IMPROVED CLUSTER CENTER ADAPTION FOR IMAGE CLASSIFICATION

Mingmin Zhen, Wenmin Wang, Ronggang Wang*

School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
mingminzhen@pku.edu.cn, {wangwm, rgwang}@ece.pku.edu.cn

ABSTRACT

The feature coding algorithm, "Vector of Locally Aggregated Descriptors (VLAD)", can be used effectively for large scale object instance retrieval. Despite its effectiveness and excellent performance, the existence of ambiguous cluster centers can reduce the performance. Though an idea to this problem has been proposed, it is not practical in fact. In this paper, we analyze possible situations that cause effect on the results and propose a novel approach to improve the VLAD method. The proposed method mainly focuses on the similarity measure between each two images. For each two images, we adapt the original cluster center to VLAD vectors. As we illustrate, our method has promising results with small vocabulary size on both datasets of 15 Scenes and VOC2007.

Index Terms— Image classification, Feature coding, Image retrieval, Image representation

1. INTRODUCTION

As a fundamental problem in computer vision, image classification has been paid vast attention to. To do this task, it is necessary to construct an effective image representation. A-mong various approaches to this purpose, the Bag-of-Features (BoF) [1] model is probably one of the most widely-used. For a given image, the BoF based method extracts a set of local patches by interest point detection or dense sampling, and use local feature descriptors (e.g. SIFT) to represent them. Then, the descriptors are voted on the vocabulary which is obtained by clustering (e.g. K-means). At last, the voting results are represented as a histogram vector for classification.

Recently, several attempts have been made to improve the accuracy of BoF, in which Fisher Vector (FV) [2, 3, 4] and its simpler variant VLAD (Vector of Locally Aggregated Descriptors) [5, 6, 7] are among the most successful methods. VLAD approach has the advantages of both BoF and FV. Since VLAD needs less memory and obtains higher performance compared with other coding methods, it can be used in the field of image search on a large scale.

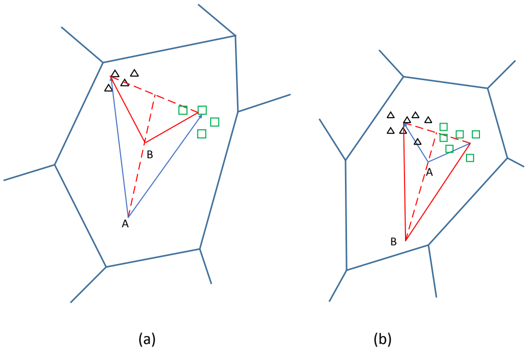

Fig. 1. **Two situations for VLAD similarity measure.** Similar to [8], the Voronoi cells represent the clusters used to construct VLAD vectors. The feature point A corresponds to the cluster center in both (a) and (b). Black triangles and green squares correspond to local features extracted from different images, while the blue arrows correspond to the sum of residuals. The original cluster center A is not consistent for both (a) and (b). For situation (a) and (b), a good approach is to change the cluster center from A to B.

However, despite the popularity of VLAD, there always exists a problem of vocabulary sensitivity, i.e. the fact that the similarity between two VLAD vectors is highly dependent on the visual vocabulary. An idea called *cluster center adaption* [8] is proposed to solve this problem and to improve the performance, but it is not practical in fact [9].

To solve the problem of vocabulary sensitivity, we first analyze two possible situations (as Figure 1 shows) that can cause ambiguity, and then propose a method to deal with these situations. Since the essence of image classification is to make the "distance" (e.g. Euclidean distance, cosine distance) between similar images shorter and different images longer, we take the similarity measure of each two images into account. We propose a method called Improved Cluster Center Adaption to improve the VLAD method (ICCA-VLAD). In

This project was supported by Shenzhen Peacock Plan (20130408-183003656).

the experiments, our method is shown to outperform the original VLAD significantly.

