CODEBOOK OPTIMIZATION USING WORD ACTIVATION FORCES FOR SCENE CATEGORIZATION

Qun Li[1,2], *Honggang Zhang*[1], *Jun Guo*[1], *Le An*[2], *Bir Bhanu*[2]

[1]Pattern Recognition and Intelligent System Laboratory,
Beijing University of Posts and Telecommunications, Beijing, China
[2]Center for Research in Intelligent Systems, University of California, Riverside, CA, USA
{ liqun, zhhg, guojun }@bupt.edu.cn, lan004@ucr.edu, bhanu@ee.ucr.edu

ABSTRACT

Visual codebook based quantization of robust appearance descriptors extracted from local image patches is an effective means of capturing image statistics for texture analysis and natural scene classification. In this paper, based on the newly proposed statistics of word activation forces (WAFs), we optimize the codebook. Currently, codebooks are typically created from a set of training images using a clustering algorithm. However, these codebooks are often functionally limited due to redundancy. We show that WAFs can remove the redundancy efficiently. In the experiment, the proposed method achieved the *state-of-the-art* performance on the Caltech-101, fifteen natural scene categories and VOC2007 databases. The optimization method also offers insights into the success of several recently proposed images classification approaches, including vector quantization (VQ) coding in the Spatial Pyramid Matching (SPM), sparse coding SPM (ScSPM), and Locality-constrained Linear Coding (LLC).

Index Terms— Scene categorization, word activation forces, codebook

1. INTRODUCTION

Currently, a typical image classification system is based on bag-of-words (BoW) model [1, 2] combined with spatial pyramid matching (SPM) [3]. The BoW model represents an image as a histogram of its local features. It is robust against spatial translations of features, and demonstrates decent performance in whole-image categorization tasks. However, due to the lack of the information about the spatial layout of the features, original BoW model is incapable of capturing shapes or locating an object. Hence, a simple BoW model shows limited capacity in more complicated tasks, such as natural scene classification.

To overcome the above limitation, various extensions of the BoW have been proposed, among them SPM is the most successful approach. The resulted "spatial pyramid" is a computationally efficient extension of the orderless BoW representation, which works well for image classification. To future achieve better performance, SPM needs to be combined with the use of nonlinear Mercer kernels, e.g., Chi-square kernel.

To improve the scalability, researchers aimed at obtaining nonlinear feature representations that work better with linear classifiers, e.g. [4, 5]. In particular, Yang *et al.* [5] proposed the ScSPM method where sparse coding (SC) was used instead of vector quantization (VQ) to obtain nonlinear codes. Yu *et al.* [6] empirically observed that SC results tend to be local-nonzero coefficients are often assigned to bases nearby to the encoded data. Wang presented a simple but effective coding scheme called Locality-constrained Linear Coding (LLC) [7] in place of the VQ coding in traditional SPM. LLC utilizes the locality constraints to project each descriptor into its local-coordinate system, and the projected coordinates are integrated by max pooling to generate the final representation. With linear classifier, the proposed approach performs remarkably better than the traditional nonlinear SPM, achieving the *state-of-the-art* performance on several benchmarks.

Having achieved these progresses, one began to pay attention to a more basic problem: codebook construction. A codebook approach gives a set of discrete visual words, used in the BoW model. Traditionally, codebooks were usually constructed by using an unsupervised method such as k-means to cluster the descriptor vectors of patches sampled either densely or sparsely from a set of training images. This kind of methods can work well for texture analysis on images containing only a few homogeneous regions, but cannot guarantee to obtain an optimal codebook for a complicated application situation, such as natural scenes. Gemert *et al.* [8] showed that such a kind of methods of codebook construction have two drawbacks: codeword uncertainty and codeword plausibility, and proposed the kernel codebooks to improve categorization performance. Separately, sparse coding SPM (ScSPM) image classification systems used sparse coding to generate codebooks, and INRIA adopted a supervised dictionary learning method [9].

