A Simple but Efficient Way to Combine VLAD with Locality-constrained Linear Coding

Zhenglin Tan, Wenmin Wang *, Yifeng Jiang, Ronggang Wang

School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
tzl@sz.pku.edu.cn, *wangwm@ece.pku.edu.cn, jiangyifeng@sz.pku.edu.cn, rgwang@pkusz.edu.cn

***Abstract*—The VLAD (vector of locally aggregated descriptors) representation, derived from BoF and Fisher kernel, has shown its efficiency in the field of image search. However, assigning local descriptors to a codeword is a hard voting process, which does not consider the uncertainty and the plausibility for single codeword. In this paper, we propose an approach to combine VLAD with locality-constrained linear coding, as opposed to the original one, considering several nearest neighbors when assigning local descriptors and computing weights. In order to evaluate our proposed method, experiments are conducted on several image classification benchmarks, using VLAD for comparison. The experimental results show that our method stably outperforms VLAD in terms of classification accuracy, while producing feature representation of the same dimension without much additional computational cost.**

***Index Terms*—VLAD, soft-assignment, image classification, feature representation, LLC**

I. Introduction

Over the last decade, the bag-of-features (BoF), derived from the bag-of-words model in document analysis, has been widely used and deeply studied in image classification system.

Generally speaking, the BoF framework used for image classification consists of five basic components: local features (e.g., SIFT descriptors [1]) extraction, codebook learning with training data, feature coding with pre-trained codebook, pooling or aggregating codes over images, and learning classifier (e.g., SVM) for classification. In this framework, the key problem is how to encode local features into a vector that can well represent the whole image.

So far, developed from the hard voting [2] which is simple but limited in representing descriptors, some coding strategies to deal with quantization losses have been proposed which retain more information by expressing features as combinations of visual words (e.g., soft voting [3], [4], sparse coding [5], locality-constrained linear coding (LLC) [6], and salient coding [7]). Another extension of this popular framework includes the use of spatial pyramids [8] to take into account some aspects of the spatial layout of the image.

Recently, an alternative patch aggregation mechanism has been proposed, that is used to record the difference between the features and the visual words (e.g., super-vector coding [9], Fisher vector (FV) [10], [11], [12], and vector of locally aggregated descriptors (VLAD) [13]). As with the earliest BoF model which adopts hard voting mechanism, VLAD associates each descriptor to its nearest neighbor in the pre-trained codebook. The only difference is that VLAD aggregates residual vectors (i.e., the difference between the descriptor and the associated visual word) instead of counting occurrences followed by average pooling.

However, we find that the VLAD representation utilizes hard assignment of local descriptors to only one nearest centroid, and there is still much room for further improvement. To obtain a more powerful vector representation of an image, our attempts are made to combine VLAD with various soft-assignment schemes.

In this paper, we propose an approach to combine VLAD with locality-constrained linear coding, called VLAD-LLC, by recording the weighted difference between the descriptor and the reconstruction vector. We use VLAD incorporating spatial information as our baseline method, where an image is divided into three pyramid levels with grids of 1×1, 2×2, 3×1, as suggested in [9]. The implementation details will be discussed in the following sections.

The remainder of this paper is organized as follows. In Section II, we discuss the related works. In Section III, we introduce our method and demonstrate how it works. In Section IV, we give our parameter settings and report experimental results on three image classification datasets. Finally, Section V concludes our paper.

II. Related Work

In this section, we briefly review VLAD that produces an image representation from a set of local descriptors, then feature coding algorithm LLC and its fast implementation approximated LLC are introduced.

A. VLAD

That the vector of locally aggregated descriptors (VLAD) [13] is proposed to represent image by aggregating local features in feature space. A visual dictionary $\boldsymbol{B} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M] \in R^{D \times M}$ of M visual words is first learned using K-means clustering, then the extracted local feature $\boldsymbol{x}_t (\boldsymbol{x}_t \in R^D)$ is assigned to its nearest visual word $\boldsymbol{\mu}_i = NN(\boldsymbol{x}_t)$ in the dictionary using a distance measure. For each visual word, the residual vectors are accumulated as:

978-1-5090-5316-2/16/$31.00 © 2016 IEEE

$$\boldsymbol{v}_i = \sum_{\boldsymbol{x}_t : NN(\boldsymbol{x}_t)=\boldsymbol{\mu}_i} \boldsymbol{x}_t - \boldsymbol{\mu}_i. \quad (1)$$

The accumulated residual vectors corresponding to M visual words are concentrated to form the VLAD image representation of MD dimension, where D is the dimension of local features.

