Information Sciences 348 (2016) 209–226

Contents lists available at ScienceDirect

Information Sciences

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

# Scene classification using local and global features with collaborative representation fusion

Jinyi Zou[a], Wei Li[a,*], Chen Chen[b], Qian Du[c]

[a] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
[b] Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA
[c] Department of Electrical and Computer Engineering, Mississippi State University, MS 39762, USA

ARTICLE INFO

*Article history:*
Received 2 November 2015
Revised 4 February 2016
Accepted 8 February 2016
Available online 13 February 2016

*Keywords:*
Scene classification
Locality-constrained linear coding
Spatial pyramid matching
Collaborative representation-based classification

ABSTRACT

This paper presents an effective scene classification approach based on collaborative representation fusion of local and global spatial features. First, a visual word codebook is constructed by partitioning an image into dense regions, followed by the typical *k*-means clustering. A locality-constrained linear coding is employed on dense regions via the visual codebook, and a spatial pyramid matching strategy is then used to combine local features of the entire image. For global feature extraction, the method called multiscale completed local binary patterns (MS-CLBP) is applied to both the original gray scale image and its Gabor feature images. Finally, kernel collaborative representation-based classification (KCRC) is employed on the extracted local and global features, and class label of the testing image is assigned according to the minimal approximation residual after fusion. The proposed method is evaluated by using four commonly-used datasets including two remote sensing images datasets, an indoor and outdoor scenes dataset, and a sports action dataset. Experimental results demonstrate that the proposed method significantly outperforms the state-of-the-art methods.



## 1. Introduction

In the last decade, scene classification has drawn increasing attention both in academia and industry [14,37,45,46,57,63]. The task is to automatically classify an image by feature extraction and label assignment. Although great effort in extracting features (e.g., hash codes [10,15,17], manifold structures [24,56,58], etc) has been made, it is still a challenging task due to many factors to be considered such as variations in spatial position, illumination, and scale.

In the early days, scene classification methods mainly concentrated on modeling [25] and using global spatial features such as color and texture histograms [41]. The global features often have simple implementation and low computational cost but offer limited performance. In [11,36,48,50,59,61], the popular bag-of-visual-words (BoVW) model was adopted, which represented an image with an orderless collection of local features. In this model, an image can be treated as a document, similar to "words". The image is usually partitioned into patches and represented by a codebook. To this end, it follows three steps: (i) feature detection (commonly, the key point detection [30,31] is used.), (ii) feature description based on key points [9,27], and (iii) codebook generation. However, the BoVW model ignores the spatial layout of features.

* Corresponding author. Tel.: +86 10 64413467, +86 18146529853; fax: +86 10 64434726.
*E-mail address:* liwei089@ieee.org, leewei36@gmail.com (W. Li).

http://dx.doi.org/10.1016/j.ins.2016.02.021
0020-0255/

To overcome this issue, spatial pyramid matching (SPM) was developed based on approximate global geometric correspondence in [16]. This approach partitions an image into increasingly fine sub-regions and computes histograms of local features in each sub-region. It is a simple extension of an orderless bag-of-features image representation and overcomes the shortcoming of the BoVW model. Based on this SPM framework, many extensions have been proposed. In [52], an improved SPM method was developed by generalizing vector quantization (VQ) to sparse coding, followed by multiscale spatial max pooling. A linear SPM kernel based on Lowes scale invariant feature transform (SIFT) [9,27] sparse codes was employed, providing excellent performance on several scene datasets. However, this method exhibits high computation complexity and is extremely time-consuming. In [49], a simple yet effective coding scheme called locality-constrained linear coding (LLC) was proposed to replace the VQ coding. LLC utilizes the locality constraint to project each descriptor onto its local-coordinate system and applies max pooling to the projected coordinates to generate the final representation. Compared with the sparse coding in [3,52], LLC not only guarantees sparsity but also solves the representation problem as a constrained least squares fitting problem, which takes both accuracy and efficiency into consideration.

