Pattern Recognition 47 (2014) 2940–2951

Contents lists available at ScienceDirect

Pattern Recognition

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

Extended Spectral Regression for efficient scene recognition

Liyuan Li*, Weixun Goh, Joo Hwee Lim, Sinno Jialin Pan

Institute for Infocomm Research, Singapore 138632, Singapore

ARTICLE INFO

Article history:
Received 5 October 2012
Received in revised form
1 October 2013
Accepted 18 March 2014
Available online 27 March 2014

Keywords:
Spectral Regression
Subspace learning
Machine learning
Image classification
Scene recognition
Computer vision

ABSTRACT

This paper proposes a novel method based on Spectral Regression (SR) for efficient scene recognition. First, a new SR approach, called Extended Spectral Regression (ESR), is proposed to perform manifold learning on a huge number of data samples. Then, an efficient Bag-of-Words (BOW) based method is developed which employs ESR to encapsulate local visual features with their semantic, spatial, scale, and orientation information for scene recognition. In many applications, such as image classification and multimedia analysis, there are a huge number of low-level feature samples in a training set. It prohibits direct application of SR to perform manifold learning on such dataset. In ESR, we first group the samples into tiny clusters, and then devise an approach to reduce the size of the similarity matrix for graph learning. In this way, the subspace learning on graph Laplacian for a vast dataset is computationally feasible on a personal computer. In the ESR-based scene recognition, we first propose an enhanced low-level feature representation which combines the scale, orientation, spatial position, and local appearance of a local feature. Then, ESR is applied to embed enhanced low-level image features. The ESR-based feature embedding not only generates a low dimension feature representation but also integrates various aspects of low-level features into the compact representation. The bag-of-words is then generated from the embedded features for image classification. The comparative experiments on open benchmark datasets for scene recognition demonstrate that the proposed method outperforms baseline approaches. It is suitable for real-time applications on mobile platforms, *e.g.* tablets and smart phones.

1. Introduction

With the proliferation of mobile computing devices such as smart phones and tablets, there is an increasing demand for real-time vision algorithms for scene recognition and image classification on these platforms. Potential applications include wearable egocentric systems [1], mobile robots [2], and tourist information access [3]. With limited memory and computational resources, efficient high performance algorithms will be core technologies for these emerging fields.

In the past decade, scene classification has attracted much attention from researchers in computer vision and pattern recognition. Recent progress has shown that approaches based on bag-of-words can achieve impressive performance [4–10]. In BOW-based methods, the first step is clustering the local visual features into small groups as codewords, *i.e.* the bag-of-words. Each image is then represented as a histogram of the bag-of-words. The next step is applying a learning model on the histogram representation for classification. The conventional BOW-based approach is simple and effective, but its performance is not satisfactory on challenging datasets. Most recent efforts focus on the extension of BOW-based representation for improving performance on a few challenging benchmark datasets. The successful approaches include Spatial Pyramid Matching (SPM) to exploit spatial information [5], various coding and pooling techniques to improve discriminative power [11], and advanced classifiers for accurate classification [12]. The introduction of these techniques has significantly improved performance on the benchmark datasets. However, it also comes at the expense of great increases in memory requirements and computational costs.

Techniques of data dimensionality reduction are frequently used to improve the efficiency of complicated algorithms based on high-dimensional features. By reducing the dimensionality of features, both memory size and computation time required by the algorithm can be reduced greatly. Recently, Spectral Regression (SR), a general framework based on graph learning for dimensionality reduction, has achieved better performance than the conventional approaches of PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) in many applications [13,14]. These recent progresses motivate us to apply SR to improve the efficiency of scene recognition. To apply spectral regression to a real-world problem, one has to solve the eigen-problem on an $m \times m$ matrix, where m is the number of samples. It becomes prohibitive to directly apply it to embed local image features, *e.g.* dense SIFT

* Corresponding author. Tel.: +65 64082509; fax: +65 67761378.
E-mail address: lyli@i2r.a-star.edu.sg (L. Li).

http://dx.doi.org/10.1016/j.patcog.2014.03.012
0031-3203/

descriptors, for image classification since there are typically over millions of dense local features extracted from all the training images. Solving such an eigen-problem on a personal computer is still infeasible now since more than 1TB memory space is required.

In this paper, a novel approach to apply SR to encapsulate low-level image features for efficient scene recognition is proposed. First, a new SR approach, called Extended Spectral Regression (ESR), is proposed particularly for feature dimensionality reduction on a vast dataset which contains a huge number of data samples. Based on ESR, we propose a novel method to map the low-level image features into an embedded manifold subspace for efficient BOW-based image classification. We propose an enhanced low-level feature representation which combines the local appearance descriptor and its spatial, scale and orientation information into a concise representation. ESR is then applied to encapsulate the enhanced low-level features into a low dimensional manifold subspace. The Bag-of-Words is then generated on the embedded manifold subspace for image classification. With the powerful manifold learning, it is possible to pull related local features of the same class closer and push the local features from different image classes apart in the manifold subspace. Hence, the BOW generated on the manifold subspace will have better descriptive power for image representation.

