Pattern Recognition 55 (2016) 58–67

Contents lists available at ScienceDirect

Pattern Recognition

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

Adaptive spatial pooling for image classification

Yinglu Liu [a], Yan-Ming Zhang [a], Xu-Yao Zhang [a], Cheng-Lin Liu [a,b,*]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguan East Road, Beijing 100190, PR China
[b] CAS Center for Excellence in Brain Science and Intelligence Technology, 95 Zhongguan East Road, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 11 February 2015
Received in revised form
15 October 2015
Accepted 27 January 2016
Available online 11 February 2016

Keywords:
Weighted pooling
Spatial layout
Distribution matrix
Image classification

ABSTRACT

In this paper, we propose an adaptive spatial pooling method for enhancing the discriminability of feature representation for image classification. The core idea is to adopt a spatial distribution matrix to define how the image patches are pooled together. By formulating the pooling distribution learning and classifier training jointly, our method can extract multiple spatial layouts of arbitrary shapes rather than regular rectangular regions. By proper mathematical transformation, the distributions can be learned via a boosting-like algorithm, which improves the efficiency of learning especially for large distribution matrices. Further, our method allows category-specific pooling operations to take advantage of the different spatial layouts of different categories. Experimental results on three benchmark datasets UIUC-Sports, 21-Land-Use and Scene 15 demonstrate the effectiveness of our method.

1. Introduction

Pooling is a crucial step in popular image classification methods, such as Bag-of-Visual-Words (BoVW [1,2]) and Convolution Neural Network (CNN [3,4]). It is used to aggregate a set of unordered local features into a vector representation. Based on this representation, discriminative classifiers (such as SVM [5,6], neural networks [7] and boosting [8,9]) can be trained for various classification tasks. Here we use $\mathbf{v} = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ to define a pooling operation, where $\mathbf{x}_i \in \mathbb{R}^d$ $(i = 1, 2, \ldots, m)$ refer to local features within a spatial region $\mathcal{R}e$ and $\mathbf{v}$ denotes the pooled vector. Here, d and m are the dimensionality and the number of local descriptors, respectively. There are two important factors for a pooling operation: One is the operator function f which defines the way to merge the local features, such as average pooling [10], max pooling [11,12] and l_p-norm pooling [13]. The other is the action region $\mathcal{R}e$ and it decides which local features will be selected for pooling. As we know, if $\mathcal{R}e$ covers the whole image as in BoVW methods, the pooled vector $\mathbf{v}$ is invariant with the spatial shifts of $\mathbf{x}_i$ because the spatial relationship is totally ignored within the action scope. This is helpful to tolerate spatial shifts, but it drops discriminative information about the spatial layout, which usually plays a very important role for image classification.

Several methods have been proposed to take advantage of the spatial layout of regions. One representative is the spatial pyramid matching representation (SPM [2]). Essentially, the SPM method partitions images into uniform sub-regions at different levels of resolution, and then applies a pooling operator on these sub-regions separately. The final representation is obtained by concatenating the pooled features of different sub-regions. With the help of spatial information, SPM achieves significantly better performance compared with the BoVW model. However, the spatial partition patterns of SPM need to be predefined, and the number and the style of the spatial partition patterns are very limited, such as the 1×1, 2×2 and 4×4 uniform grids. Some methods [14,15] improve the SPM by adopting abundant random action regions in the image, but these methods often suffer from exhaustive search from a large region pool and the action regions are still constrained to regular shapes. Other methods [16,17] are proposed to pool the features corresponding to foreground and background separately with the help of object detection. More details are discussed in Section 2.

In recent years, weighted pooling has been widely used in order to capture spatial layouts in images more flexibly. By giving each local feature a weight and then representing an image as the weighted sum of local features, the method can extract information from regions of arbitrary shapes (rather than rectangular regions), and define very flexible pooling operators (other than average pooling and max pooling). It is easy to see that both BoVW and SPM are special cases of weighted pooling, with globally uniform weights and rectangular region uniform weights, respectively. The design of weights for pooling is influential to the image classification performance. Harada et al. [18] select weights by maximizing the Partial Least Squares and Fisher criteria, while Huang et al. use multiple Gaussian distributions to depict the spatial structures [19]. Similar to our proposed method, the method of [20] (abbreviated as LSPR) optimizes weights along with the training of classifiers.

[*] Corresponding author. Tel. +86 13811659042; fax: +86 10 8254 4594.
E-mail addresses: ylliu@nlpr.ia.ac.cn (Y. Liu), ymzhang@nlpr.ia.ac.cn (Y.-M. Zhang), xyz@nlpr.ia.ac.cn (X.-Y. Zhang), liucl@nlpr.ia.ac.cn (C.-L. Liu).

http://dx.doi.org/10.1016/j.patcog.2016.01.030
0031-3203/

However, the LSPR models the relationships between the weights and the classifiers using a multi-layer perceptron (MLP), which is computationally expensive because the weights corresponding to different structures need to be optimized simultaneously. The previous methods also have the limitation that the weights for pooling are shared by all categories. This leads to the under-utilization of discriminative spatial information because different categories usually have different spatial layouts.

