Neurocomputing 184 (2016) 188–195

Contents lists available at ScienceDirect

Neurocomputing

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

Image classification based on saliency coding with category-specific codebooks

Zhen Yang, Huilin Xiong *

Department of Automation, Shanghai Jiao Tong University, 800,Dongchuan Rd., Shanghai 200240, China

A R T I C L E I N F O

Article history:
Received 17 December 2014
Received in revised form
28 July 2015
Accepted 30 July 2015
Available online 28 November 2015

Keywords:
Image classification
Feature coding
Dictionary learning

A B S T R A C T

This paper presents a feature encoding scheme for image classification by combining the salient coding method with the category-specific codebooks, which are generated separately using the training images of each category. Different from the usual way of concatenating or merging the category codebooks to form a global dictionary, we employ the category codebooks to calculate a type of category-sensitive saliency feature, and then, encode the saliency features to form a representation of image content. Compared to the state-of-the-art methods such as LC-KSVD, the dictionary generation and feature encoding in our scheme are pretty simple, and no complicated optimization is involved. However, our scheme can achieve better, in some cases, significantly better results, in terms of the classification accuracy, than the state-of-the-art methods. Extensive experiments are carried out to show the effectiveness of our method in comparing with various image classification methods.

1. Introduction

In recent years, the image classification based on the local feature coding has attracted a lot of attentions in the communities of computer vision, pattern recognition, and machine learning, and a variety of feature coding methods have been proposed [2–4,12,13,15–17,19,20,22], which promote broad applications of image classification in image and video retrieval [1,6], video surveillance [5,7], and web content analysis [8]. Basically, image classification is to assign a category label to a given image, according to the image content, which is usually represented via "encoding" a type of local feature (*e.g.*, SIFT), densely extracted from the image. Typically, the image classification based on the local feature coding is composed of the following four steps: 1) extraction of the local image features; 2) generating a codebook or dictionary from training images; 3) encoding of the local features to form a representation of the image content; and 4) classification of the image representations by a classifier (*e.g.*, SVM). Among them, the codebook generation and the local feature encoding are the crucial components for the success of an image classification approach, and characterize the major difference of various image classification methods.

The codebook or dictionary is usually generated by the conventional clustering algorithm, *i.e.*, *K*-means. However, recently, dictionary learning or codebook generation via a supervised way, rather than an unsupervised clustering, has attracted much interest. The simplest way of constructing a codebook, using the supervised information (category labels), is to generate one codebook for each category, and then, combine or merge the category codebooks to form a global codebook [9–11]. More sophisticated way of codebook generation adopts some optimization techniques to learn good codebook or dictionary. For example, in [12], a dictionary learning scheme, referred to as K-SVD, is proposed, in which the dictionary learning is transformed to a problem of l_0-norm sparse representation (SR), and an efficient optimization algorithm is developed. Based on K-SVD, the label constraint K-SVD scheme [13], referred to as LC-KSVD, is proposed to learn a discriminative dictionary and encode the local features, simultaneously.

Quantization of the local features using a codebook is the most important step in the feature encoding, which is in essence to approximate each local feature by a linear combination of the codewords in a codebook. The combination coefficients are usually called quantization responses of the local feature on codebook. A simple way of feature quantization is the conventional vector quantization (VQ), in which each feature is quantized just by its nearest codeword, and the quantization responses only contain one non-zero elements. Instead of the "hard" style of VQ, Soft-VQ [14] quantizes each feature with a Gaussian-weighted linear

* Corresponding author. Tel.: +8602134204022; fax: +8602134204022.
E-mail addresses: yangzhen5771@sjtu.edu.cn (Z. Yang), hlxiong@sjtu.edu.cn (H. Xiong).

http://dx.doi.org/10.1016/j.neucom.2015.07.124
0925-2312/

combination of multiple nearest codewords. Moreover, ScSPM [15] employs the sparse representation technique to quantize each of the local features. However, due to the nature of highly computational complexity in SR, the ScSPM method is very time-consuming in practice. In [16], an efficient feature coding method, Local-constraint Linear Coding, referred to as LLC, proposes to apply the locality constraint, rather than the sparsity constraint, to approach the problem of feature quantization. LLC is shown to be very efficient in computation, and more importantly, the image representation based on the LLC coding usually leads to good results of image classification.

In LLC, local features are quantized by multiple nearest codewords, and max pooling method is adopted to encode the quantization responses, where only the strongest response is preserved. The strength of a response of a local feature on a codeword indicates the proximity or saliency of the codeword to the local feature. If there is no dominant response, only preserving the relatively largest response will inevitably lead to a loss of discriminative information. In view of this, a method called Salient Coding is introduced in [17,31], in which a relative proximity, rather than the absolute proximity, is incorporated in the LLC scheme to further improve the performance of LLC in image classification.