Our method not only generates lower dimension visual words but also integrates various aspects of the low-level feature into the compact representation. Therefore, we can obtain an effective image representation of much lower dimension while achieving better results compared with PCA and SPM. We have evaluated our method on two challenging datasets for scene recognition, namely Scene-15 and UIUC-Sports. We also implemented our method with OpenCV for easy deployment and tested on a new indoor scene dataset. The memory requirement and computational cost indicate that our technique is suitable for deployment on portable computing devices.

1.1. Related works

To achieve efficient image classification for real-time tasks, it is desirable to reduce the dimension of image features, or visual words. This speeds up the computation of histogram representation and reduces the memory requirements for visual vocabulary. PCA has been applied on SIFT (Scale-Invariant Feature Transform) and HOG (Histogram of Oriented Gradients) features. In some papers [15,16], almost no loss of performance has been observed when PCA is used, while in other papers [17], it is found that PCA degrades performance considerably. In this paper, we propose to use SR to reduce the dimension of low-level image features for efficient scene recognition since recent papers have shown the superiority of SR over PCA and LDA in many applications [13,14]. As mentioned above, SR cannot be applied directly to the dataset of low-level image features since there may be over millions of local features extracted from all training images. Hence, we propose the ESR for this purpose.

Since its introduction [18,19], the bag-of-words model has become the most effective representation for image classification. Subsequent research has focused on extensions of BOW representation to achieve improved performance on challenging benchmark datasets. The progresses are achieved along two general directions, namely (a) enhanced encoding and (b) spatial layout representation.

In usual BOW methods, simple hard histogram encoding is used where a local feature is assigned to the nearest visual word in the vocabulary. On the bag-of-words generated by k-means, enhanced encoding techniques have been proposed. In [16], Gemert et al. proposed to replace hard quantization by soft quantization, or kernel codebook encoding. Using soft assignment, a local feature may be assigned to a few closer visual words in the dictionary with weights within [0,1] according to its distance to the words. Sparse coding approaches are further proposed to improve the soft assignment [11,20,21]. Sparse coding uses a linear combination of a small number of visual words to approximate the local feature, and the coefficients of projecting the feature down to the local linear subspace spanned by the set of visual words are pooled in the histogram. An appealing encoding approach, Fisher encoding, has been proposed for image classification recently [15,22]. With a dictionary (BOW) of K words, Fisher encoding captures the average first and second order differences between the local features and the visual words on a learned GMM (Gaussian Mixture Models) model for the codewords. Hence, it leads to an extended image representation of $K(2D+1)$ dimensions where D represents the feature dimension. Improvements over state-of-the-arts have been obtained through these enhanced encoding techniques, in particular the Fisher encoding. However, image representation is greatly extended and extra computation is required for feature encoding.

In the basic BOW framework, the image representation is a frequency histogram of quantized local features, where the spatial layout of the local features is completely ignored. Clearly, spatial information of low-level features is useful since the compositions of particular visual objects and context regions typically share common spatial layout properties. Various approaches to encode spatial information in BOW representation have been explored [10,23,24]. Among them, the most effective way is to extend the basic BOW representation by using Spatial Pyramid Matching (SPM) [5]. SPM partitions the image into increasingly finer cells, up to 3 layers, and concatenates the BOW histograms of the cells. For a 3-layer pyramid, the image representation is extended to $\sum_{i=0}^{2} 4^i K = 21K$, where K is the dictionary size. The SPM strategy is used in most state-of-the-art approaches [11,15,17,20,21,25–27]. More recently, Krapac et al. [28] proposed to extend the BOW representation by using Fisher kernel to encode the spatial layout of visual words, which is represented by learned spatial MoG (Mixture of Gaussians, *i.e.* GMM) models. It can reduce the image representation from 64,500 dimensions by SPM to 13,300 dimensions and obtain comparable results. In our method, we propose a concise low-level feature representation which includes the spatial, scale, orientation, and appearance information of the local feature. Such enhanced descriptor is then mapped into a compact manifold subspace learned by ESR. Hence, there is no need to extend the BOW representation to encode the spatial information of local features.

It is worth to note that some sample selection methods, such as Editing and Condensing algorithms [29], also generate a compact sample set for classification. These methods aim at selecting a sufficiently small set of samples from the whole training set by removing outliers. The compact sample set can reduce the computational burden of classification. However, as reported in [30], discarding any features, even the most non-informative features will result in the deterioration of image classification performance. Different from these methods, our method embeds all the local image features into a compact and effective manifold subspace for efficient image classification.

1.2. Contributions

Our method is illustrated by Fig. 1, where the gray blocks indicate the novel parts. It contains two stages, *i.e.* Training Stage and Testing Stage. There are three steps in the Training Stage, as illustrated by the three columns of blocks from the left to the right in the figure. In the first step, we first cluster the huge number of enhanced low-level local features from all training