Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

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

Towards effective codebookless model for image classification

Qilong Wang [a], Peihua Li [a,*], Lei Zhang [b], Wangmeng Zuo [c]

[a] School of Information and Communications Engineering, Dalian University of Technology, Dalian 116024, China
[b] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
[c] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

A R T I C L E I N F O

Article history:
Received 23 July 2015
Received in revised form
28 February 2016
Accepted 2 March 2016

Keywords:
Codebookless model
Image classification
Bag-of-features
Riemannian manifold

A B S T R A C T

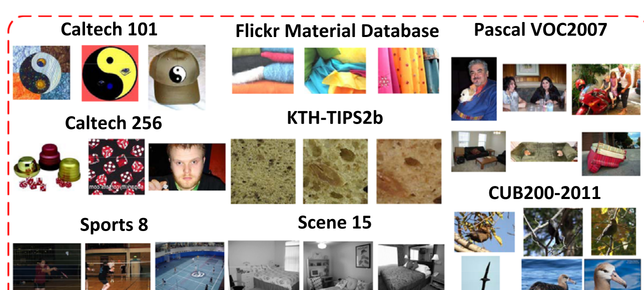
The bag-of-features (BoF) model for image classification has been thoroughly studied over the last decade. Different from the widely used BoF methods which model images with a pre-trained codebook, the alternative codebook-free image modeling method, which we call codebookless model (CLM), attracts little attention. In this paper, we present an effective CLM that represents an image with a single Gaussian for classification. By embedding Gaussian manifold into a vector space, we show that the simple incorporation of our CLM into a linear classifier achieves very competitive accuracy compared with state-of-the-art BoF methods (e.g., Fisher Vector). Since our CLM lies in a high-dimensional Riemannian manifold, we further propose a joint learning method of low-rank transformation with support vector machine (SVM) classifier on the Gaussian manifold, in order to reduce computational and storage cost. To study and alleviate the side effect of background clutter on our CLM, we also present a simple yet effective partial background removal method based on saliency detection. Experiments are extensively conducted on eight widely used databases to demonstrate the effectiveness and efficiency of our CLM method.

1. Introduction

Image classification has been attracting massive attentions in computer vision and pattern recognition communities in recent years. It is one of the most fundamental but challenging vision problems because images, as illustrated in Fig. 1, often suffer from significant scale, view or illumination variations (e.g., in texture classification [8] and material recognition [23]), and pose changes, background clutter, partial occlusion (e.g., in scene categorization [31,32] and object recognition [17,18,22,52]).

For a long time the bag-of-features (BoF) model [46] has been almost given priority to image classification. As shown in Fig. 2(a), the BoF-based methods generally consist of five components: local features extraction, learning codebook with training data, coding local features with pre-trained codebook, pooling or aggregating codes over images, and finally, learning classifier (e.g., SVM) for classification. With this processing pipeline, the BoF-based methods can be seen as a hand-crafted five-layer hierarchical feed-forward network [49] with a pre-trained feature coding template (codebook) [7]. The learned codebook depicts the distribution of feature space, and makes coding of high dimensional features possible. This architecture has achieved very promising performance in a variety of image classification tasks.

The codebook as a reference for feature coding serves as a bridge between local features and global image representation. However, it is well known that segmentation of feature space involved in building of codebook brings on quantization error [6], and leads to continuous striving for this side effect (e.g., soft coding methods [44,19,55] alleviate but cannot completely eliminate it). Though offline, training of codebook, particularly large



	Caltech 101	Caltech 256	KTH-TIPS2b	FMD	VOC2007	CUB200-2011	Scene 15	Sports 8
FV	83.7	50.1	71.3	58.9	61.8	27.3	89.4	91.3
Our CLM	86.3	53.6	75.2	64.2	61.7	36.0	89.4	91.5

Fig. 1. Some example images and accuracy comparison (in %) between Fisher vector (FV) and our codebookless model (CLM) on various image databases.

* Corresponding author.
E-mail address: peihuali@dlut.edu.cn (P. Li).

http://dx.doi.org/10.1016/j.patcog.2016.03.004
0031-3203/

size ones, is time consuming. In addition, in general the pre-trained codebook on one database cannot naturally adapt to other databases [58].

An alternative approach is to estimate the statistics directly on sets of local features from input images [10,38,50], as illustrated in Fig. 2(b), which is called codebookless model (CLM) in this paper. It is clear from Fig. 2 that the major difference is that the BoF model learns a codebook to explore the statistical distribution of local features and then performs coding of descriptors, while the CLM represents images with descriptors directly, requiring no pre-trained codebook and the subsequent coding. Conceptually, the codebookless model has the potential to circumvent the aforementioned limitations of the BoF model, however, which has received little attention in image classification community. The main reasons may be that such methods have not yet shown competitive classification performance, and that they often need to utilize inefficient and unscalable kernel-based classifiers.

In this paper, we propose an effective CLM scheme, and argue that the CLM can be a competitive alternative to the BoF methods for image classification. The comparison between state-of-the-art BoF method, Fisher Vector (FV) [44], and our CLM on various image databases is shown in Fig. 1. First and foremost, we extract a set of local features (e.g., SIFT [37]) on a dense grid of image, and simply model them with a single Gaussian model to represent the input image. Then, we employ a two-step metric for matching Gaussian models. By using this metric, Gaussian models can be fed to a linear classifier for ensuring efficient and scalable classification while respecting the Riemannian geometry structure of Gaussian models. Moreover, we introduce two well-motivated parameters into the used metric. One is to balance the effect between mean and covariance of Gaussian, and another is for eigenvalue power normalization on covariance.

