Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Separable vocabulary and feature fusion for image retrieval based on sparse representation

Yanhong Wang[a,b], Yigang Cen[a,b,*], Ruizhen Zhao[a,b], Yi Cen[c], Shaohai Hu[a,b], Viacheslav Voronin[d], Hengyou Wang[e]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Key Laboratory of Advanced Information Science and Network Technology of Beijing, Beijing 100044, China
[c] School of Information Engineering, Minzu University of China, Beijing 100081, China
[d] Department of Radio-electronic Systems, Don State Technical University, Shakhty 346500, Russia
[e] School of Science, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

ARTICLE INFO

Keywords:
Separable vocabulary
Sparse representation
Feature fusion
Image retrieval

ABSTRACT

Visual vocabulary is the core of the Bag-of-visual-words (BOW) model in image retrieval. In order to ensure the retrieval accuracy, a large vocabulary is always used in traditional methods. However, a large vocabulary will lead to a low recall. In order to improve recall, vocabularies with medium sizes are proposed, but they will lead to a low accuracy. To address these two problems, we propose a new method for image retrieval based on feature fusion and sparse representation over separable vocabulary. Firstly, a large vocabulary is generated on the training dataset. Secondly, the vocabulary is separated into a number of vocabularies with medium sizes. Thirdly, for a given query image, we adopt sparse representation to select a vocabulary for retrieval. In the proposed method, the large vocabulary can guarantee a relatively high accuracy, while the vocabularies with medium sizes are responsible for high recall. Also, in order to reduce quantization error and improve recall, sparse representation scheme is used for visual words quantization. Moreover, both the local features and the global features are fused to improve the recall. Our proposed method is evaluated on two benchmark datasets, i.e., Coil20 and Holidays. Experiments show that our proposed method achieves good performance.

1. Introduction

In recent years, content-based image retrieval (CBIR) is a very hot research issue of computer vision and multimedia information. Although it has achieved rapid development, researchers have not yet to standardize various image retrieval systems [1]. Image retrieval still remains as a challenging problem. It is the fact that effects of image retrieval are failed due to occlusion, distortion, corrosion and the different lighting conditions.

Image retrieval means that, for a given query image, we will retrieve all the similar images from the database. Similar images are defined as images contain the same objects or a scene viewed under different imaging conditions [2]. In the past years, the BOW model [3,4] has achieved great effect in image retrieval area. This model is inspired by the text retrieval system [3–5]. It contains four major steps: (1). Local features are extracted from each image, such as the SIFT descriptor [6], rootSIFT descriptor [7] and SURF descriptor [8] etc. (2). Each local descriptor is quantized to a visual word according to a pre-trained vocabulary by an unsupervised clustering approach. (3). Each image is represented by a frequency histogram of visual words. (4). Retrieval results are returned according to the similarities between the query image and the images of dataset.

Vocabulary plays a very important role in the BOW model. For a large number of local features, in order to ensure the retrieval accuracy, we need to train a large visual vocabulary. But a large visual vocabulary will lead to a low recall and other issues [9,10]. In order to improve the recall, in previous works, there are two main types of solutions: Firstly, the size of the vocabulary is changed. For examples, in [2], Jegou et al. proposed to use the vocabulary with medium size to improve recall. However, this will lead to a low accuracy [10]. In [11,12], the author represented images with vector of locally aggregated descriptors (VLAD), which can be viewed as a simplification of the fisher vector (FV) [13] representation. Moreover, the VLAD method only requires a small vocabulary in the retrieval process. Secondly, multiple vocabularies based strategies are used. The vocabularies are usually generated by an independent training dataset. In [14], the author proposed a Bayes merging approach to down-weight the indexed features in the intersection set. In [15], instead of computing the multiple vocabul-

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
E-mail address: ygcen@bjtu.edu.cn (Y. Cen).

http://dx.doi.org/10.1016/j.neucom.2016.08.106
Received 27 February 2016; Received in revised form 17 July 2016; Accepted 8 August 2016
Available online xxxx
0925-2312/

aries independently, they presented a method that created the vocabularies jointly.

It is well-known that the SIFT descriptors are robust against image local translation, scaling, rotation and partial invariant for both illumination changes and affine projection [16]. In BOW model, each local descriptor is quantized to a visual word by finding the nearest center in the feature space. However, in practice each local descriptor may relate to multiple visual words of vocabulary. Thus, it is not reliable to rely solely on the frequency histogram of BOW model. Recently, sparse representation [17–20] is widely used in image processing, which can reduce the quantization error and improve recall. The basic idea of sparse representation is seeking an optimal linear combination to approximate the original signal [21]. Thus, sparse representation based scheme is also used in our proposed algorithm for visual words quantization.

The GIST descriptor was initially proposed in [22] and a computational model was proposed for scene recognition, which did not require segmentation and pre-processing of individual objects or regions. In addition, a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) was proposed to represent the dominant spatial structure of a scene. The GIST descriptor aggregates oriented edge responses at multiple scales into very coarse spatial bins. It has been successfully used for scenes classification and image retrieval [23–25]. In [25], the GIST descriptor was used to retrieve an initial set of images of the same landmarks. Then image point based matching algorithm was used to refine the results and a 3D model of the landmark was built. In [26], the GIST descriptor was used to create a GIST indexing structure (GISTIS) as the local SIFT descriptor.