The rest of the paper is organized as follows. In Section 2, we introduce the Fisher vector and VLAD. And we analyze the similarity measure and illustrate the ICCA-VLAD method. Then experiments and results are shown in Section 3. At last, our conclusions are shown in Section 4.

2. METHODS

For an image classification task, firstly a set of low-level feature descriptors $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^D, i = 1, ..., N\}$ are extracted from an image in traditional BoF method, where D is the feature descriptor dimension and N is the number of local features. Then a vocabulary or codebook $\mathbf{C} = \{\mathbf{c}_j \in \mathbb{R}^D, j = 1, ..., K\}$, where K is the number of codewords, is obtained by a clustering method. At last the image representation is the responses on the vocabulary. In the following subsections we will briefly introduce the Fisher vector and VLAD, and then illustrate the similarity measure between two VLAD vectors and the improved cluster center adaption.

2.1. Fisher vector

In order to take advantage of both generative and discriminative models, the Fisher kernel is used to construct Fisher vector (FV) for image representation [2, 3]. Instead of using BoF, FV includes the high order statistic information which describes the probability distribution of low level feature descriptors.

In the vocabulary constructing step, Gaussian Mixture Model (GMM) is used. Suppose there are K Gaussians, $\Theta = \{\boldsymbol{\alpha}_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, j = 1, ..., K\}$ denotes the parameters of the GMM. In the parameters, $\boldsymbol{\alpha}_j, \boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ denote the weight, the mean vector and the covariance matrix of the j-th Gaussian, respectively. The main idea of FV is computing the gradient vectors with respect to the mean and the variance vector of the j-th Gaussian

$$\mathcal{G}^j_{\boldsymbol{\mu},i} = \frac{1}{N\sqrt{\boldsymbol{\alpha}_j}} \sum_{i=1}^{N} \omega_{i,j} (\frac{\mathbf{x}_i - \boldsymbol{\mu}_j}{\boldsymbol{\sigma}_j}) \tag{1}$$

$$\mathcal{G}^j_{\boldsymbol{\sigma},i} = \frac{1}{N\sqrt{2\boldsymbol{\alpha}_j}} \sum_{i=1}^{N} \omega_{i,j} (\frac{(\mathbf{x}_i - \boldsymbol{\mu}_j)^2}{\boldsymbol{\sigma}_j^2} - 1) \tag{2}$$

where $\omega_{i,j}$ is the soft-assignment of the i-th descriptor $\mathbf{x}_i$ to the j-th Gaussian and $\boldsymbol{\sigma}_j$ is the variance vector of the diagonal covariance matrix $\boldsymbol{\Sigma}_j$ as we assumed. Thus, the final representation of FV in [3] is a 2DK-dimensional vector if the Spatial Pyramid Matching (SPM) [7] is not considered.

2.2. VLAD

When considering the efficiency in large scale image search problems, VLAD is proposed to simplify FV. VLAD uses a clustering algorithm to generate a vocabulary. Then the descriptors are voted on their nearest codewords. The response on each codeword is the accumulation of the difference between the nearest descriptors of the codeword and itself. For a descriptor $\mathbf{x}_i$, the vector differences between $\mathbf{x}_i$ and K codewords are calculated. Then the vector difference based coding result is constructed as a DK-dimensional feature vector

$$\phi(\mathbf{x}_i) = [\lambda_1(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_1)^T, ..., \lambda_K(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_K)^T] \tag{3}$$

where $\lambda_j(\mathbf{x}_i) = 1$ if $\mathbf{c}_j$ is the nearest codeword of $\mathbf{x}_i$ and $\lambda_j(\mathbf{x}_i) = 0$ otherwise. Based on SPM [7], the descriptors belongs to the same block are accumulated by using sum method to generate a block vector representation.

