Pattern Recognition 82 (2018) 118–129

Contents lists available at ScienceDirect

Pattern Recognition

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

Improved spatial pyramid matching for scene recognition

Lin Xie[a,1], Feifei Lee[a,1,*], Li Liu[b], Zhong Yin[a], Yan Yan[a], Weidong Wang[a], Junjie Zhao[a], Qiu Chen[c,*]

[a] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China
[b] School of Information Engineering, Nanchang University, China
[c] Major of Electrical Engineering and Electronics, Graduate School, Kogakuin University, Japan

ARTICLE INFO

Article history:
Received 19 May 2017
Revised 20 February 2018
Accepted 26 April 2018
Available online 27 April 2018

Keywords:
Spatial pyramid matching (SPM)
Spatial partition
Histogram of oriented gradients (HOG)
Autoencoder
Scene recognition

ABSTRACT

A scene image is typically composed of successive background contexts and objects with regular shapes. To acquire such spatial information, we propose a new type of spatial partitioning scheme and a modified pyramid matching kernel based on spatial pyramid matching (SPM). A dense histogram of oriented gradients (HOG) is used as a low-level visual descriptor. Furthermore, inspired by the expressive coding ability of autoencoders, we also propose another approach that encodes local descriptors into mid-level features using various autoencoders. The learned mid-level features are encouraged to be sparse, robust and contractive. Then, modified spatial pyramid pooling and local normalization of the mid-level features facilitate the generation of high-level image signatures for scene classification. Comprehensive experimental results on publicly available scene datasets demonstrate the effectiveness of our methods.

1. Introduction

As one of the most challenging problems in the field of computer vision, scene recognition has received considerable attention due to the rapid development of intelligent machines. The approach of placing low-level descriptors (e.g., colour histogram, Local Binary Pattern, and Scale-Invariant Feature Transform) into a classifier directly has been shown to perform poorly [1] because scene images often contain many objects of interest under various backgrounds. Moreover, low-level descriptors are mainly dependent on edges or corner points, and they cannot provide adequate semantic information for scene recognition. Therefore, many researchers have focused on identifying the intermediate semantic representations to narrow the gap between computers and humans with respect to understanding scenes.

Many researchers have attempted to transform low-level descriptors into richer intermediate representations to improve recognition performance and help computers understand more abstract concepts. One extremely popular method is the Bag-of-Visual-Words (BoVW) [2], which is derived from text analysis. BoVW usually involves the following steps. First, local visual descriptors are extracted from image patches, and then dictionary learning produces the codebook, which includes representative visual words. Finally, the image can be characterized by the frequency histogram of the visual words. BoVW discards the spatial structure information in scene images, which restricts the power of the image representations. To overcome this problem, spatial pyramid matching (SPM) [3] based on the BoVW was proposed as a method of incorporating the spatial information of local visual descriptors into the histograms, and it has achieved significant success.

In this paper, the SPM method is used to identify generic spatial structure information within scene images and learn mid-level features. We adopt the histogram of oriented gradients (HOG) as the underlying descriptor because HOG descriptors can be easily and rapidly extracted. To incorporate the generic spatial structure information into the traditional SPM, a new spatial partitioning scheme is proposed to capture a greater degree of local sensitivity in scene images. Partitions in the horizontal and vertical directions are added to preserve consistent structure information. We also modify the pyramid matching kernel to alleviate the influence of viewpoints. This modified SPM achieves better performance and is superior to the conventional SPM in its computational and storage requirements. The steps of this modified SPM are shown in Fig. 1(a). After the K-means clustering on local visual descriptors, the spatial distribution histograms can be calculated. By applying the modified pyramid matching kernel, the histogram representation of the whole image can be obtained. Finally, the intersection kernel SVM (Support Vector Machine) is used to realize the classification.

Another approach named modified spatial pyramid pooling based on various autoencoders is proposed in this paper. Many

* Corresponding authors.
E-mail addresses: feifeilee@ieee.org (F. Lee), q.chen@ieee.org (Q. Chen).
[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.patcog.2018.04.025
0031-3203/