In this paper, we propose an adaptive spatial pooling (ASP) method for image classification with the objective of overcoming the under-utilization of spatial information in previous methods. Our core idea is to adopt a spatial distribution matrix to define how the image patches are pooled together. It avoids the prior definition of how to partition images or how to design action regions, and learns a flexible pooling scheme on the whole image directly from the training data. We formulate the pooling distribution learning and classifier training into a unified framework and optimize the joint learning problem via a boosting-like algorithm. Compared with existing methods, our method has three advantages: (1) Since the pooling operator is parameterized as a matrix (each column denotes a distribution of patches), our method can extract various spatial layouts of flexible shapes embedded in images; (2) By proper mathematical transformation, our problem is efficiently solved via a boosting-like algorithm, especially for a distribution matrix of large size; (3) Category-specific pooling operator can be learned by discriminative training. This endows more discriminative power to our model. Fig. 1 shows some examples of the distributions learned by our method. It is obvious that the learned distributions reflect the spatial layout of images.

The rest of this paper is organized as follows: Section 2 reviews the related work about pooling; Section 3 introduces the proposed ASP method in detail; Section 4 presents our experimental results on several benchmark datasets. Finally, Section 5 gives concluding remarks.

2. Related work

Bag-of-Visual-Words (BoVW) and Convolutional Neural Network (CNN) are two popular image representation methods for image classification and object recognition. CNN learns image representations by performing convolution and pooling operation alternately on the whole image. It has achieved the state-of-the-art performance on many datasets, such as MNIST [21], NORB [22] and ImageNet [23]. However, it is computationally expensive and needs large training datasets to avoid over-fitting. On the contrary, The BoVW framework, based on hand-craft features, is much cheaper in computation and also achieves good performance on many real-world problems [24,25]. The proposed method is under the BoVW framework and can be combined with CNNs in the future, because the convolution outputs of CNNs can be taken as local features for adaptive pooling. The BoVW framework involves several steps, each with many techniques proposed:

- *Local feature extraction*: In this stage, the patches of interest are located by either sparse sampling or dense sampling, and features are extracted from the sampled regions. In sparse sampling, patches are selected by interest point/region detectors, such as Harris detector [26,27], DoG [28] and MSER [29]. These methods are usually time-consuming and may miss some important regions, however. A simple and popular technique, called dense sampling, is to sample patches with a fixed step and patch size on the whole image. To extract features, we apply hand-craft feature descriptors, such as HoG [30], SIFT [28], SURF [31] and LBP [32], to each patch.
- *Codebook learning*: The codebook in computer vision is analogous to the vocabulary in natural language processing (NLP). It consists of some representative codewords from local features, which can be learned in either unsupervised or supervised manner. While k-means clustering [33] is most widely used for codebook learning, many advanced methods have been proposed for improving the discriminative ability of the codebook [34–37].
- *Encoding*: This step is to map the local features from the original feature space to a new space describing the weights of codewords. Accordingly, the dimensionality after encoding for each local patch equals the number of codewords. A simple coding scheme is the hard coding (HC [1]), which encodes each local feature with the most similar codeword. Unlike the HC, many advanced coding methods make full use of the codebook, such as the soft coding [38,39], sparse coding [12], local linear coding [40] and salient coding [41].
- *Pooling*: This is an operation to aggregate the codes of local patches into a vector representation of the image. Since our work improves the classification performance by proposing a new pooling method, we discuss the existing pooling methods in more details below.

The idea of feature pooling dates back to the research in 1960s [42]. Huber et al. discovered in the cat's visual cortex that the responses of high complex cells which receive signals from simple cells are insensitive to small spatial shift. This inspired the pooling operation widely used in vision recognition systems [43,44]. Many previous works aimed at finding a good pooling operator. The most popular operators are the average pooling and the max pooling. The average pooling [10] takes the average value of all local features **x** within a region as the pooled feature **v**, usually used along with hard coding. In max pooling [11], each dimension of **v** is the maximum value of the corresponding dimension of set of **x** in the region. Many advanced coding schemes, such as sparse coding [12] and localized soft coding [39], are combined with max pooling and have achieved considerable performance gain.

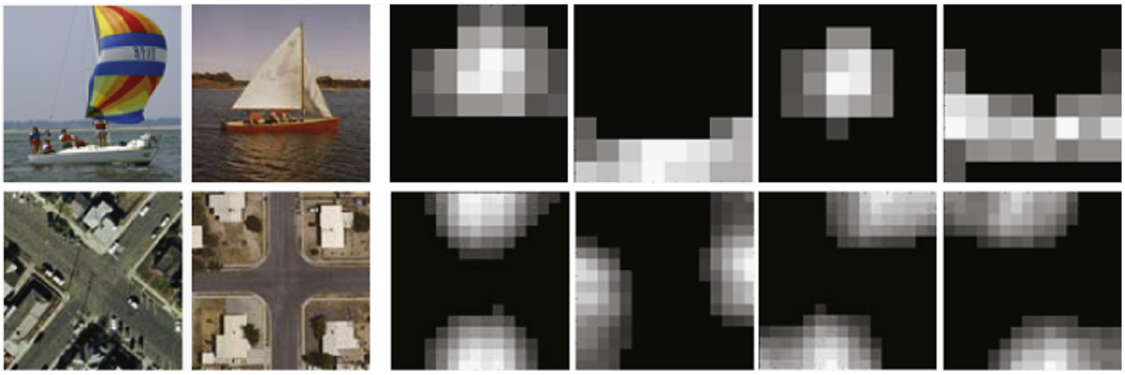

Fig. 1. Visualization of some learned distribution matrices. Each row stands for one specific category: two original images and four learned pooling distributions.