Image and Vision Computing 38 (2015) 65–74

Contents lists available at ScienceDirect

Image and Vision Computing

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

Editor's Choice Article

How to use Bag-of-Words model better for image classification☆

Chong Wang [1], Kaiqi Huang *

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 18 September 2014
Accepted 28 October 2014
Available online 4 December 2014

Keywords:
Image classification
Bag-of-Words
Visual vocabulary
Coding
Pooling

ABSTRACT

The Bag-of-Words (BoW) framework is well-known in image classification. In the framework, there are two essential steps: 1) coding, which encodes local features by a visual vocabulary, and 2) pooling, which pools over the response of all features into image representation. Many coding and pooling methods are proposed, and how to apply them better in different conditions has become a practical problem. In this paper, to better use BoW in different applications, we study the relation between many typical coding methods and two popular pooling methods. Specifically, complete combinations of coding and pooling are evaluated based on an extremely large range of vocabulary sizes (16 to 1*M*) on five primary and popular datasets. Three typical ones are 15 Scenes, Caltech 101 and PASCAL VOC 2007, while the other two large-scale ones are Caltech 256 and ImageNet. Based on the systematic evaluation, some interesting conclusions are drawn. Some conclusions are the extensions of previous viewpoints, while some are different but important to understand BoW model. Based on these conclusions, we provide detailed application criterions by evaluating coding and pooling based on precision, efficiency and memory requirements in different applications. We hope that this study can be helpful to evaluate different coding and pooling methods, the conclusions can be beneficial to better understand BoW, and the application criterions can be valuable to use BoW better in different applications.

1. Introduction

Image classification is a fundamental problem in computer vision. It plays a key role in many applications such as image analysis and visual surveillance. In recent years, the Bag-of-Words (BoW) model has been widely used on many popular datasets and competitions, e.g., 15 Scenes [1], Caltech 101 [2], Caltech 256 [3], PASCAL VOC [4] and ImageNet [5]. In BoW, local features are first extracted to construct image representation, which is then fed into a classifier, as shown in Fig. 1. Specifically, the representation is an essential part, which includes two steps:

Coding: Coding means that local features are encoded by a vocabulary and the response of the feature on the vocabulary is generated. The probabilistic strategies [6–9] describe the distribution of local features, while sparse coding methods [10–15] better reconstruct the features. Recently, superior performance has been obtained by some high-dimensional coding methods [16–19].

Pooling: Pooling transforms the response of all local features on a vocabulary into image representation, which is fed into a classifier. Average pooling [6] and maximum pooling [10] are widely used. Recently, weighted average pooling [17] and local pooling [20] have shown better results.

Although many coding and pooling methods have been proposed, there are limited guidelines about how to use them in different applications [21–24]. Boureau et al. [21,22] analyze theoretically how coding and pooling are related based on sample cardinality (the number of local features) under small vocabulary sizes; Chatfield et al. [23] and Huang et al. [24] evaluate typical coding methods under relatively larger vocabulary sizes, but without considering different pooling schemes. Besides, all studies do not evaluate coding and pooling on large-scale datasets for generalization, such as the ImageNet database [5]. Different from the previous studies, in this paper, we consider four aspects:

- To provide systematical user guidelines, the complete combinations of more popular coding methods [15,14] and two popular pooling methods (average, maximum) under an extremely large range of vocabulary sizes (16 to 1*M*) are considered. The maximum vocabulary size (1*M*) is 1000 and 40 times larger than 1024 in [21,22] and 25*k* in [23] respectively.
- Given the fact that large-scale image classification has become much more active in recent years [25–27], we consider two large-scale datasets, namely Caltech 256 [3] and ImageNet [5]. Furthermore, combined with three typical ones including 15 Scenes, Caltech 101 and PASCAL VOC 2007, the evaluation on these primary datasets can provide strong support and generalization for the conclusions and guidelines.

