Information Sciences 422 (2018) 271–281

Contents lists available at ScienceDirect

Information Sciences

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

Image-level classification by hierarchical structure learning with visual and semantic similarities

Chunjie Zhang[a,d,*], Jian Cheng[b,c,d], Qi Tian[e]

[a] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[c] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China
[d] University of Chinese Academy of Sciences, 100049, Beijing, China
[e] Department of Computer Sciences, University of Texas at San Antonio TX, 78249, U.S.A

ARTICLE INFO

Article history:
Received 30 June 2017
Revised 6 September 2017
Accepted 8 September 2017
Available online 9 September 2017

Keywords:
Image classification
Hierarchical structure learning
Image-level modeling
Object categorization

ABSTRACT

Image classification methods often use class-level information without considering the distinctive character of each image. Images of the same class may have varied appearances. Besides, visually similar images may not be semantically correlated. To solve these problems, in this paper, we propose a novel image classification method by automatically learning the image-level hierarchical structure (ILHS) using both visual and semantic similarities. We try to generate new representations by exploring both visual and semantic similarities of images. Images are clustered hierarchically to explore their correlations. We then use them for image representations. The diversity of image classes within each cluster is used to re-weight visual similarities. The re-weighted similarities are aggregated to generate new image representations. We conduct image classification experiments on the Caltech-256 dataset, the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset. Experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

Many methods have been proposed to classify images based on their visual contents [18,29,29]. Although these methods have greatly improved classification accuracies, the distinctive character of each image is often ignored. Besides, images may have varied appearances because of class variances. It would be more effective to model image-level information for reliable classification.

Some methods try to model distinctive characters of images for classification [2,42] and detection [35]. However, these methods treat images equally and independently without fully exploring their correlations. The exploration of sub-classes [43] helps to alleviate this problem. However, the number of sub-classes is often predefined. Besides, only visual features are used. The class information should also be used for image-level modeling.

The hierarchical structure has been widely used to model image correlations [15,50]. The structure can either be pre-defined using domain knowledge or automatically learned from the data. For image classification tasks,

* Corresponding author.
E-mail addresses: chunjie.zhang@ia.ac.cn, zhangchunjie1983@gmail.com (C. Zhang), jcheng@nlpr.ia.ac.cn (J. Cheng), qitian@cs.utsa.edu (Q. Tian).

http://dx.doi.org/10.1016/j.ins.2017.09.024
0020-0255/

the learning approach is widely used. However, many of these methods only use the visual information without considering semantic correlations of images. The class information should also be used to improve classification accuracies.

To jointly model visual and semantic correlations of images, in this paper, we propose a novel image-level hierarchical structure learning method for classification. First, we measure similarities of images using visual features and class labels jointly. Training images are then clustered hierarchically to explore their correlations. We calculate visual similarities between images and clusters. Class diversities of clusters are used to re-weight visual similarities. The re-weighted similarities are aggregated to generate image representations. We conduct image classification experiments on the Caltech-256 dataset [8], the PASCAL VOC 2007 dataset [7] and the PASCAL VOC 2012 dataset [6]. Experimental results demonstrate the effectiveness of the proposed method.

The main contributions of this paper lie in three aspects:

- First, we generate discriminative image representations for classification by jointly modeling the visual and semantic correlations.
- Second, a hierarchical structure is automatically constructed to represent images discriminatively.
- Third, the proposed method can be combined with various image representation strategies to further improve classification performances.

The rest of this paper is organized as follows. We discuss the related work in Section 2 and give the details of the proposed image-level hierarchical structure learning method for classification in Section 3. In Section 4, we conduct experiments on the Caltech-256 dataset, the PASCAL VOC 2007 dataset and the PASCAL VOC 2012 dataset. Finally, we give the conclusions in Section 5.

2. Related work

Many methods were proposed for image classifications [14,18,20,29,36,45]. To alleviate the quantization loss of the bag-of-visual-words model (BoW) [45], the sparse coding method [29] was proposed by softly assigning one local feature to a number of visual words. Max pooling was then used to extract the largest response for image representation. The fisher vector [18] was proposed to explore the first-order and second-order information. Instead of extracting local features, the convolutional neural network based method [20] worked directly on image pixels with good performances. However, these methods treated images of the same class jointly. The distinctive character of each image was ignored.

Exemplar based methods were used both for classification [2,42] and detection [10,35]. Zhang et al. [42] modeled contextual relationships of exemplar classifiers and improved classification accuracies. Boiman et al. [2] used local features directly to classify images. Each exemplar classifier was trained to separate one sample from the other samples. However, different exemplar classifiers were not aligned. To alleviate this problem, Zepeda and Perez [35] used exemplar SVMs as visual encoder. To make use of the correlations of images, the sub-class based method [45] was proposed. Zhang et al. [43] also learned general and class-specific codebooks with low-rank constraint. However, how to determine the number of sub-classes was still an open problem. Besides, the hierarchical information was not fully considered by these methods.

Hierarchical representation strategies were also used to model image correlations [15,50]. Li et al. [15] used ImageNet as auxiliary information to hierarchically represent images with pre-defined structure. Zhang et al. [50] proposed a saliency prediction method by learning graphlet hierarchies. Bo et al. [1] tried to solve the multipath sparse coding problem with hierarchical matching pursuit. However, pre-defined hierarchical structure could not capture the specific characters of images. Besides, only the visual information was used while semantic correlations were discarded. Instead of directly using the initial image representations, researchers also made various transformations and combinations [23,37]. Torresani et al. [23] used classesmes for object recognition while Zhang et al. [37] made classification in sub-semantic space.

In recent years, more and more discriminative image representation methods were proposed, e.g., the fisher vector based method [18] and the convolutional neural network (CNN) based methods [5,20,34]. Motivated by these works, many works [4,14,17,26,27] were made. Cinbis et al. [4] leveraged fisher kernels with non-iid patches. Oquab et al. [17] tried to learn mid-level image representations using CNN. Wei et al. [26] proposed a novel method for multi-label classification while Xu et al. [27] explored multi-loss regularized deep networks. Deep neural network base strategies [3,11–13] were also used for human pose analysis with good performances. However, these methods only made use of the visual information, leaving semantic correlations of images unconsidered. Instead of only using the Euclidean distance, other more efficient similarity measurement methods [22,30–32] were also proposed and greatly improved the performances.

Researchers also proposed many efficient methods [19,24,25,28] for classifications. Wang et al. [24] improved the sparse coding algorithm with locality constraint and boosted classification performances. Sermanet et al. [19] combined localization, detection and classification in the deep convolutional neural network framework. Yang et al. [28] targeted object categorization problem using the bounding box information. Zhang et al. [36–49] conducted a series of works on image classifications. Gonzalez et al. [9] used multi-resolution patterns for image classification while Yuan et al. [33] explored high-order local patterns. Many of these methods trained class-level classifiers which could not handle images with large variations. It would be more effective to consider the image-level information.