In this paper, we present a feature coding scheme by combining the salient coding method with the category-specific codebooks, generated separately using the training images of each category. Different from the usual way of concatenating or merging the category codebooks to form a global dictionary, we use the category-specific codebooks to calculate a type of category-sensitive saliency feature, and then, input the saliency features to the feature encoding pipeline. In our scheme, the dictionary generation and feature encoding are pretty simple compared to the state-of-the-art methods such as LC-KSVD, and no complicated optimization is involved. But, surprisedly, our feature coding scheme can achieves better, in some cases, significantly better results, in terms of the classification accuracy, than the state-of-the-art methods. Extensive experiments are carried out to show the effectiveness of our method in comparing with various image classification methods.

The remainder of this paper is organized as follows: In Section 2, we briefly introduce some of the conventional feature coding methods; Section 3 presents in detail our feature coding method based on the category codebooks and the category saliency feature coding; and Section 4 gives the experimental results in comparing with various feature coding approaches, on three widely used image classification databases. Finally, Section 5 concludes the paper.

2. Relate works

The image classification based on local feature coding usually adopts the densely extracted local features, e.g., SIFT, to construct image representation. Let $Y=(y_1,y_2,\cdots,y_N)\in\mathfrak{R}^{D\times N}$be N D-dimensional local features extracted from the densely segmented grid in an image. A codebook or dictionary with K codeword is denoted as $C=(c_1,c_2,\cdots,c_K)\in\mathfrak{R}^{D\times K}$, which is usually generated from the local features of the training images using the K-means clustering.

Given an image with its local features, $Y=(y_1,y_2,\cdots,y_N)\in\mathfrak{R}^{D\times N}$, the feature quantization is to quantize each of the local feature vector with the codebook, aiming to obtain a more compact and discriminative representation of the image content. Let $W=(w_1, w_2,\cdots,w_N)\in\mathfrak{R}^{K\times N}$ denote the quantization responses of the local features on the codebook. The conventional quantization methods can be described as follows.

2.1. Vector quantization (VQ)

The simplest feature quantization is the classical vector quantization, which can be expressed as,

$$w(i)=\begin{cases}1, & \text{if } i=\arg\min_j(\|y-c_j\|_2)\\ 0, & \text{otherwise}\end{cases},\quad i=1,2,\cdots,K \tag{1}$$

The soft-VQ [14] is a more precise VQ method, in which Gaussian weights are used to describe the quantization responses,

$$w(i)=\frac{\exp\left(\|y-c_i\|_2^2/\sigma\right)}{\sum_{k=1}^{M}\exp\left(\|y-c_k\|_2^2/\sigma\right)},\quad i=1,2,\cdots,K \tag{2}$$

where σ is the Gaussian parameter, M denotes the number of the nearest neighbors in computation. For the sake of computational efficiency, in the scheme of LSA [18], M is set to be far smaller than K.

2.2. Sparse coding (SC)

Feature quantization approximates each of the local feature by a linear combination of the codewords in a predefined codebook. While the VQ method approximates each local feature by one codeword, the sparse coding method [15] uses multiple codeworks to approximate each local feature with the sparsity constraint, aiming to make a good trade-off between the quantization precision and the sparsity of the linear approximation. SC can be formulated as the following optimization problem,

$$\begin{aligned}&\arg\min_w \|y-Cw\|_2^2+\lambda\sum_{i=1}^{K}|w(i)|\\ &s.t.\sum_{i=1}^{K}w(i)=1\end{aligned} \tag{3}$$

where λ is a regularization constant for the trade-off. Although SC could lead to a good feature quantization, its highly computational complexity prevents it from being a feasible quantization scheme for a large dataset.

2.3. Locality-constrained Linear Coding (LLC)

A further development of the sparse coding is the Locality-constrained Linear Coding method (LLC) [16], in which the locality rather than the global sparsity is emphasized. To address the problem of highly computational complexity in the optimization procedure, LLC proposes to adopt a new locality constraint to replace the l_1-norm sparsity constraint,

$$\begin{aligned}&\arg\min_w \|y-Cw\|_2^2+\lambda\sum_{i=1}^{M}\left(w(i)\exp(\|y-c_i\|_2/\sigma)\right)^2\\ &s.t.\sum_{i=1}^{M}w(i)=1\end{aligned} \tag{4}$$

which makes the LLC method having an analytical solution. Besides its computational efficiency, the LLC based feature coding has been shown to be very effective for image classification.

In the LLC coding method, the max pooling strategy is used to encode the quantization responses, where only the strongest response is preserved. The strength of a response of a local feature on a codeword indicates the proximity or saliency of the codeword to the local feature. If a codeword is very close to a local feature, the response on this codeword will be the dominant one over those on the other codewords, which means the codeword can independently describe this feature perfectly. On the other hand, if there is no dominant response, only preserving the relatively largest response will inevitably lead to a loss of discriminative