☆ Editor's Choice Articles are invited and handled by a select rotating 12 member Editorial Board committee. This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author at: Zhong Guan Cun East Road #95, Haidian District, Beijing, China.
E-mail addresses: cwang@nlpr.ia.ac.cn (C. Wang), kqhuang@nlpr.ia.ac.cn (K. Huang).

[1] Permanent address: Hai Dian Qu Zhong Guan Cun Dong Lu 95 Hao, Beijing, China.

http://dx.doi.org/10.1016/j.imavis.2014.10.013
0262-8856/© 2014 Elsevier B.V. All rights reserved.

- Based on experimental results, the relation between coding and pooling is analyzed from different regimes of classification performance. In different regimes, the combinations of coding and pooling have different influence on the classification performance. Besides, these conclusions and guidelines are validated on various vocabulary construction methods for their strong generalization.
- To use the BoW model conveniently in practical applications, we provide detailed application criterions by selecting the appropriate pairs of coding and pooling methods. These criterions are given based on precision, efficiency and memory requirements, and we summarize these three factors as guidelines for some typical applications.

There are three contributions in this paper:

- *Systematic evaluation*. In this paper, complete combinations of more coding and pooling methods, an extremely large range of vocabulary sizes, primarily typical and large-scale datasets constitute a systematic evaluation. This evaluation compares many coding and pooling methods on primary datasets, and it is convenient to use appropriate methods.
- *Interesting conclusions*. In this paper, we draw some conclusions about coding and pooling. Some of them are different from the previous viewpoints [21–23], while some have never been found before but have shown importance in better applying the BoW model in practice. Particularly, extremely large sizes and large-scale datasets are important to draw these conclusions and improve the generalization ability.
- *Application criterions*. In this paper, based on the conclusions, the detailed application criterions of the BoW model are provided based on precision, efficiency and memory requirements. These criterions can be helpful for researchers and industry community to use appropriate coding and pooling methods in different applications.

The rest of this paper is organized as follows. Section 2 first introduces the related work on coding and pooling. Then, detailed experimental setups are presented in Section 3, and conclusions are drawn in Section 4. Besides, application criterions are provided in Section 5. Finally, Section 6 gives conclusive remarks.

2. Related work

In this section, the related work on coding and pooling is presented. Let $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}] \in \mathfrak{R}^{D \times N}$ be a set of N local features, and $\mathbf{C} = [\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_M}] \in \mathfrak{R}^{D \times M}$ be a visual vocabulary with M visual words. For a local feature $\mathbf{x_i}$, the response on $\mathbf{C}$ is $\mathbf{R_i} = [r_1, r_2, \ldots, r_M] \in \mathfrak{R}^{1 \times M}$. For a visual word $\mathbf{c_j}$, the cluster weight and covariance matrix are w_j and $\sigma_\mathbf{j}$ respectively. Besides, λ is a penalty term in sparse coding based methods. Table 1 summarizes some popular coding methods, and some other variables are explained in the footnote below Table 1.

2.1. Coding

In the past decade, many feature encoding methods have been proposed in the literature of image classification. Hard Quantization (HQ) [6] efficiently represents each local feature by the nearest visual word, but it obtains good performance only under large vocabulary sizes [16]. To overcome the limitation, Fisher Kernel (FK) [16] extends HQ by applying a Gaussian mixture model to approximate the distribution of local features, and it shows good results under small sizes. However, assigning continuous local features to discrete visual words causes ambiguity [8]. To model the ambiguity, Soft Quantization (SQ) [8] describes each local feature by applying a Gaussian kernel on the Euclidean distance between the feature and a vocabulary. Recently, Liu et al. have observed the locality of local features in underlying manifolds, so localized SQ [9] is proposed by only considering each feature's neighbor words. However, with average pooling, HQ and SQ may not reconstruct local features precisely, which can be important in feature encoding [10–12,17,18].

