Neurocomputing 120 (2013) 277–286

Contents lists available at ScienceDirect

Neurocomputing

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

Visual word coding based on difference maximization

Yu-Bin Yang [a,*], Ling-Yan Pan [a], Yang Gao [a], Guang-Nan He [a], Yao Zhang [b]

[a] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
[b] Jin-Ling College, Nanjing University, Nanjing 210089, China

ARTICLE INFO

Article history:
Received 25 December 2011
Received in revised form
10 June 2012
Accepted 7 August 2012
Available online 27 March 2013

Keywords:
Bag of words
Visual words
Codebook
Difference maximization
Image classification

ABSTRACT

Image classification is an important topic in computer vision, which becomes more and more challenging due to the rapid increase of the amounts and categories of images, as well as the different geometric deformations and illumination variations existing in image objects. "Bag-of-Features" model, also known as "Bag-of-Words" model, plays a fundamental and crucial role in generating efficient and discriminative image content representations by using local descriptors such as visual words, making it widely used in solving image classification problems. However, because of the weak discriminative power and strong ambiguity of the low-level visual features, the visual codebook, i.e., a set of visual words, generated in this model is usually over-completed and inconsistent for capturing image semantics. To address this issue, we propose a novel visual word coding algorithm in this paper based on difference maximization technique to improve the generated codebook model. Instead of mapping an image feature vector to one or multiple nearest visual words, the proposed approach utilizes a group of the nearest and the farthest visual words together in the coding process. Consequently, the representative variations of different image features are well kept and strengthened, which can then improve the discriminative power of the visual word descriptor significantly. We examine the performance of our visual word coding model extensively on four standard real-world image datasets, demonstrating that it captures image semantic content more accurately and achieves superior classification performance.

1. Introduction

Image classification is an important and challenging topic in computer vision, which can be directly used in versatile tasks including image understanding, scene classification, image retrieval, and image content filtering, etc. However, due to the rapid increase of the amounts and categories of images, as well as the different geometric deformations and illumination variations existing in image objects, it becomes more and more difficult to accurately identify the correct semantic category of an image from diverse classes. To solve the above problem, many well-performed image classification techniques have been proposed, among which two models are most widely applied: (1) *bag-of-features* (BoF, also known as "bag-of-words", BoW) [1], and (2) *spatial pyramid matching* (SPM) [2].

The bag-of-features methodology was first proposed for text document analysis and further transferred to be used in computer vision applications [1]. The BoF model uses a visual analogue of a textual "word", constructed by quantizing low-level visual features such as color, shape, and texture, etc. Recent studies have indicated that local features represented by bag-of-features are suitable for image classification tasks. In the BoF model, an image is firstly partitioned into local patches, based on which low-level feature descriptors such as SIFT [3,4], HOG [5–7], SURF [8] or CENTRIST [9] can be extracted. In this manner, an image is represented as a "bag" of local features. Consequently, a set of visual code words, that is, the codebook, is constructed from numerous low-level features via various clustering algorithms such as *k-means* or *random trees*. The next step, called as coding process, assigns all low-level features to the built codebook, after which an image is represented as a histogram of visual words. Then, the final coding result is learned by adopting machine learning techniques, among which SVM [10] is one of the most widely used classifiers with great generalization ability. However, the BoF Model ignores the spatial information contained in an image and thus is incapable of capturing location-related features of an object. In this case, SPM is a good extension for BoF model, which divides an image into increasingly finer spatial local patches, and then concatenates all the histograms of local features of each local patch.

From the above discussion, it can be seen that the coding approach is crucial for guaranteeing good quality of the mapping between the built codebook and low-level features in order to achieve promising classification performance. From this point, there are already many extensions of BoF model aiming to construct better codebook [11,12], and better coding method [13–16]. The common point of those achievements is that the

* Corresponding author.
E-mail address: yangyubin@nju.edu.cn (Y.-B. Yang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.050

classification performance of feature histograms mainly depends on the representative power of the visual words. However, because of the weak discriminative power and strong ambiguity of the low-level visual features, image features with different high-level semantics are probably mapped onto the same visual words. This disadvantage makes the generated visual codebook, i.e., the set of visual words, usually over-completed [17] and inconsistent to capture image semantics.

To address the above issue, we propose a novel visual word coding algorithm in this paper, based on difference maximization technique to improve the generated codebook model. Instead of mapping an image feature vector strictly to one or multiple nearest visual words as the current coding methods usually do, the proposed approach utilizes a group of the nearest and the farthest visual words together in the coding process. This strategy may greatly increase the distribution of the feature's discrepancy over the coded visual words, that is, even if image features belong to the same cluster center, their coding results are not necessarily same. Consequently, the representative variations of different image features are well kept and strengthened, which can then improve the discriminative power of the visual word descriptor significantly. We examine the performance of our visual word coding model extensively on four standard real-world image datasets, demonstrating that it captures image semantic content more accurately and achieves superior classification performance.