Up till now, some improvement methods have been proposed, which can be roughly divided into three kinds:

- The first is trying to deal with the burstiness phenomenon of visual elements [14]. This phenomenon is alleviated by power-law normalization [15] motivated in [11], which can discount large values in the feature vector.
- The second is dedicated to overcoming the losses in quantization: [16] has considered multiple vocabularies with a joint dimensionality reduction to ameliorate the quantization losses, and [17] proposes vocabulary adaption algorithm to address the problem of vocabulary sensitivity.
- The last is enforcing equal norms for the residual vectors, as proposed in [18], modifying the per-word aggregation step to solve the problem that local descriptors of a given image do not contribute equally to the original VLAD representation.

B. LLC

Locality-constrained linear coding (LLC) [6], inspired by the viewpoint of local coordinate coding (LCC) [19] which illustrates that locality is more essential than sparsity, uses the following criteria:

$$\min_{\boldsymbol{C}} \sum_{t=1}^{N} \|\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{c}_t\|^2 + \lambda \|\boldsymbol{d}_t \odot \boldsymbol{c}_t\|^2 \quad (2)$$

$$s.t. \mathbf{1}^\top \boldsymbol{c}_t = 1, \forall t$$

where $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_N]$ is a set of codes (reconstruction coefficients) for each descriptor. $\odot$ denotes the element-wise multiply operation, and $\boldsymbol{d}_t \in R^M$ is used to measure the distance between $\boldsymbol{x}_t$ and each visual word enforcing locality constraint.

In practice, a simplified version of LLC is used to enhance the coding speed. Ignoring the regularization term in Eq.(2), the approximated LLC directly selects $K (K \ll M)$ nearest basis vectors of each descriptor $\boldsymbol{x}_t$ to reconstruct it by minimizing the first term only.

Let $\sigma_1, \ldots, \sigma_K$ be the indices of the K visual words closer to $\boldsymbol{x}_t$ in the dictionary and represent them collectively as $\tilde{\boldsymbol{B}} = [\boldsymbol{\mu}_{\sigma_1}, \ldots, \boldsymbol{\mu}_{\sigma_K}] \in R^{D \times K}$, then the reconstruction coefficients can be computed by solving the following optimization problem:

$$\min_{\tilde{\boldsymbol{C}}} \sum_{t=1}^{N} \|\boldsymbol{x}_t - \tilde{\boldsymbol{B}}\tilde{\boldsymbol{c}}_t\|^2 \quad (3)$$

$$s.t. \mathbf{1}^\top \tilde{\boldsymbol{c}}_t = 1, \forall t.$$

Approximated LLC is fast in that K is usually a very small number. Meanwhile, both locality representation and sparsity representation can be achieved.

III. Proposed Method

As for the BoF framework, performance gain is achieved by replacing vector quantization with soft-assignment schemes. We speculate that, in VLAD, this sort of replacement will also make sense if the aggregation step considers several nearest local bases instead of only the nearest visual word.

Based on this assumption, we extend VLAD to a vector representation which assigns descriptors to several nearest centroids in feature space. The detailed procedure is illustrated in Algorithm 1.

Algorithm 1 The procedure of VLAD-LLC

Input:
the set of descriptors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ extracted from an image;
the set of centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M$ learned on a trainging set using K-means;