$$\boldsymbol{\omega}_m = \sum_{i=1}^{N} \phi(\mathbf{x}_i) * \nu_i \tag{4}$$

where $\nu_i = 1$ if the descriptor $\mathbf{x}_i$ is in the m-th block and $\nu_i = 0$ otherwise. At last, all block vectors are concatenated to construct a final MDK-dimensional vector $\mathbf{z}$, where M is the number of SPM blocks (including all levels) as image representation.

$$\mathbf{z} = [\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M] \tag{5}$$

2.3. Analysis of similarity measure

For a binary image classification problem, an SVM (Support Vector Machine) aims to learn a decision function

$$f(\mathbf{z}) = \sum_{i=1}^{n} \alpha_i \kappa(\mathbf{z}_i, \mathbf{z}) + b \tag{6}$$

where $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$ is the training set, and $y_i \in \{-1, +1\}$ indicates labels. Given a test image represented by $\mathbf{z}$, if $f(\mathbf{z}) > 0$ then the image is classified as positive, otherwise negative.

In kernel function theory, $\kappa(.,.)$ can be any reasonable kernel. For computation efficiency, linear kernel is used for VLAD. So linear kernel is as follows

$$\kappa(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j = \sum_{l=1}^{M} \sum_{s=1}^{K} \langle \mathbf{z}_i(l,s), \mathbf{z}_j(l,s) \rangle \tag{7}$$

where $\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \mathbf{z}_i^T \mathbf{z}_j$, and $\mathbf{z}_i(l,s)$ is the sum of the residual vectors $\mathbf{x} - \mathbf{c}_i$ of the feature $\mathbf{x}$ assigned to each codeword $\mathbf{c}_i$. Since each vector $\mathbf{z}$ is l_2-normalized, $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$ can be seen as the cosine used to measure similarity between corresponding two images. The similarity between VLAD vectors can be decomposed as the sum of scalar products of residuals for each cluster independently.

When considering the contribution to the similarity for one cluster, it's obvious that the similarity measure induced by the VLAD descriptors increases if $\langle \mathbf{z}_i(l,s), \mathbf{z}_j(l,s) \rangle$ is positive, and decreases otherwise. In [8] one situation is described which is similar to Figure 1a. The set of descriptors are deemed to be very different thus generating a negative contribution to the similarity of the two images, but a positive contribution is obtained. So the *cluster center adaption* is proposed to adapt the cluster center to this situation. But in fact, there also exists an opposite situation in Figure 1b. The two sets of descriptors should generate a positive contribution but a negative contribution is obtained because of the ambiguous cluster center.

For these two situations, though the idea of *cluster center adaption* can be partly useful, it is not practical. We propose a method to deal with these situations. This will be detailed in the next subsection.

2.4. Improved Cluster Center Adaption

Through the introduction of FV and VLAD, we can see that both of them depend on the relationship between descriptors and vocabulary. For example, the vocabulary in FV can be the means of GMM and we denote $\mathbf{c}_j = \boldsymbol{\mu}_j, j = 1, ..., K$. As shown in Equation 1 and 2, the final representation of FV can be seen as a function of $(\mathbf{x}_i - \mathbf{c}_j)$ where $i = 1, ..., N; j = 1, ..., K$. Thus, the FV depends on the relationships between descriptors **X** and vocabulary **C**. The same is true for VLAD and intuitively, VLAD and FV are both involved in this relationship. Therefore, sensitivity to vocabulary is the drawback of both FV and VLAD. In *cluster center adaption* method, its main thought is moving the cluster center to an appropriate position based on the whole dataset. Different from it, we propose a method that is also adjusting the cluster center's position but based on just the two images.