The rest of this paper is organized as follows. In Section 2, we discuss the related work on improving codebook model. Section 3 presents the proposed visual word coding method based on difference maximization in details. Section 4 illustrates and analyzes the extensive experimental results on four standard real-world image datasets. Afterwards, in-depth discussion on the experimental results is presented in Section 5. Finally, conclusion remarks are provided in Section 6.

2. Related work

As mentioned in Section 1, the codebook's quality and the coding approach itself are two fundamental factors that affect the final classification performance. Therefore, researchers have carried out extensive studies on both topics. Examples are as follows. Jurie et al. [11] described a scalable acceptance-radius cluster based on mean shift approach, associating with densely sampled local patches. The algorithm is able to generate better codebooks and significantly outperforms the traditional codebook model constructed by using *k-means* techniques when applied on several popular image datasets. Wu et al. [12] demonstrated that the Histogram Intersection Kernel (HIK) is more effective than the Euclidean distance to serve as a similarity measurement in creating codebook. As a consequence, the approach can be used to improve the generation of visual codebooks.

The coding result represented as a histogram over visual codes can be applied in the following machine learning process and then directly influences the image classification accuracy. Intuitively, Csurka et al. [1] mapped each local feature onto its nearest code. This strategy may perform unfavorably because of the codebook's over-fitting nature or the large quantization error caused by hard assignment. Yang et al. [13] proposed a spatial pyramid matching approach based on SIFT sparse codes (ScSPM) for image classification. It used selective sparse coding instead of traditional vector quantization (VQ) technique to extract salient properties of appearance descriptors of local image patches. Yu et al. [14] revealed that sparse coding is helpful for learning only if the codes are local. They presented a *Local Coordinate Coding* (LCC) approach which encouraged the coding to be local, and turned a difficult nonlinear learning problem into a simple global linear learning problem. Experimental results have shown that locality is more essential than sparsity. Motivated by LCC, Wang et al. [15] presented an effective coding scheme called *Locality-constrained Linear Coding* (LLC), which can be seen as a fast implementation of LLC. It firstly performed a *k-nearest-neighbor* (KNN) search and then solved a constrained least square fitting problem. LLC algorithm finally projected each feature descriptor onto several similar visual words from a codebook. It overcomes the drawback of sparse coding which neglects features' locality, and also speeds up the computation process. Gemert et al. [16] indicated that traditional codebook model ignored the ambiguity of visual feature, that is, codeword uncertainty and codeword plausibility. It successfully introduced uncertainty modeling technique in the coding process by using kernel density estimation, and remarkably improved the categorization performance. Huang et al. [18] explored the relationships of the codes and constructed a codebook graph. The richer information contained in the graph is helpful to achieve excellent performance stably and robustly.

Although the above studies are able to generate better visual codebooks than the basic BoF model, they still suffer from weak discriminative power and strong ambiguity of the low-level visual features. As we may notice, the existing approaches all deal with the coding process by computing the nearest code or codes to the image low-level features. Therefore, the classification performance of feature histograms mainly depends on the representative power of the visual words. Nevertheless, by simply using low-level image features, we may not assure that image features with different high-level semantics can be certainly mapped onto different visual words. This disadvantage makes the generated visual codebook usually over-completed and inconsistent to the high level semantics.

3. Visual word coding based on difference maximization

Constructing the bag-of-features from the images consists of the following steps: (1) Detect local patches (regions/points of interest) automatically, (2) compute local descriptors over these patches, (3) quantize the local descriptors into “words” to construct the visual vocabulary, i.e., codebook, and (4) find the occurrences of each visual word in the codebook in each image to build the bag-of-features (in fact, histogram of words).

Let $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]\in R^{D\times N}$ denote a set of D-dimensional local features extracted from an image, where $\mathbf{x}_i$ stands for the ith feature vector. Given a codebook with M clustered visual words, $\mathbf{B}=[\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_M]\in R^{D\times M}$, an image is then encoded by a set of codes $\mathbf{C}=[\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_N]$, where M-dimensional vector $\mathbf{c}_i$ indicates the code for $\mathbf{x}_i$.

3.1. Vector quantization

Traditional hard assignment coding method is generally implemented by solving a constraint problem described in Eq. (1). The constraint ensures that only a single component of the coding vector $\mathbf{c}_i$ for $\mathbf{x}_i$ equals to 1 and all other components equal to 0. The goal is to find the nearest code for each feature.

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{N} \|\mathbf{x}_i-\mathbf{B}\mathbf{c}_i\|^2,$$
$$s.t.\forall i, \mathbf{c}_i\geq 0, \|\mathbf{c}_i\|_{l^0}=1, \|\mathbf{c}_i\|_{l^1}=1. \quad (1)$$

Many recent studies, for example Refs. [13,14], have shown that vector quantization suffers from its computational complexity and large quantization error. Another drawback was also indicated in Ref. [16] that it may raise word uncertainty and word plausibility problems.