Neurocomputing 156 (2015) 79–85

Contents lists available at ScienceDirect

Neurocomputing

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

Joint image representation and classification in random semantic spaces

Chunjie Zhang [a], Xiaobin Zhu [b], Liang Li [a,*], Yifan Zhang [c], Jing Liu [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] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[d] Key Lab of Intell. Info. 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, USA

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form
22 December 2014
Accepted 29 December 2014
Communicated by Rongrong Ji
Available online 8 January 2015

Keywords:
Image representation
Image classification
Semantic space
Random sampling
Sparse representation

ABSTRACT

Local feature based image representation has been widely used for image classification in recent years. Although this strategy has been proven very effective, the image representation and classification processes are relatively independent. This means the image classification performance may be hindered by the representation efficiency. To jointly consider the image representation and classification in an unified framework, in this paper, we propose a novel algorithm by combining image representation and classification in the random semantic spaces. First, we encode local features with the sparse coding technique and use the encoding parameters for raw image representation. These image representations are then randomly selected to generate the random semantic spaces and images are then mapped to these random semantic spaces by classifier training. The mapped semantic representation is then used as the final image representation. In this way, we are able to jointly consider the image representation and classification in order to achieve better performances. We evaluate the performances of the proposed method on several public image datasets and experimental results prove the proposed method's effectiveness.

1. Introduction

In recent years, the use of local features for image classification is widely used by researchers [1]. Typically, local features are quantized or encoded to get the image representation with spatial pyramid matching [2]. SVM classifiers are then trained to predict image categories. To improve the performance of this strategy, many works have been done either by using various local features [3–7] or by generating more discriminative codebooks with the corresponding encoding strategies [8–14]. However, the image representation of these methods and the classifier training are relatively independent. In other words, the image representation is not task-driven which may not be able to serve the final classification task well.

To overcome this problem, a lot of works have been done. on one hand, the use of local features directly [15,16] has been explored. However, the computation cost is relatively high compared with encoding based methods. Instead of extract pre-defined features, the automatic learning of features is also been studied [17] with hundreds of thousands of parameters to tune. This requires a lot of experiences and is also time consuming. On the other hand, the use of semantic based image representations has also been proposed. Attribute [18–20] is widely used as it depicts some semantic aspects of images for various image classification tasks. One problem with the attribute based image representation is that it has to be pre-defined. Besides, the design of proper attributes is very hard even for experts which limits its generalization ability. Although researchers have also explored the learning of attributes [21] from images directly, the performances are still far from satisfactory.

The construction of semantic spaces from training images either by generative models [22,23] or discriminative models [24–26] have also been studied. The generative models often assume that the images follow some probabilistic distributions and try to map images to the corresponding spaces for semantic representation. The discriminative models try to train various classifiers instead. Basically, the performances of discriminative models are often better than generative models, especially when we do not have too much training images.

* Corresponding author.
E-mail addresses: zhangcj@ucas.ac.cn (C. Zhang), brucezhucas@gmail.com (X. Zhu), liliang2013@ucas.ac.cn (L. Li), yfzhang@nlpr.ia.ac.cn (Y. Zhang), jliu@nlpr.ia.ac.cn (J. Liu), qmhuang@jdl.ac.cn (Q. Huang), qitian@cs.utsa.edu (Q. Tian).

http://dx.doi.org/10.1016/j.neucom.2014.12.083
0925-2312/

Motivated by the success of discriminative models for semantic modeling, we propose to jointly consider the image representation and classification task in random semantic spaces. Images are first represented using local features with sparse coding technique. We then randomly select images from the training set to generate the semantic spaces. For each random selection strategy, we construct the corresponding semantic space and images are then mapped into this semantic space with the learned classifiers. Finally, we train SVM classifiers to predict images' classes in these randomly generated semantic spaces. Since the generated semantic spaces are task dependent, we can combine the image representation and classification task into a unified framework. Besides, by randomly generating a series of semantic spaces instead of using one particular semantic space [25,26], we can model images more adequately and jointly make use of both the visual information and semantic based representation. We evaluate the proposed joint image representation and classification method in random semantic spaces on several public image datasets and compare with several the state-of-the-art methods to show its effectiveness.

Compared with the work of [26], the contributions of the proposed method in this paper lie in three aspects. First, instead of generating the semantic representation and then using it for classification, we jointly consider the image representation and classification into a unified framework. Second, by randomly selecting training images for semantic spaces construction, we can generate a series of semantic spaces which can model the images more adequately than [26] that only uses one semantic space. Third, the classification accuracies of the proposed method are able to outperform [26] on several public datasets.