Many experiments by researchers suggest that a codebook constructed by the traditional method may be redundant. This

problem not only increases the computation time of the system, but it also corrupts the representation of image, decreasing the accuracy for classification. Therefore, removing the redundancy in the codebook is important. To this end, we utilize our newly proposed statistics of word activation forces (WAFs) to optimize the codebook. Through this step, we reduced the codebook to almost half of the original size and kept the accuracy equal or higher. The improved efficiency significantly enhances the practical value of the image classification method for real applications.

The remainder of the paper is organized as follows: Section 2 introduces the basic idea of WAFs and a codebook optimization algorithm, using WAFs, is proposed to reconstruct the codebook; Experimental results on three widely used datasets are reported in Section 3; and Finally in Section 4 conclusions are made.

2. WAFS BASED CODEBOOK OPTIMIZATION

2.1. Word activation forces

The WAFs [10] is defined as:

$$WAF_{ij} = (f_{ij}/f_i)(f_{ij}/f_j)/d_{ij}^2. \quad (1)$$

Specifically, given the frequencies f_i and f_j and co-occurrence frequency f_{ij} of a pair of words i and j in the corpus, we predict the strength of the activation that word i exerts on word j through the statistic WAF_{ij}, where d_{ij} is the average distance by which word i precedes word j in their co-occurrences. Seeing the ratios of f_{ij} to f_i and f_{ij} to f_j as masses, we identify that the statistic is defined in the same form of the universal gravitation.

Therefore, we name it as word activation force from i to j, shortly WAF_{ij}. According to the definition, the magnitude of WAFs is restricted to [0, 1]. WAFs is a type of statistics to weight the links of a complex network and thereby developing a desired affinity measure. It is shown that the approach is superior in facilitating the analysis through experiments on a large-scale word network. The experiment on the word network verifies that the measured word affinities are highly consistent with human knowledge.

In [10], the authors give the affinity measure between the words in the WAFs. For a pair of words i and j in the directed word network WAFs, we define their affinity as:

$$A_{ij}^{waf} = \left[\frac{1}{|K_{ij}|}\sum_{k\in K_{ij}} OR(waf_{ki}, waf_{kj}) \cdot \frac{1}{|L_{ij}|}\sum_{l\in L_{ij}} OR(waf_{il}, waf_{jl})\right]^{1/2}, \quad (2)$$

where
$K_{ij} = \{k|waf_{ki} > 0 \quad or \quad waf_{kj} > 0\}, L_{ij} = \{l|waf_{il} > 0 \quad or \quad waf_{jl} > 0\}$, and $OR(x, y) = min(x, y)/max(x, y)$, which is defined as the geometric average of the mean overlap rates of the in-links and out-links of the inquired two words.

Algorithm 1 WAFs based codebook optimization algorithm

Input: $B_{init} \in \mathbb{R}^{D\times M}$, $thresh1$, $thresh2$.
Output: B.

```
B ← B_init.
for each i ∈ [1, D] do
    for each j ∈ [1, D] do
        calculate occurrence f_i of visual word i using thresh1;
        freq_index(i) ← j;
    end for
end for
for each i ∈ [1, D] do
    calculate co-occurrence f_ij of visual words i and j;
end for
for each i ∈ [1, D] do
    for each j ∈ [1, D] do
        WAF_ij ← (f_ij/f_i)(f_ij/f_j)/d_ij^2;
    end for
end for
for each i ∈ [1, D] do
    for each j ∈ [1, D] do
        if WAF_ij < thresh2 then
            remove visual word j of B_init to B_opti;
            B ← B_opti.
        end if
    end for
end for
```

2.2. WAFs based codebook optimization algorithm

In this paper, we adapt the definition of WAFs to optimize the codebook in an image classification system, which is named WAFs based codebook optimization algorithm. The scheme is described as follows:
(1) Generate initial codebook B_{init} using K-means clustering or other methods.
(2) Count occurrences f_i, f_j, and co-occurrences f_{ij} of visual words i and j. In this step, similar visual words are used to calculate the occurrences of the codeword. We give a threshold *thresh1* to decide similar, such as twenty percent of the average value of all visual words distances.
(3) Calculate WAFs matrix of codebook according to Eq. (1), where d_{ij} is Euclidean distance, and remove one from a pair of visual words with high affinities. We give another threshold *thresh2* to decide high affinities. Finally, take the remained codebook *B* as new codebook.
(4) Classify using new codebooks and compare the results with original ones.

