Improving VLAD with Regional PCA Whitening

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*—In recent yeas, VLAD has been used to represent an image effectively and efficiently by just a few bytes in large-scale image retrieval. In spite of its remarkable performance, a series of modification methods have been presented. In addition, the redundancy between the features corresponding to the same cluster center could be improved. In this paper, a regional PCA Whitening method is proposed to decorrelate the features and reduce the dimensionality for each cluster with the consideration of mapping the descriptor into high dimensionality explicitly. Our method can also be embedded into original VLAD pipeline with global PCA very well. The experimental results on both Holidays and UKbench dataset show that our approach improves VLAD significantly.**

***Index Terms*—VLAD, image retrieval, PCA, feature representation, normalization**

I. Introduction

For the task of image search, the bag-of-features (BoF) approach [1] has become very popular during the last decade. Many state-of-the-art retrieval systems rely on the BoF representation. For BoF, an image is described by local features such as SIFT [2] where the corresponding descriptors are quantized into discrete visual words. The visual words are generally cluster centres of local descriptors generated by K-means algorithm. The image is then represented by a histogram of weighted visual word occurrences. This representation is simple and robust to occlusion, clutter and other image transformations.

Recently, several attempts have been made to enhance the performance of BoF, in which Vector of Locally Aggregated Descriptors (VLAD) [3], [4] is among the most successful methods. By using this approach, an image can be represented by just a few dozen bytes after the dimensionality reduction and compression. Meanwhile, it is also advantageous with less computation resource compared with other methods such as Fisher Vector (FV). Both the retrieval accuracy and efficiency can still preserve high while searching even in hundred millions images.

Some modification methods of VLAD are proposed, which can be roughly divided into three kinds. The first is to solve the problem of bursty visual elements [5]. Power-law normalization [6] can be used to address this problem, which can discount large values in the vector. Intra-normalization [7] is also an effective solution which normalizes the sum of residuals within a cluster independently. The second is dedicated to overcome the problem of quantization losses, which is more serious when the data distribution between the training dataset and the retrieval dataset is very inconsistent. The vocabulary is usually trained off-line and on a sampled dataset. Jgou and Chum [8] have used multiple vocabularies and the trick of joint dimensionality reduction to alleviate this problem. Arandjelovic and Zisserman [7] propose a vocabulary adaptation algorithm to reduce the problem of vocabulary sensitivity. The last is incorporating other useful information into VLAD vectors. The dominant orientation was introduced by Zhao and Jgou [9], which used oriented pooling method to maintain invariance to orientation. Besides, using dense features also gets better results both in [9] and [10]. Based on explicit feaure mapping [11], nonlinear feature mapping with consideration of the vector difference between descriptors and vocabulary in a high-dimensional space is proved to be effective [12].

For retrieving large scale datasets, the dimensionality of VLAD vectors is always a problem, especially when considering spatial information. In order to address this problem, PCA is generally used to reduce the VLAD vector into a low dimensionality. And it is observed that the performance of VLAD is even improved by PCA [8]. However, the PCA method is generally used globally. In this paper, we propose a method called regional PCA Whitening (RPCAW) to further improve the performance of VLAD. In our proposed approach, we firstly map the obtained features into high dimensional space explicitly. Then we compute the projection matrix based on each cluster center, which is generated by K-means. Finally we project each component of VLAD vector into a new space. In addition, the global PCA on VLAD can also be performed directly. The experiments show that our proposed method is of better performance than both regular VLAD and VLAD with global PCA.

The remainder of the paper is organized as follows. In Section II, the regular VLAD and explicit feature mapping are introduced. Additionally the proposed method is presented. The experimental results will be shown in Section III. At last, we will conclude our work in Section IV.

II. RPCAW: AN IMPROVEMENT OF VLAD

In this section, we firstly introduce the original VLAD pipeline, then present the explicit feature mapping, and finally illustrate the proposed regional PCA Whitening.

A. Original VLAD

The VLAD describes an image by the difference of its local feature descriptors from a learned codebook. For that, a set of low-level features are extracted from each image. VLAD utilizes a coarse visual codebook $q : X \rightarrow C$, $C = \{c_1, c_2, ..., c_k\}$, that has been learned offline and maps image descriptors to a set of centroids of size k. Here, X denotes the descriptor space and C the set of centroids. Typically, such a visual codebook is obtained by k-means clustering of descriptors of a training dataset.

Given an image represented by a set of m local descriptors $I = \{x_1, x_2, ..., x_m\}$, the original VLAD representation is obtained by encoding the descriptors in the following way:

$$v_i = \sum_{x_j \epsilon I : q(x) = c_i} x_j - c_i \tag{1}$$

That is, each local descriptor is assigned to its nearest centroid and the residual with this centroid is computed. The residuals of all descriptors with centroid c_i are accumulated. Each centroid c_i in the codebook contributes a vector of aggregated residuals. The final VLAD signature v is obtained by concatenating the residual vectors v_i forming a $D = kd$ dimensional image signature where d is the dimensionality of the original descriptors.