Compute:
for $i = 1, \ldots, M$ **do**
$\boldsymbol{v}_i := \mathbf{0}_D$
for $t = 1, \ldots, N$ **do**
find K centroids nearest to $\boldsymbol{x}_t$, with the indices $\sigma_1, \ldots, \sigma_K$
% compute the weights $W = \{w_1, \ldots, w_K\}$
$W = \text{argmin}_{\sum_j w_j = 1} \|\boldsymbol{x}_t - \sum_{j=1}^{K} w_j \boldsymbol{\mu}_{\sigma_j}\|$
% reconstruction vector
$\boldsymbol{y} := \sum_{j=1}^{K} w_j \boldsymbol{\mu}_{\sigma_j}$
% accumulate descriptor on nearest K centroids
for $j = 1, \ldots, K$ **do**
$\boldsymbol{v}_{\sigma_j} := \boldsymbol{v}_{\sigma_j} + w_j \times (\boldsymbol{x}_t - \boldsymbol{y})$
% accumulated residual vectors are concentrated
$\boldsymbol{V} = [\boldsymbol{v}_1^\top \ldots \boldsymbol{v}_M^\top]$
% apply power normalization
for $u = 1, \ldots, MD$ **do**
$V_u := sign(V_u)|V_u|^\alpha$
% apply L2 normalization
$\boldsymbol{V} := \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_2}$

Output:
representation $\boldsymbol{V}$.

Our method can be summarized as follows: First, each local descriptor is associated with $K (K > 1)$ nearest visual words in the dictionary; then the weights that best linearly reconstruct the descriptor from its neighbors can be computed; finally, multiplied by the corresponding weight, the difference between the descriptor and the reconstruction (reconstructed by K local bases) is accumulated for each visual word.

Benefiting from the combination of VLAD and LLC, VLAD-LLC produces an image representation of the same dimension MD as VLAD at a low extra computational cost and thus brings no increase in the complexity of the final classification step.

Compared to VLAD, another improvement of VLAD-LLC is the strategy of the aggregation step. We accumulate the weighted difference between the descriptor and the reconstruction vector rather than the associated visual word directly, because the latter could achieve little performance gain. The reconstruction step can be seen as automatical generation of a new centroid using local bases to deal with codeword ambiguity, and leads to smaller coefficients for the basis vectors farther away from a local feature at the same time.

VLAD is a particular case of our proposed method where the number of nearest neighbors is equal to 1. We show experimentally that the additional reconstruction step can ameliorate quantization losses and bring stable improvement in terms of classification accuracy.

IV. Experimental Results

In this section, we begin with an illustration of our experiments setting, then evaluate the classification performance of VLAD-LLC compared with VLAD based on three widely used datasets: Caltech256 [20], Caltech101 [21], and Sports8 [22]. Throughout the experiments, our implementation uses only SIFT descriptor and linear SVM classifier.

A. Experiments Setting

In our implementation, we first resize the maximum height of each image to be no more than 480 pixels with preserved aspect ratio, then employ the 128 dimensional SIFT descriptor densely extracted on a grid with step size of 4 pixels under three scales: 16×16, 24×24, 32×32. Next, totally a million of SIFT descriptors randomly drawn from the training images are taken as input for K-means clustering to learn a dictionary with 64 visual words. During the coding processing, the approximated LLC is actually used, and the number of nearest codes (K) is set to 4 except on Caltech-256 dataset, where the number of neighbors varies from 2 to 5. For power normalization applied in VLAD, the parameter α is set to 0.2 as suggested in [18].

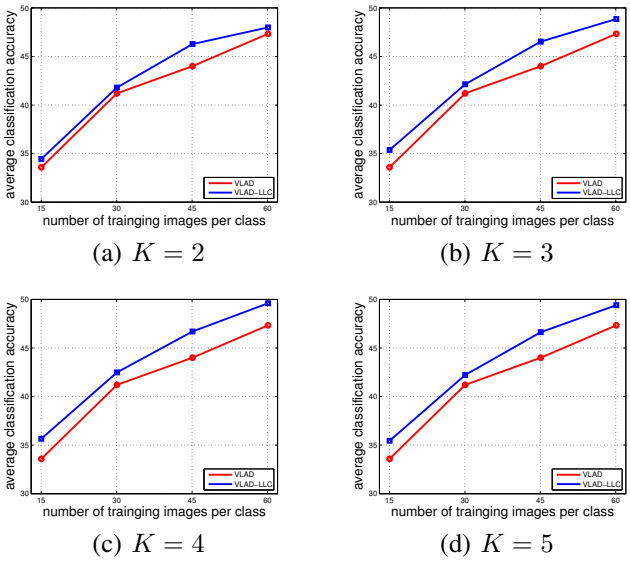
To incorporate spatial information, an image is partitioned into totally eight sub-regions in three levels of spatial pyramid as 1×1, 2×2, 3×1, and the feature vectors are computed on the respective sub-regions and then concatenated into a long feature vector for the whole image that is finally fed into the linear SVM classifier. Eventually, the performance is measured by averaging classification accuracies over all categories.