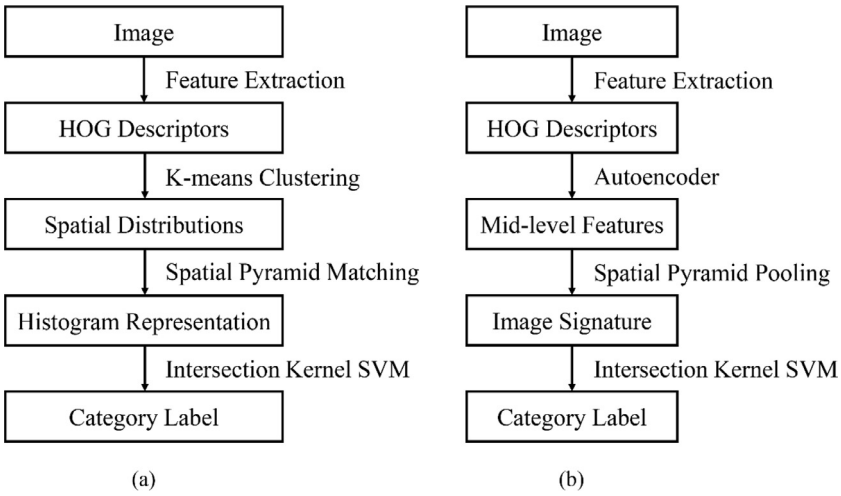

Fig. 1. (a). Major steps of the modified SPM. (b). Major steps of the modified spatial pyramid pooling based on various autoencoders.

models learn representations directly from pixels; in contrast, we explore the encoding of local visual descriptors. As an unsupervised learning technique, the autoencoder is designed to learn an over-complete mid-level feature. A single autoencoder has fewer parameters than other deep architectures, and the directed model facilitates its training. Interesting properties of local visual descriptors are exploited by using three types of autoencoder variants: sparse autoencoder, denoising autoencoder and contractive autoencoder. The learned mid-level features are encouraged to be sparse, robust and contractive. The training process of the autoencoder corresponds to the dictionary learning in the BoVW framework. This method makes the inference of the mid-level features more efficient. Then, the modified spatial pyramid pooling and local normalization on the mid-level features map produce the high-level image signature for scene recognition. This architecture merges the complementary strength of the BoVW framework and autoencoders. Compared with that of other unsupervised means, such as sparse coding, the inference of this model is simple and fast. The main steps of this approach mentioned above are shown in Fig. 1(b).

The remainder of this paper is organized as follows. We review related works in Section 2. The basic techniques are introduced in Section 3. The details of our proposed methods are described in Section 4, including our modified SPM based on HOG, the modified spatial pyramid pooling based on various autoencoders and intersection kernel SVM for scene classification. The experimental results and a discussions are provided in Section 5. Finally, we conclude this paper and offer the suggestions for future work in Section 6.

2. Prior work

Due to the limited power of local visual descriptors, many global features for scene recognition, including GIST [4], CENTRIST [5] and LDBP [6], have been proposed to describe the holistic appearance of a scene. Yu et al. [7] used a unified low-dimensional subspace to effectively fuse multiple features for scene recognition. Scene images can be described as a set of meaningful visual attributes [8,9], although defining of these visual attributes requires significant manual effort and the performance is restricted.

In addition to these elaborate global features, other scene representations are obtained by the popular BoVW model. Many improved versions have emerged over the past few years. For example, Zhou et al. [10] presented a novel Gaussianised vector representation using a global Gaussian Mixture Model (GMM). Qin and Yung [11] extended the BoVW model by introducing contextual information that provides useful cues about the region of interest. Zhou et al. [12] incorporated a multi-resolution representation into the BoVW model. Hotta proposed the local autocorrelation (LAC) feature [13] and local co-occurrence feature [14] based on the subspaces obtained by a Kernel Principal Component Analysis (KPCA) of visual words. Similar to the efforts in the framework of the BoF model, our modified SPM is designed to construct a more discriminative spatial pyramid representation by combining the histograms of visual words from different regions.

In the BoVW framework, many representative visual words constitute the dictionary, feature encoding is used to map the local features to richer intermediate features describing the weights of visual words. Many methods of dictionary learning and feature encoding have been proposed to improve the discrimination of visual words. Wu et al. [15] presented a modified K-means algorithm by incorporating the histogram kernel, and their approach generated a better dictionary because local descriptors were compared in the histogram intersection kernel space instead of the Euclidean space. Yang et al. [16] proposed an extension of SPM utilizing sparse coding of local visual descriptors. Gao et al. [17] introduced the Laplacian matrix in sparse coding to address the problem in which similar local visual descriptors are transformed into different codes. Locality-constrained Linear Coding (LLC) [18] leads to local smooth sparsity and better reconstruction. In our study, we attempt to encode local visual descriptors by autoencoder variants to accelerate the coding process.

In addition to the methods of feature encoding, some researchers have also focused on spatial structure information. For example, Harada et al. [19] estimated the optimal weights of each cell for the most discriminative power, and Jiang et al. [20] developed two classifiers to select the most discriminative pattern from randomized spatial partition schemes. Compared with these models, the method of determining weights and the spatial partition scheme in the approach proposed here is simpler and more effective.

In recent years, deep learning has rapidly developed because of its promising performance for many problems. Deep architectures allow for the exploitation of the potential semantic information behind inputs and provide benefits for classification tasks. However, many deep learning models, such as the Convolutional Neural Network (CNN), Deep Belief Network (DBN) and Deep Boltzmann Machine (DBM), require a large amount of labelled training samples and the consideration of underfitting and overfitting problems during the training process.