Normalization: When a VLAD vector is obtained, normalization is generally necessary. Though there are several normalization methods [13], the two steps including power normalization and L_2 normalization are mainly considered in this paper.

For power normalization, we apply the function f for each component v_i of VLAD vector:

$$f(v_i) = sgn(v_i)|v_i|^{\alpha} \tag{2}$$

where α is typically set to 0.5. Compared with power normalization, another normalization is intra-normalization [7]. After power normalization, we do the L_2 normalization to the obtained vector v'.

$$y = \frac{v'}{\parallel v' \parallel} \tag{3}$$

where y is the final VLAD vector that can be used to represent an image.

B. Explicit feature mapping

When considering similarity metric, it is proved that additive kernels [11] are effective to map the feature to a high-dimensional space implicitly. For VLAD vectors x and y, the similarity metric is denoted by $K(x, y)$ and additive kernels can be described as

$$K(x, y) = \sum_{1}^{D} \kappa(x_k, y_k) \tag{4}$$

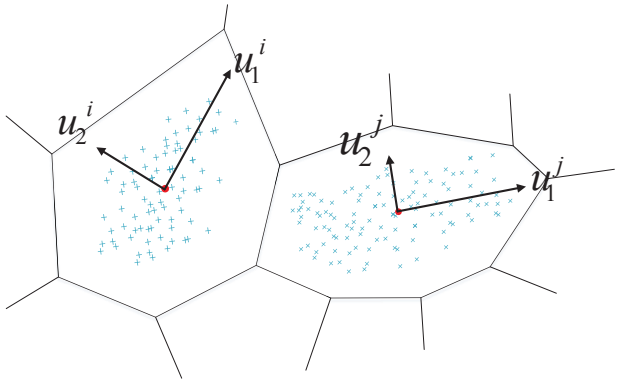

Fig. 1. **The illustration of regional PCA Whitening**. The Voronoi cells represent different clusters generated by K-means algorithm and the sign "+" is the descriptor assigned to a word. u_1^i and u_2^i indicate the principal direction for i-th cluster. The same is for u_1^j and u_2^j. It is observed that the direction of variance for different cluster subspace is also different obviously.

where the function $\kappa(x_k, y_k)$ is used to compute the similarity between the k-th elements of x and y. For χ^2 and intersection kernel, they have the form as $\kappa(x_k, y_k) = \frac{(x_k - y_k)^2}{x_k - y_k}$ and $\kappa(x_k, y_k) = min(x_k, y_k)$ respectively. Thus, the χ^2 and intersection kernels are example of additive kernel.

Generally, kernel function is calculated through an implicit feature map $\psi(x)$ of data points from a low-dimensional space to a high-dimensional space. But an approximate feature map can be used to map the original feature explicitly

$$\kappa(x_k, y_k) \approx \langle \psi_\kappa(x_k), \psi_\kappa(y_k) \rangle \tag{5}$$

where $\psi_\kappa(x)$ is the explicit mapping of feature x. In this paper, we map the feature explicitly by using χ^2 kernel, namely

$$\psi_\kappa(x) = e^{i\tau \log x} \sqrt{x sech(\pi\tau)} \tag{6}$$

where τ is considered as the index of the implicitly mapped feature vector $\psi_\kappa(x)$ and function $\psi_\kappa(x)$ map the original D dimensional feature into M dimensional feature. In details, $\psi_\kappa(x)$ is approximated by a finite number of samples by Fourier sampling theorem. Each element of feature vector in the original space becomes $2n + 1$ elements in the mapped space and the number of samples n is set to 1. Therefore, the final representation of the high dimensional VLAD vector is a feature vector with $M = (2n + 1)D$ dimension after explicit mapping.

C. regional PCA Whitening

The regional PCA Whitening is performed on each cluster subspace. For i-th cluster, we firstly compute the covariance matrix G_i as follows

$$G_i = \frac{1}{D} \sum_{j=1,k=1} (x_j - c_i)(x_k - c_i)^T \tag{7}$$

where x_j and x_k are the descriptors assigned to the cluster center c_i. In fact, c_i is also the mean of descriptors assigned

to the i-th cluster. Then, we compute the eigenvalues and eigenvectors of G_i:

$$(\lambda_1^i, \lambda_2^i, ..., \lambda_D^i) = eigval(G_i) \quad (8)$$

$$(u_1^i, u_2^i, ..., u_D^i) = eigvect(G_i) \quad (9)$$

with $\lambda_1^i \geq \lambda_2^i \geq ... \geq \lambda_D^i$ and u_t associated with λ_t^i. In fact, u_t^i describes the principal direction of features distribution in the cluster subspace. As shown in Fig. 1, the features distribution is different from each other for different cluster subspace.

