Information Sciences 369 (2016) 160–170

Contents lists available at ScienceDirect

Information Sciences

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

Boosted random contextual semantic space based representation for visual recognition

Chunjie Zhang[a], Zhe Xue[a], Xiaobin Zhu[b,*], Huanian Wang[c], Qingming Huang[a,d], Qi Tian[e]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
[b] Beijing Technology and Business University, Beijing, China
[c] Central University of Finance and Economics, Changping District, Beijing, China
[d] Key Lab of Intelligent Information Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[e] Department of Computer Sciences, University of Texas at San Antonio, TX, 78249, U.S.A

ARTICLE INFO

Article history:
Received 16 October 2015
Revised 12 May 2016
Accepted 19 June 2016
Available online 21 June 2016

Keywords:
Pattern recognition
Image processing
Visual representation

ABSTRACT

Visual information has been widely used for image representation. Although proven very effective, the visual representation lacks explicit semantics. However, how to generate a proper semantic space for image representation is still an open problem that needs to be solved. To jointly model the visual and semantic representations of images, we propose a boosted random contextual semantic space based image representation method. Images are initially represented using local feature's distribution histograms. The semantic space is generated by randomly selecting training images. Images are then mapped into the semantic space accordingly. Semantic context is explored to model the correlations of different semantics which is then used for classification. The classification results are used to re-weight training images in a boosted way. The re-weighted images are used to construct new semantic space for classification. In this way, we are able to jointly consider the visual and semantic information of images. Image classification experiments on several public datasets show the effectiveness of the proposed method.

1. Introduction

Visual information [35] has been widely adopted for various tasks, e.g. classification [35], segmentation [22] and retrieval [33]. Local feature based strategy is widely used for its efficiency. Usually, local features are encoded with pre-learned codebooks. A number of local features (*e.g.* SIFT [25], HoG [6], SURF [2], MROGH [10] and KAZE [1]) have been proposed whose effectiveness has been proven by researchers.

Although effective, visual feature has no explicit semantic correspondence with human perception. Many methods [7,54,56] only use visual cues for representations. However, due to the semantic gap [37], only using visual information cannot model semantics well. Hence, how to explore the semantic information of images becomes urgent. The state-of-the-art semantic based methods try to solve this problem with training images [4,14,20,23,28,30,43,45,52,55] or using images

[*] Corresponding author.
E-mail addresses: zhangcj@ucas.ac.cn (C. Zhang), xuezhe10@mails.ucas.ac.cn (Z. Xue), brucezhucas@gmail.com (X. Zhu), huanianwang06@gmail.com (H. Wang), qmhuang@jdl.ac.cn (Q. Huang), qitian@cs.utsa.edu (Q. Tian).

http://dx.doi.org/10.1016/j.ins.2016.06.029

from other sources [21,31,39,49,51]. Using the training images can help to generate semantic spaces discriminatively. However, most of them only consider the visual information for semantic space construction. The initially generated semantic space may not be able to represent images well. Besides, it is hard to learn effective classifiers for generic image classes. Moreover, object often exhibits visual polysemy. The leverage of extra information [21,31,39,49,51] can make use of more information along with the training images. However, this approach is still dataset dependent once the images are collected. Besides, the collected images may be improper for the specific task.

To alleviate the semantic gap, attribute based image representation also becomes popular [15–17,36]. Attributes are concepts that can be understood by human being and are also easy to be distinguished by computers. This helps to represent images in an understandable way. However, attributes have to be pre-defined by experts. Besides, there are many concepts that cannot be well represented by pre-defined attributes.

To solve the problems mentioned above, in this paper, we propose a novel boosted random contextual semantic space based image representation method for visual classification (BRCSS). Images are initially represented with the bag-of-visual-words (BoW) model. The semantic space is then generated by random selection. Contextual semantic representations of images are used to model the correlations of different semantics. We then train classifiers for prediction. We use the results in a boosted way by re-weighting images which are then used for semantic space construction. We conduct classification experiments on several public available image datasets, experimental results prove the effectiveness of the proposed method.