Given two vectors $\mathbf{z}_r^k$ and $\mathbf{z}_s^k$ corresponding to a cluster k, we firstly compute the mean vector and use a parameter λ to adjust the cluster center as follows

$$\mathbf{z}_{center} = \lambda \frac{\mathbf{z}_r^k + \mathbf{z}_s^k}{2} \tag{8}$$

Then we recompute the improved adaptive VLAD representation

$$\mathbf{z}_r^k(ICCA) = \mathbf{z}_r^k - \mathbf{z}_{center} \tag{9}$$

$$\mathbf{z}_s^k(ICCA) = \mathbf{z}_s^k - \mathbf{z}_{center} \tag{10}$$

where $\mathbf{z}_r^k(ICCA)$ and $\mathbf{z}_s^k(ICCA)$ are the final ICCA-VLAD representation. Finally we compute the similarity between two images by using Equation 7 replacing $\mathbf{z}_i$ with $\mathbf{z}_i(ICCA)$.

When we compare the ICCA-VLAD representation and Figure 1, we can see that situation (a) corresponds to $\lambda > 0$, i.e. moving the cluster center from A to B. And situation (b) is opposite corresponding to $\lambda < 0$. Thus, when $\lambda = 0$, ICCA-VLAD falls back to standard VLAD.

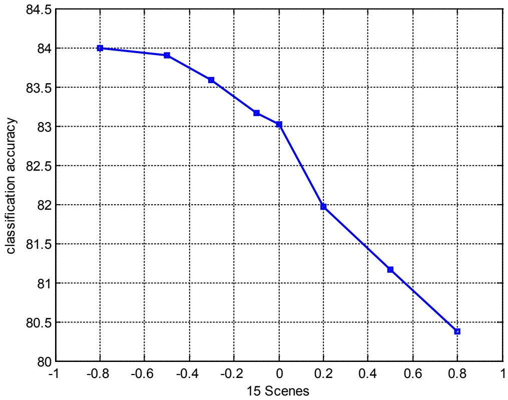

Fig. 2. The classification accuracy for different parameter λ. The corresponding vocabulary size is 64.

3. EXPERIMENTAL RESULTS

3.1. Datasets and experimental settings

We perform a series of experiments on two datasets, i.e. 15 Scenes [7] and VOC 2007 [10]. The 15 Scenes consists of 4,485 images spread over 15 categories, each of which contains 200 to 400 images. We randomly pick out 100 images from each category for training, and keep the remaining images for testing. The experiments are repeated for 10 times and average accuracy is reported. The VOC 2007 dataset is one of the most challenging databases for image classification with 9,963 images distributed in 20 classes of objects. The training and testing samples have been well divided. We follow the official experimental settings and report the mean average precision.

In all experiments, we adopt the 128-dimensional SIFT features densely extracted from images on a grid with a step of 4 pixels under three scales: 16 × 16, 24 × 24, and 32 × 32. To generate vocabulary, we use the standard K-means clustering algorithm. After all features are encoded, spatial pyramid matching is performed on three levels, i.e. 1×1, 2×2 and 3×1 for both 15 Scenes and VOC 2007. For fair comparison, all the experiments use L_2 normalization.

3.2. Experimental results

The λ in Equation 8 is an important parameter. We first conduct experiments with different λ on 15 Scenes, as reported in Figure 2. In an overall view, our proposal can achieve better performance as λ decreases and $\lambda < 0$. And when $\lambda = 0$, it falls back into standard VLAD. From the results, we can conclude that the situation (b) shown in Figure 1 affects the results more than situation (a). So we can select an optimum parameter λ for improved cluster center adaption.

We also compare the results on different vocabulary size for 15 Scenes and VOC 2007 as shown in Figure 3 and Table