The rest of this paper is organized as follows. We give the related work in Section 2. In Section 3, we show the details of the proposed joint image representation and classification in random semantic spaces method. The experimental results are given in Section 4 and finally we conclude in Section 5.

2. Related work

The local feature is often used with the bag-of-visual-words (BoW) model for classification [1]. To make use of the spatial layout of local features, spatial pyramid matching (SPM) [2] is used. However, the use of *k*-means clustering and hard assignment causes information loss. To solve this problem, researchers have tried to extract various types of local features [3–7]. Bay et al. [3] proposed to use Speeded up robust features (SURF) while Dalal and Triggs [4] used the histogram of gradients (HoG) for local feature description which can be computed faster than SIFT. In order to make use of the color information, Sande et al. [5] decomposed images into different color channels and extracted SIFT features for each channel. Zhang et al. [6] used a two-dimensional model instead of SIFT features while Ke and Sukthankar [7] imposed PCA transform to the raw SIFT features and proposed PCA-SIFT which requires less storage. Besides, generating more discriminative codebooks for local feature encoding has also been widely studied [8–14]. Gemert et al. [8] used soft assignment with kernels instead of nearest neighbor assignment to reduce the quantization loss. Yang et al. [9] explored the sparse coding technique and experimentally found that sparse coding can be combined with max pooling for image classification. Wang et al. [10] added the neighbor information for sparse coding to speed up computation and improved the performances. Zhang et al. [11] proposed to generate codebook spatially to combine the spatial information during the codebook generation process. Gao et al. [12] used local feature's similarity as constraints to ensure encoded parameters' consistency. Zhou et al. [13] proposed the super vector coding while Zhang et al. [14] combined tilt and orientation consistency with Laplacian sparse coding and improved the final performances. Since the encoding of local features may cause information loss, many works have been done to alleviate this problem. Yang et al. [15] used local features directly by training classifiers for classification while Boiman et al. [16] used simple nearest neighbor information of local features. Besides, the learning of features directly from images was also proposed by Farabet et al. [17] with good performance. However, the computational cost of the above mentioned methods is very high.

To overcome the semantic gap problem caused by sole visual feature based image representation, the use of semantic based image representation is proposed. This strategy can be broadly divided into two categories. The first approach used attribute based image representation with attributes pre-defined [18–20] or learned from the training images [21]. Farhadi et al. [18] used attribute to describe objects while Lampert et al. [19] tried to detect unseen object classes by between-class attribute transfer. Parikh and Grauman [20] interactively constructed a discriminative vocabulary of nameable attributes. Li et al. [21] tried to learn attribute from Internet images and applied it for image classification. The second approach used semantic representation directly by both generative and discriminative models [22–26]. Rasiwasia and Vasconcelos [22] used low-dimensional semantic spaces generated by GMM model for scene classification and then applied it for image retrieval [23]. Malisiewicz et al. [24] used exemplar-SVMs for object detection while Zhang et al. [25] applied it for object categorization. To increase the semantic information, the use of sub-semantic space is proposed [26]. However, the sub-semantic space has to be generated by calculating the eigenvalues of the visual-semantic similarity matrix which is time consuming. Besides, only generating one semantic space may be not able to model the image classes very well, especially when the images have large intra-class variations.

The use of randomness for image classification is also very popular. Zhang et al. [27] found the randomly selected codebook performs as good as the codebook generated by k-means clustering. The random forest was proposed by Breiman [28] for classifier training. Inspired by the random forest, Moosmann et al. [29] used random clustering forests to construct a series of trees for local feature encoding. These encoding parameters are then concatenated to represent images. The binary classifier [30,31] was also used for face recognition with good performances. Kumar et al. [30] used attribute classifiers and 'simile' classifiers to separate faces with reference people. Berg and Belhumeur [31] used the reference set of faces for 'identity-preserving' alignment and then used the outputs of binary classifiers for representation. Although very effective, only using binary classifiers is not enough for generic image classification. Besides, the computational cost of constructing the binary classifiers is high. For n classes of images, $n*(n-1)/2$ classifiers are needed while the proposed method only needs to training $O(n)$ classifiers. Moreover, for generic images, the large inter-class variations also makes it difficult to choose a proper reference set. The use of more discriminative representations [32] is needed for improving the performance or speed up the computation [33].

3. Random semantic space based joint image representation and classification

In this section, we give the details of the proposed joint image representation and classification in random semantic space algorithm. The flowchart of the proposed method is given in Fig. 1.

3.1. Local feature based raw image representation

To take advantage of the discriminative power of local features, we use the local features for raw image representation. This is achieved by encoding local features with sparse coding and then