The main contributions of this paper lie in three aspects. First, compared with visual feature based image representations, the proposed BRCSS can alleviate the semantic gap by using semantic space based image representations. Second, compared with other semantic space based image representations, BRCSS uses the contextual semantic representations of images in an iterative way by generating semantic spaces with random sampling and re-weighting. Third, the semantic representation is also task dependent.

Although both BRCSS and BCSS [57] use the boosting strategy, they are fundamentally different for three reasons. First, BRCSS explores semantic relationships of exemplar classifiers with mixture Dirichlet distributions. The use of semantic representation helps to alleviate the semantic gap while BCSS only uses visual information for classification. Second, BRCSS re-weights the training images while BCSS treats local features differently. The aim of local feature encoding process focuses on minimizing the reconstruction error. This is different from the classification task which makes the updating of BCSS not as efficient as BRCSS. Third, the performance of BRCSS is better than BCSS not only because of the semantic relationship modeling but also because of the updating of training images.

BRCSS also differs from GraphSC [58] which explores local manifold structure with graph regularized sparse coding. GraphSC concentrates on the efficient usage of visual information while BRCSS explores the correlations of semantic representation. Both BRCSS and GraphSC are able to improve classification performances.

The rest of this paper is organized as follows. We give the related work in Section 2. The details of the proposed boosted random contextual semantic space based image representation for classification method are given in Section 3. To evaluate the effectiveness of the proposed method, we conduct image classification experiments in Section 4. Finally we conclude in Section 5.

2. Related work

Visual information had been widely used for the classification task. Sivic and Zisserman [35] used SIFT features [25] for video retrieval with good performance. There were also many other local features [1,2,6,10] that were proposed. Dalal and Triggs [6] proposed the histograms of oriented gradients (HoG) and applied it for human detection with less computational complexity compared with SIFT feature. The speeded up robust features algorithm (SURF) was proposed by Bay et. al. [2] while Fan et. al. [10] proposed the MROGH feature. Alcantarilla et. al. [1] proposed the KAZE feature. Many works had been done using these features [7,40,42,50,54,56]. Zhang et. al. [54,56] used the SIFT feature for image classification with the sparse coding technique. Datta et. al. [7] used it for image retrieval.

Although the visual based image representation had been proven very effective, it still had one problem. The visual information had no explicit semantic meanings. To solve this problem, many works had been done [4,14,20,21,23,28,30,31,37,39,41,43,45,49,51,52,55,58]. Training images were often used for semantic space construction [4,14,20,23,28,30,43,45,52,55]. Rasiwasia and Vasconcelos [30] used low-dimensional semantic spaces for scene classification while Hauptmann et. al. [14] used high-level concepts for video retrieval. Zhang et. al. [55] used exemplar classifiers for weak semantic space construction and then extended with sub-semantic space [52]. Pereira et. al. [28] used semantic queries for retrieval. Torresani et. al. [43] proposed the classesmes as a way to semantically represent images and applied it for object category recognition. Vogel and Schiele [45] used this technique for content-based image retrieval. Bosch et. al. [4] used a hybrid generative/discriminative approach for scene classification while Li and Perona [20] used a Bayesian hierarchical model. A co-clustering based method was proposed by Liu and Shah [23]. For specific tasks, using the training images for semantic space representation was very efficient. However, this strategy failed when the semantics were hard to classify. Besides, visual polysemy also hindered the discriminative power of the learned semantic representation.

In order to make use of other information along with the training data, researchers also leveraged extra data [21,31,39,49,51]. Li et. al. [21] proposed to construct ObjectBank by collecting images from the Internet while Russell et. al. [31] used LabelMe to make use of human labor for image annotation. Zhang et. al. [51] tried to learn discriminative codebooks using the information of other datasets while Yang et. al. [49] used web images for semantic video indexing. A human