images for visual BOW refinement.

Unlike our idea of utilizing the tags of images to refine the visual BOW model and then improve the performance of image classification, this semantic information can also be directly used as features for image classification. For example, by combining the tags of images with the global (e.g. color histogram) and local (e.g. BOW) visual features, one influential work [10] has reported the best classification results so far (to the best of our knowledge) on the PASCAL VOC'07 [23] and MIR FLICKR [24] benchmark datasets. However, when the global visual features (actually much weaker than those used in [10]) are also considered for image classification in this paper, our later experimental results demonstrate that our method performs better than [10] on these two benchmark datasets.

It should be noted that, besides the tags of images, other types of information can also be used to bridge the semantic gap associated with the visual representation. In [25,14,26,1], local or global spatial information is incorporated into the visual representation, which leads to obvious performance improvements in image classification. In [27,28], extra depth information is considered for image classification in the ImageCLEF 2013 Robot Vision Task. In [29], inspired from the biological/cognitive models, a hierarchical structure is learnt for computer vision tasks. Although the goal of these approaches is the same as that of our method, we focus on utilizing the tags of images to bridge the semantic gap in this paper. In fact, the spatial, depth, or hierarchical information can be similarly added to our refined visual BOW model. For example, our refined visual BOW model can be used just as the original one to define spatial pyramid matching kernel [1].

2.2. Visual BOW reduction

The goal of visual BOW reduction is to reduce the visual BOW model of large vocabulary size to a much smaller size. This is mainly motivated by the fact that a large visual BOW model causes sheer amount of time in image classification although it can achieve better performance on a relatively large image dataset. In

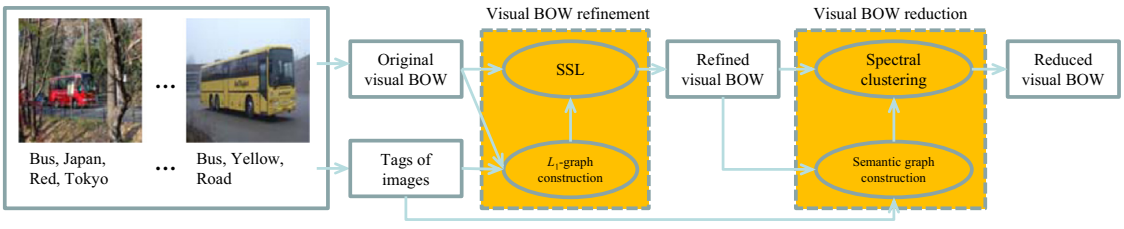

Fig. 1. Illustration of the proposed framework for visual BOW refinement and reduction by exploiting the tags of images. The visual BOW refinement is newly proposed in the present work (i.e. our main contribution), while the visual BOW reduction is mainly proposed in our short conference version [22].