Recently, deep features have shown remarkable performance in many applications of computer vision, such as image classification [27], object recognition [28], speech recognition [29], etc. For image retrieval, most of previous works use off-the-shelf deep features that extracted from convolutional neural networks (CNNs) for image classification task [30,31], however, their performance is significantly below the state-of-the-art methods. Moreover, several works use convolutional networks to produce descriptors suitable for retrieval within the Siamese architectures [32,33].

In this paper, we propose a new method for image retrieval based on feature fusion and sparse representation over separable vocabulary. In [1], the author also proposed combination of sparse representation and feature fusion. They extracted the CLOG (color boosting laplacian-of-gaussian) [34] and the SURF (speeded-up robust feature) features from images. However, we fuse the local texture feature (SITF) and the global scene feature (GIST) at the stage of similarity measurement. For image retrieval of a same scene, the GIST feature will improve the retrieval accuracy. Also, [1] used sparse representation model to match features. However, we adopt sparse representation to select the best vocabulary and quantize features. The main contributions of our paper are as follows.

1. Vocabulary generation and selection: For the vocabulary, separable vocabulary is applied in our algorithm. According to our proposed two steps of clusters and visual words assignment strategy, the correlations within a vocabulary and among the vocabularies will be decreased, which improve the representation ability of the vocabularies such that the retrieval accuracy can be improved. Also, sparse representation is used for the best vocabulary selection, which will reduce the running time for image retrieval by using a small vocabulary.

2. Feature quantization: Sparse representation scheme is employed for visual words quantization. By considering that the non-zero elements in the sparse representation coefficients represent the frequency of occurrence of the visual words for the query image, non-negative constrain of the sparse representation coefficients is added in the optimization model of the quantization process. This non-negative constrain leads to a more reasonable sparse representation result. Then, NNOMP (non-negative orthogonal matching pursuit) method is used to solve the optimization problem, which decreases the time consuming of the sparse representation.

3. Similarity measurement: The local SIFT features based on sparse representation quantization and global GIST features are fused at the stage of similarity measurement, which improves the image retrieval accuracy of our proposed algorithm.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 introduces our proposed method. The experimental results are given in Section 4 and the conclusions are presented in Section 5.

2. Related work

2.1. Vocabulary generation

For a training dataset, we can extract a set of features $Y = (y_1, y_2, \ldots, y_e)$. The vocabulary containing K visual words is usually generated on Y by an unsupervised clustering approach such as k-means, approximate k-means (AKM) [4], or hierarchical k-means (HKM) [35] etc. The vocabulary is generated by minimizing the quantization distortion [36]:

$$\min_{\{c_k\}} \sum_{k=1}^{K} \sum_{y \in C_k} d(y, c_k) \tag{1}$$

$$with: C_k = \{y | d(y, c_k) < d(y, c_{k'}), \forall\, k' \neq k\},$$

where C_k denotes the k^{th} cluster, c_k is the center of the cluster C_k, i.e., c_k is a visual word. Also, $d(\cdot)$ represents the squared Euclidean distance.

In the BOW model, the vocabulary is also called as the quantizer. It is used to quantize SIFT descriptors to visual words. For a SIFT descriptor, the quantization is to find the nearest center in the feature space. In order to reduce the quantization error and improve recall, multiple vocabularies are often generated. Each SIFT descriptor is quantized to multiple visual words by multiple vocabularies. Mathematically, the L vocabularies are trained according to minimize the total quantization distortion [15,36]:

$$\min_{\{c_k^l\}} \sum_{l=1}^{L} \left(\sum_{k=1}^{K} \sum_{y \in C_k^l} d(y, c_k^l) \right) \tag{2}$$

$$with: C_k^l = \{y | d(y, c_k^l) < d(y, c_{k'}^l), \forall\, k' \neq k\},$$

where $C_k{}^l$ denotes the k^{th} cluster of the l^{th} vocabulary, and $c_k{}^l$ is the center of the cluster $C_k{}^l$, i.e., $c_k{}^l$ is the k^{th} visual word of the l^{th} vocabulary. Also, $c_{k'}^l$ is the k'^{th} visual word of the l^{th} vocabulary.

The vocabularies generated by Eq. (2) may be correlative, which may affect the retrieval results. To reduce the vocabulary correlations, separable vocabulary is proposed in this paper, which reduces the correlations among the vocabularies and the visual words within a vocabulary.

2.2. Feature quantization

In BOW model, feature quantization may generate two problems [7]: 1) The original information of descriptors may be lost. 2) The corresponding descriptor may be assigned to other visual word. In order to reduce the quantization error and improve the recall, soft vector quantization (soft-VQ) scheme [9], multiple assignment [37] or sparse representation are usually adopted. Among these methods, each descriptor is assigned to multiple visual words. Moreover, quantization can be regarded as a process of encoding for each descriptor. Accordingly, for hard vector quantization (VQ), each code has only one non-zero element, while for soft-VQ, a small group of elements can be non-zero [21]. Moreover, it has been proved that the coefficient vectors of sparse representation have the discriminant ability in favor of classification purpose [16,38].