All the experiments are repeated five times with different random selected training and test samples to obtain reliable results. The average classification accuracy and the standard deviation is reported.

B. Experimental Datasets

Three datasets are chosen for evaluation, which are, respectively:

- The Caltech-256 dataset [20] holds 30607 images falling into 256 categories, each category contains at least 80 images. According to the standard experimental setup, we randomly pick up 15, 30, 45, and 60 training images per category, and at most 50 images for test.
- The Caltech-101 dataset [21] contains 9144 images belonging to 101 categories, where the number of images in each category varies from 31 to 800. Following the common setup during experiment, we use 15 and 30 images per category for training while leaving the rest for test.
- The UIUC sports event dataset [22] is collected for image-based event classification. It contains 8 categories and 1579 images in total, and the size of each category ranges from 137 to 250. These 8 categories are rowing, badminton, polo, bocce, snowboarding, croquet, sailing, and rock climbing. Following the experimental setup used in [22], we randomly select 70 training images and 60 test images from each class.



(a) $K = 2$ (b) $K = 3$ (c) $K = 4$ (d) $K = 5$

Fig. 1. VLAD-LLC compared with VLAD under different K neighbors (Caltech-256)

C. Analysis and Discussion

We explore the effect of different small number of neighbors used for approximated LLC in our proposed method. The detailed performance comparison is shown in Fig.1.

As can be seen, the results are very impressive: under all the cases (four types of training number and four values selected for K), our method can consistently outperform VLAD in terms of classification accuracy. Since the best result on Caltech-256 dataset is obtained when K is set to 4, it seems a fairly good choice for K considering 64 clustering centers. Thus, the experiments on other datasets simply use this setting.

We also compare our method with several existing coding schemes [4], [5], [6], [23] reported in the literature. The comparison results are shown in Table I. Although the implementation details may be quite different, it can be concluded that the baseline method VLAD incorporating spatial information used in this paper is favorably competitive with other feature

TABLE I
PERFORMANCE COMPARISON ON CLASSIFICATION DATASETS.

(a)Caltech-101

Methods	**Acc.(%)(Tr.=15)**	**Acc.(%)(Tr.=30)**
ScSPM [5]	67.0 ± 0.45	73.2 ± 0.54
LLC [6]	65.43	73.44
VLAD	67.27 ± 0.51	75.21 ± 0.62
VLAD-LLC	69.01 ± 0.42	77.35 ± 0.57

(b)UCUI-Sports

Methods	**Acc.(%)**
localized soft-assignment coding [4]	84.56 ± 1.5
laplacian sparse coding [23]	85.31 ± 0.51
VLAD	88.79 ± 0.99
VLAD-LLC	89.50 ± 0.83

coding algorithms on classification tasks, and replacing VLAD with VLAD-LLC can bring considerable performance gain on these two datasets.

V. CONCLUSION

In this paper, we first use VLAD incorporating spatial pyramid as an image representation for classification tasks and evaluate it on several benchmark datasets. Besides, we propose an approach to combine VLAD with approximated locality-constrained linear coding which uses local basis to linearly reconstruct each descriptor.

The experiments on different benchmark datasets verify the effectiveness of VLAD-LLC under small dictionary size and demonstrate that our proposed method outperforms original VLAD at a low extra computational cost in the field of image classification.

As an indication of our work, VLAD-LLC might serve as a better image representation not exclusively for image classification. Further researches can be made by integrating VLAD-LLC into other image processing tasks.

VI. ACKNOWLEDGEMENT

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

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[3] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 696–709.

[4] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2486–2493.

[5] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[7] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1753–1760.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[9] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 141–154.

[10] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 143–156.

[12] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[13] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3304–3311.

[14] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1169–1176.

[15] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1704–1716, 2012.

[16] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 774–787.

[17] R. Arandjelovic and A. Zisserman, "All about vlad," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1578–1585.

[18] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the vlad image representation," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 653–656.

[19] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223–2231.

[20] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[21] 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," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[22] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[23] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3555–3561.