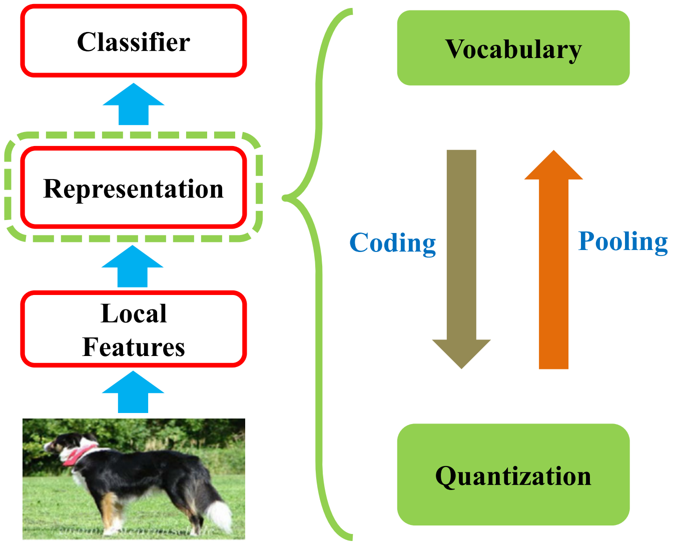

Fig. 1. The framework of the BoW model. Firstly, local features are extracted and clustered to obtain a vocabulary. Then, the features are encoded on the vocabulary to generate feature response. Finally, all the response is pooled over to construct image representation, which is fed into a classifier.

To reconstruct local features precisely, Sparse Coding (SC) [10] represents the features by a visual vocabulary sparsely. Combined with maximum pooling and the spatial pyramid matching (SPM) [1] model, SC can work well with the efficient linear SVM. Recently, empirical studies have shown that the high-dimensional representation constructed by SC can obtain superior performance [14], thus Over-complete Sparse Coding (OSC) [14] enhances the efficiency of SC by softly partitioning the feature space into some sub-manifolds. Based on SC, Yu et al. observe that locality is essential, so Local Coordinate Coding (LCC) [11,28] considers feature encoding in a local manifold, but it has high computational complexity. To implement LCC efficiently, Wang et al. propose the Local-constrained Linear Coding (LLC) [12], which has analytical solution. However, Gao et al. observe that SC, LCC and LLC do not consider the dependence of local features, thus Laplacian Sparse Coding (LSC) [13] enhances the robustness of feature encoding. Recently, some other sparse coding methods [29–36] have also shown good results.

Except for the above-mentioned methods, Huang et al. observe that the sparse coding based methods are saliency oriented, so they propose the Salient Coding (SaC) [15] which has shown competitive performance. To further reduce the reconstruction error of local features, Super Vector Coding (SV) [17] extends HQ to a much higher dimensional feature space, and Improved Fisher Kernel (IFK) [18] enhances FK by power normalization, and both SV and IFK have obtained superior performance [23]. To exploit these high-dimensional methods practically, Hervé Jégou et al. propose the Vector of Locally Aggregated Descriptors (VLAD) [19], which enjoys the high efficiency and low memory requirements jointly by the optimization of dimensionality reduction and an indexing algorithm. Based on VLAD, Picard and Gosselin [37] propose the Vectors of Locally Aggregated Tensors (VLAT) to improve image similarity, and it has shown better performance against VLAD. To analyze these high-dimensional coding methods in a general way, Zhao et al. [38] propose a unified framework to perform coding via vector difference.

Table 1[2] summaries some typical coding methods based on sparsity, locality and efficiency. Sparsity means that only a few words have large

[2] In SQ, β is the Gaussian smoothing factor, which is also used in LLC with the number of nearest words set to be K. In LLC, $\odot$ denotes the element-wise multiplication. In LSC, α penalizes the feature similarity, in which the similarity matrix $\mathbf{S}$ is included. In OSC, p_j is the posterior probability that $\mathbf{x_i}$ is assigned to $\mathbf{c_j}$ in GMM, and L denotes the number of primary clusters. Besides, $\mathbf{C_j}$ is the jth secondary cluster and $\mathbf{R_i^j}$ is the corresponding response. Finally, in SVC, s is a small constant determined by cross-validation.