The above process is illustrated in Alg. 1. Fig. 1 shows the overall classification process. Left is the flowchart of clas-

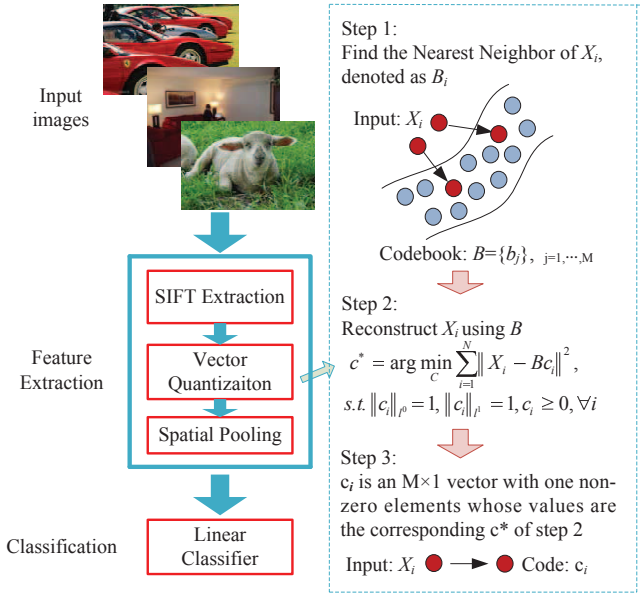

Fig. 1: Left: flowchart of classification system. Right: VQ coding process.

sification process and right is the VQ coding process in the traditional SPM, denoted as VQSPM.

3. EXPERIMENTAL RESULTS

In this section, we report results on three diverse datasets: fifteen scene categories [11], Caltech-101 [12], and Pascal VOC 2007 [13]. We compare our results with several *state-of-the-art* methods [3, 5, 7].

The classification methods we used are VQSPM, ScSPM, and LLC. We use only a single descriptor, the SIFT descriptors of 16×16 pixel patches computed over a grid with a spacing of 8 pixels, and 4×4, 2×2, 1×1 sub-regions for SPM, throughout all the experiments. Our decision to use a dense regular grid instead of interest points was based on the comparative evaluation of Fei-Fei and Perona [11], who have shown that dense features work better for scene classification. In our setup, we use linear SVM as the classifier. We partition the whole dataset of Caltech-101 into 30 training images per class and the rest for testing images, except 100 training images per class for the Scene 15.

3.1. Caltech-101

Our first set of experiments is on the Caltech-101 database, which contains 9144 images in 101 classes. The number of images per category varies from 31 to 800. Most images are of medium resolution, *i.e.*, about 300×300 pixels.

We perform k-means clustering of a random subset of patches from the training set to form a visual vocabulary. Codebook sizes for our experiments are 1024 and 2048, and we use VQSPM and LLC classification algorithms. In order to show that the conclusions can be generalized to other codebook generation method as well, specially, we trained a codebook with 1024 bases using sparse coding as described in [5] for ScSPM.

We compare our result with un-optimized codebooks. Detailed results are shown in Table 1. It can be seen that in most cases, while keeping the accuracy higher than or equal to the initial values, the proposed method reduces the codebook size by $\sim 50\%$. As a result, the average processing time reduces to 1/2 of the original at least, and the method has a significant reduction in running memory. This efficiency significantly adds to the practical values of image classification methods for real applications. We can also see that the larger the codebook there is more necessity of the codebook optimization. Furthermore, our method reduces the size of codebook to 612 cases from 1024 cases, which is generated by sparse coding, but 588 or 539 by the K-means clustering. This confirms that SC training method generates better codebook than the K-means clustering.

3.2. Scene Category Recognition

The second dataset is composed of fifteen scene categories: thirteen were provided by Fei-Fei and Perona [11], and two were collected by Svetlana. Each category has 200 to 400 images, and the average image size is 300×250 pixels. The major sources of the pictures in the dataset include the COREL collection, personal photographs, and Google image search. This is one of the most complete scene category dataset used in the literature so far.