With the success of the BoVW and SPM models in image scene classification, more algorithms have been developed based on these frameworks. For example, a multi-resolution representation was incorporated into the BoVW model [69,70], which constructed images with multiple resolutions and extracted local features from all the images with dense regions, utilized the *k*-means clustering to generate visual codebook, and then represented each sub-region as a histogram of codeword occurrences by mapping the local features to the codebook. In [8], the pyramid histogram of multi-scale block local binary pattern (PH-MBLBP) descriptor was employed, which can encode micro-and macro-structures of images, and the PH-MBLBP descriptor was verified to be a powerful texture descriptor with low computational complexity. In [66], a concentric circle-based spatial-rotation-invariant representation strategy for describing spatial information of visual words and a concentric circle-structured multi-scale BoVW method using multiple features were proposed to enhance image rotation invariance, leading to excellent scene classification results. In [65], a scene image was transformed into multi-features by 2-D wavelet decompositions, and the resulting feature maps were then employed by the BoVW and SPM models; after that, all the features of different feature maps were stacked as inputs to a classic support vector machine (SVM) classifier.

Although the BoVW [54] and SPM, [6,55] models are popular in scene classification and have gained great success, one of disadvantages is that they cannot well capture global structures in an image scene. Compared with the BoVM and SPM models, some global feature based scene classification methods actually have achieved satisfying performance. In [4], a global feature called completed local binary pattern (CLBP) was employed and multi-scale resolution was adopted to generate multi-scale global features for land-use scene classification. In [5], a global Gabor-filtering-based CLBP was applied to generate multi-features. Some other global feature representation methods for scene classification can be found in [19,20,38,40,68]. It is known that the BoVW model with SPM can capture spatial information with ordered block partitions but it is sensitive to rotation variations. On the other hand, the global feature representation focuses on global texture, such as texture depth and global contrast, but ignores local details or small objects.

Since each type of features (i.e., local features such as BoVW, and global features such as CLBP) has its own advantages and limitations, we propose to fuse the global and local features together to characterize both local fine details and global structures of scene images. In traditional methods, the fusion strategy mainly includes feature and decision level fusion [67]. Nevertheless, it is difficult to combine different features in feature level because different features may not be compatible. Moreover, it may impose a challenge of high feature dimensionality in feature level fusion. In decision level fusion, voting may lead to a rough result. Thus, in this work, both feature and decision level-fusion methods are considered to mitigate their individual shortcomings in the proposed framework. We first use the BoVW and SPM to generate local features and employ the multi-scale CLBP (MS-CLBP) to extract global features. We then employ the feature representation method (e.g., kernel collaborative representation-based classification, KCRC [53]) which is different from a conventional training-testing classifier (e.g., SVM or extreme learning machine (ELM) [22]). The features in training images can be presented as a dictionary, then testing images are reconstructed by the training dictionary. After obtaining the representation residuals from using two types of features, the sum of weighted residuals is calculated and the label is assigned according to the minimal residual class. Experimental results on four benchmark datasets demonstrate that our approach gains a remarkable classification improvement over the state-of-the-art methods.

The main contributions made in this paper can be summarized as follows. (1) To the best of our knowledge, it is the first time that local and global features are employed together for scene image classification. The complementary nature of these two types of features can effectively mitigate the shortcomings of local feature representation based methods (e.g., BoVW and SPM) and global feature representation based methods (e.g., MS-CLBP). (2) A weighted sum of approximation residuals of local and global features with collaborative representation fusion are proposed, which can overcome the difficulties in feature or decision level fusion.

The rest of the paper is organized as follows. Section 2 describes the proposed approaches, including local and global feature extraction and the collaborative representation fusion strategy. Section 3 introduces four experimental datasets. Section 4 represents the experimental results and provides some discussion. Section 5 draws the conclusions.

## 2. Proposed approach

The flow chart of the proposed method is illustrated in Fig. 1. First, the local and global feature dictionaries are prepared in advance. Then, for each testing image, the local and global features are extracted and collaboratively represented by the