method	VLAD			Ours (ICCA)		
Size	16	32	64	16	32	64
aeroplane	64.5	65.8	69.4	66.0	68.7	**71.5**
bicycle	49.9	52.5	55.6	52.2	**56.5**	56.2
bird	38.6	37.6	40.4	37.8	41.1	**43.6**
boat	63.1	64.8	67.1	63.8	66.7	**70.2**
bottle	22.3	23.8	23.3	22.1	23.8	**23.9**
bus	55.7	58.6	**63.1**	48.9	53.8	59.0
car	68.6	71.3	72.9	70.0	71.6	**73.2**
cat	51.2	52.7	55.0	53.2	53.0	**57.2**
chair	46.0	48.3	**49.0**	46.2	47.6	48.5
cow	30.9	37.5	40.2	34.4	39.9	**42.1**
diningtable	40.3	47.3	48.1	43.1	47.0	**50.7**
dog	31.2	32.7	36.1	30.8	33.8	**37.1**
horse	71.5	72.7	74.7	72.5	74.0	**75.4**
motorbike	57.5	61.6	63.1	56.2	61.9	**65.8**
person	77.0	78.6	**80.2**	78.0	79.2	80.1
pottedplant	21.5	22.6	24.3	20.9	21.4	**24.7**
sheep	40.8	37.9	42.1	37.5	40.9	**43.4**
sofa	41.7	42.0	46.4	42.8	41.7	**48.4**
train	70.4	71.4	**73.9**	70.2	71.9	73.3
tvmonitor	38.9	41.1	45.4	40.9	43.3	**46.7**
mAP	49.1	51.0	53.5	49.3	51.9	**54.6**

Table 1. Comparison of mAP between VLAD and our approach for different sizes of vocabularies on VOC 2007 dataset.

1. We use very small vocabulary size in experiments to prove the effectiveness of our method. The vocabulary size varies from 16 to 64. Our performance can also be further improved by larger vocabulary size. From the Figure, the ICCA-VLAD consistently outperforms standard VLAD. This result validates the effectiveness of the proposed method.

4. CONCLUSION

In this paper, we propose an improved method for cluster center adaption. And we also analyze two possible situations for VLAD and propose a novel approach to solve it. We obtain promising results on the popular image classification datasets 15 Scenes and VOC 2007 especially when the vocabulary is very small. Moreover, our analysis about the ambiguous cluster center can also be used to improve FV or extensions of FV and VLAD. Thus our method can be applied to this kind of problem well and is encouraging to further improvements.

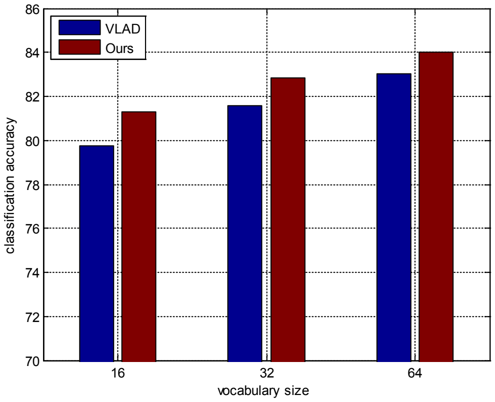

Fig. 3. The classification accuracy with vocabulary size of 16, 32 and 64 respectively.

5. REFERENCES

[1] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Bray, “Visual categorization with bags of keypoints,” in *ECCV*, 2004.

[2] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *CVPR*, 2007.

[3] Florent Perronnin, Jorge Snchez, and Thomas Mensink, “Improving the fisher kernel for large-scale image classification,” in *ECCV*. 2010.

[4] Jorge Snchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, “Image classification with the fisher vector: Theory and practice,” *International Journal of Computer Vision*, 2013.

[5] H. Jegou, M. Douze, C. Schmid, and P. Perez, “Aggregating local descriptors into a compact image representation,” in *CVPR*, 2010.

[6] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *Pattern Analysis and Machine Intelligence*, 2012.

[7] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, 2006.

[8] R. Arandjelovic and A. Zisserman, “All about vlad,” in *CVPR*, 2013.

[9] E. Spyromitros-Xioufis, S. Papadopoulos, I.Y. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, “A comprehensive study over vlad and product quantization in large-scale image retrieval,” *Multimedia*, 2014.

[10] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman, “The pascal visual object classes (voc) challenge,” *IJCV*, 2007.