The initial codebooks were generated by K-means clustering with 400 bases and 1000 bases for VQSPM and LLC classify algorithms. With the same purpose of before, we trained a codebook with 400 bases using sparse coding.

Table 2 shows that the larger the codebook is not always better, sometimes it increases the redundancy. Our method can remove the redundancy significantly.

3.3. Pascal VOC 2007

The PASCAL 2007 dataset consists of 9,963 images from 20 classes. The dataset is an extremely challenging one because all the images are daily photos obtained from Flicker where the size, viewing angle, illumination, etc appearances of objects and their poses vary significantly, with frequent occlusions. The classification performance is the standard metric used by PASCAL challenge [13]. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance.

We try our algorithm on LLC to classify with 1024 bases which are trained by K-means clustering. In Table 3, we list scores for all the 20 classes. The results are very impressive: our method optimizes the codebook to 497 bases. In addition,

Table 1: Codebook optimization and Image classification results on Caltech101 database

Method	Codebook size		Accuracy	
	before optimization	after optimization	before optimization	after optimization
VQSPM	1024	**539**	47.6	**51.8**
	2048	**871**	48.8	**52.7**
LLC	1024	**588**	69.2	**69.5**
	2048	**968**	68.1	**70.7**
ScSPM	1024	**612**	71.7	**71.8**

Table 2: Codebook optimization and Image classification results on Scene15 database

Method	Codebook size		Accuracy	
	before optimization	after optimization	before optimization	after optimization
VQSPM	400	**301**	74.2	**74.2**
	1000	**280**	73.9	**74.9**
LLC	400	**322**	80.2	**80.3**
	1000	**344**	73.5	**80.6**
ScSPM	400	**321**	74.0	**74.2**

under almost all the cases, the scores are equal or higher than those before optimization.

4. CONCLUSIONS

In this paper, we presented a new approach that optimizes visual codebook to improve computing efficiency and performance of a classification system. The approach uses a new criterion called word activation forces (WAFs) to guide the codebook optimization, in order to improve the classification and it's efficiency. We performed experiments on various image databases to indicate the benefits of the proposed method. Experimental results showed that the proposed method obtained higher efficiency while keeping the accuracy of classification compared to the *state-of-the-art* methods.

Acknowledgements. This work was partially supported by National Natural Science Foundation of China under Grant No.61005004 and 61175011, Chinese 111 program Advanced Intelligence and Network Service under Grant No.B08004 and a key project of the Ministry of Science and Technology of China under Grant No.2011ZX03002-005-01.

5. REFERENCES

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003, pp. 1470–1477.

Table 3: Image classification results using VOC2007 database

object class	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow
before optimization	62.0	46.2	29.6	49.4	19.2	35.6	62.9	37.6	41.3	29.5
after optimization	65.0	46.1	29.7	49.4	19.4	36.1	63.2	37.6	41.2	29.5

object class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
before optimization	30.2	29.9	68.1	46.9	72.1	15.2	27.8	39.6	58.9	36.7
after optimization	30.8	30.0	68.0	47.2	72.0	17.9	26.8	39.8	58.8	37.5

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. of ECCV*. Workshop on Statistical Learning in Computer Vision, 2004, pp. 1–22.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of CVPR*, 2006, pp. 2169 – 2178.

[4] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang, "Hierarchical gaussianization for image classification," in *Proc. of ICCV*, 2009, pp. 1971 – 1977.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of CVPR*, 2009, pp. 1794 – 1801.

[6] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. of NIPS*, 2009.

[7] J. Wang *et al*, "Locality- constrained linear coding for image classification," in *Proc. of CVPR*, 2010, pp. 3360 – 3367.

[8] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. of ECCV*, 2008, pp. 696–709.

[9] J. Mairal and F. Bach, "Supervised dictionary learning," in *Proc. of NIPS*, 2008, pp. 1033–1040.

[10] J. Guo, H. Guo, and Z.Wang, "An activation force-based affinity measure for analyzing complex networks," in *Sci. Rep. 1, 113; DOI:10.1038/srep00113 (2011)*. http://www.nature.com/srep/2011/111012/srep00113/full/srep00113.html.

[11] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. of CVPR*, 2005, pp. 524 – 531.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004, p. 178.

[13] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*.