Our codebookless model usually is of high dimension, by incorporating low-rank learning with SVM, we propose a joint learning method to effectively compress Gaussian models while respecting their Riemannian geometry structure. It is mentionable that, to the best of our knowledge, we make the first attempt to perform joint learning of low-rank transformation and SVM on Gaussian manifold. Finally, to alleviate the side effect of background clutter, a saliency-based partial background removal method is proposed to enhance our CLM. The experimental results show that partial background removal is helpful to CLM when images are heavily cluttered (e.g., CUB200-2011 and Pascal VOC2007).

2. Related work

The codebookless model for directly modeling the statistics of local features has been studied in past decades. Rubner et al. [43] introduced signatures for image representation, and proposed the Earth Mover's Distance for image matching which is robust but has high computational cost. Tuzel et al. [50] for the first time used covariance matrices for representing regular image regions, and employed Affine-Riemannian metric which suffers from high computational cost [40]. Gaussian model as image descriptor has been used for visual tracking [20], in which Gaussian models are matched based on the Riemannian metric, involving expensive operations to solve generalized eigenvalue problem. Going beyond Gaussian, Gaussian mixture model (GMM) is more informative and is used in image classification and retrieval [3,41]. However, GMM suffers from some limitations, such as high computational cost of matching methods and lacking of general criteria for model selection.

Our work is motivated by [9,10] and [38]. Carreira et al. [9,10] modeled the free-form regions obtained by image segmentation with estimating the second-order moments. By using Log-Euclidean metric [2], the method in [9,10] can be combined with a linear classifier, which has shown competing recognition performance on images with less background clutter (e.g., Caltech101 [18]). Different from [9,10], we employ a Gaussian model to represent the whole image. It is well-known that a covariance matrix can be seen as a Gaussian model with fixed mean vector. Compared to [9,10], our CLM contains both the first-order (mean) and second-order (covariance) information. Note that the first-order statistics has proved to be important in image classification [26,44]. Moreover, the manifold of Gaussian models and that of covariance matrices are quite different, and the embedding method in our CLM makes Gaussian models can be handled flexibly and conveniently.

Nakayama et al. [38] also represented an image with a global Gaussian for scene categorization. However, they matched two Gaussian models by using the Kullback–Leibler (KL) divergence, and hence kernel-based classifiers have to be used. This method is

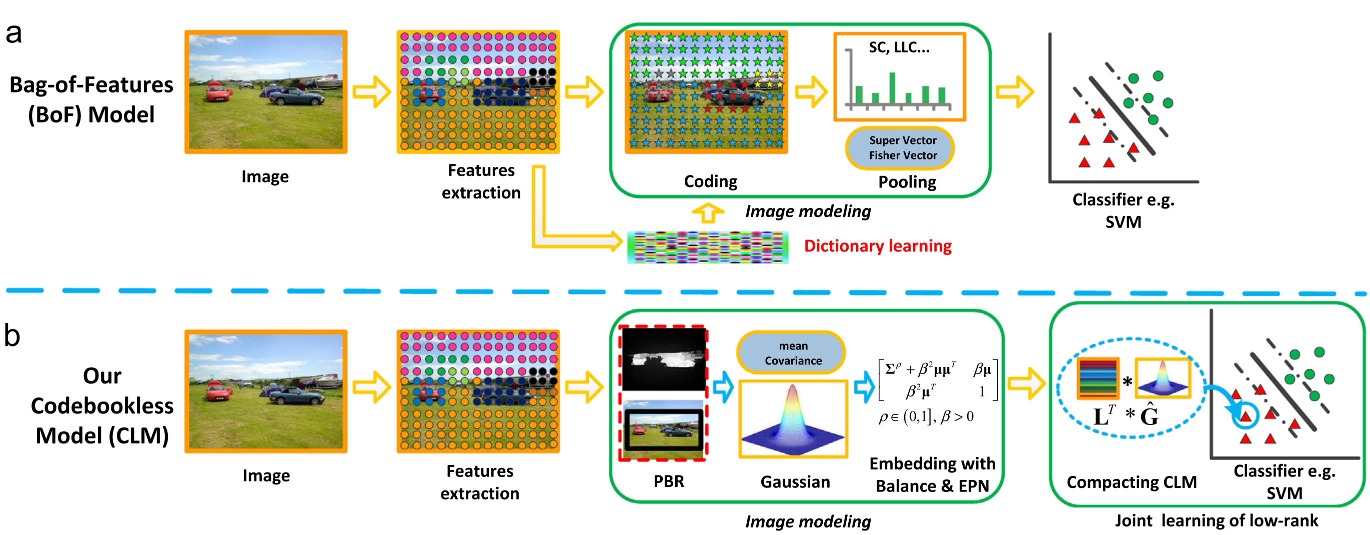


Fig. 2. Comparison between (a) the BoF model and (b) our CLM. The major difference between them is that whether there is a pre-trained codebook & coding or not. Our CLM mainly consists of a Gaussian model for image representation and a joint low-rank learning with linear SVM classifier.

Please cite this article as: Q. Wang, et al., Towards effective codebookless model for image classification, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.03.004