In order to compute the projection matrix, we firstly denote the matrix by L_t^i :

$$L_t^i = diag(\frac{1}{\lambda_1^i + \epsilon}, \frac{1}{\lambda_2^i + \epsilon}, ..., \frac{1}{\lambda_t^i + \epsilon}) \quad (10)$$

where ϵ is a regularization value and is set to 0.00001 in practice. The matrix of t largest eigenvectors is then denoted by U_t^i:

$$U_t^i = [u_1^i, u_2^i, ..., u_t^i] \quad (11)$$

At last, we obtain the projection matrix P_t^i:

$$P_t^i = L_t^i U_t^i \quad (12)$$

For each feature in i-th cluster, we project it into the low dimensional space by using matrix P_t. Then the final obtained VLAD vector is

$$y = [P_t^1 x^1, P_t^2 x^2, ..., P_t^k x^k] \quad (13)$$

In fact, after performing regional PCA Whitening, the dimensionality of VLAD vector is tk. Because t is less than $(2n+1)d$, the dimension of VLAD is reduced significantly.

In fact, our proposed regional PCA Whitening is very different from the global PCA method. Compared to distribution feature of global VLAD space, The RPCAW method is mainly focused on the feature distribution of each cluster. It is very significant and crucial. In addition, we preserve the regional dominant information and decorrelate the components in each cluster space. In the following experiments, it will be shown that RPCAW improves VLAD significantly.

III. Experimental results

The performance of proposed method is evaluated by several experiments. The experimental benchmark datasets include Holidays [14] and Ukbench [15].

TABLE I
DETAILS OF THE DATASETS IN THE EXPERIMENTS

Dataset	# images	# queries	# descriptors	Evaluation
Holidays	1491	500	4,455,091	mAP
UKbench	10200	10200	19,415,079	N-S score

A. Datasets and evaluation

INRIA Holidays [14] contains 1491 high resolution personal holiday photos (Fig. 2 and Table I) of a variety of scene types. The retrieval accuracy is measured by mean average precision (mAP) after removing the query image from the ranked list.

UKbench [15] includes 2,550 different objects (Fig. 2 and Table I), and each one has 4 images taken from different viewpoints and illuminations. All 10,200 images are indexed as both database images and queries. The retrieval performance is measured by 4 $\times$ recall at the first 4 retrieved images, which is referred as the N-S score (maximum is 4).

The local descriptors of all methods are extracted with Hessian-affine detector [16] and described by SIFT descriptor [2]. Following [17], rootSIFT is used on every point since it is shown to be effective in image retrieval. In our experiments, all the vocabularies are trained independently on the Flickr60K dataset [14] using AKM [18] with different initial seeds.

B. standard VLAD with RPCAW

In order to evaluate the performance of VLAD with RPCAW, we compare standard VLAD and VLAD with RPCAW under the same circumstances. In addition, we also present the retrieval result for VALD with just explicit feature mapping. As presented in TABLE II and TABLE III, the RPCAW method with different parameter t, which varies from 32 to 256, is used for regular VLAD. It can be observed that the performance of VLAD with explicit feature mapping (HVLAD) is better than standard VLAD (SVLAD) and VLAD with RPCAW outperforms both SVLAD and HVLAD, especially when $t = 128$.

TABLE II
THE PERFORMANCE OF RPCAW ON HOLIDAYS(MAP)

	SVLAD	HVLAD	VLAD+RPCAW			
Size	-	-	32	64	128	256
64	**58.09**	60.50	59.71	62.51	**64.28**	59.74
128	**61.12**	62.72	61.13	64.3	**65.24**	62.59
256	**61.68**	63.08	63.34	65.58	**66.93**	63.88

TABLE III
THE PERFORMANCE OF RPCAW ON UKB(N-S SCORE)

	SVLAD	HVLAD	VLAD+RPCAW			
Size	-	-	32	64	128	256
64	**3.34**	3.45	3.41	3.46	**3.50**	3.38
128	**3.38**	3.49	3.46	3.51	**3.52**	3.43
256	**3.43**	3.52	3.5	3.54	**3.54**	3.49

C. VLAD with RPCAW and GPCA

When taking the scale of dataset into consideration, the global PCA method is used generally. In order to further prove the performance of RPCAW method, the standard VLAD with global PCA and VLAD with both RPCAW and global PCA are compared. Fig. 3 and Fig. 4 show the performance of VLAD with embedding of both RPCAW and global PCA is much

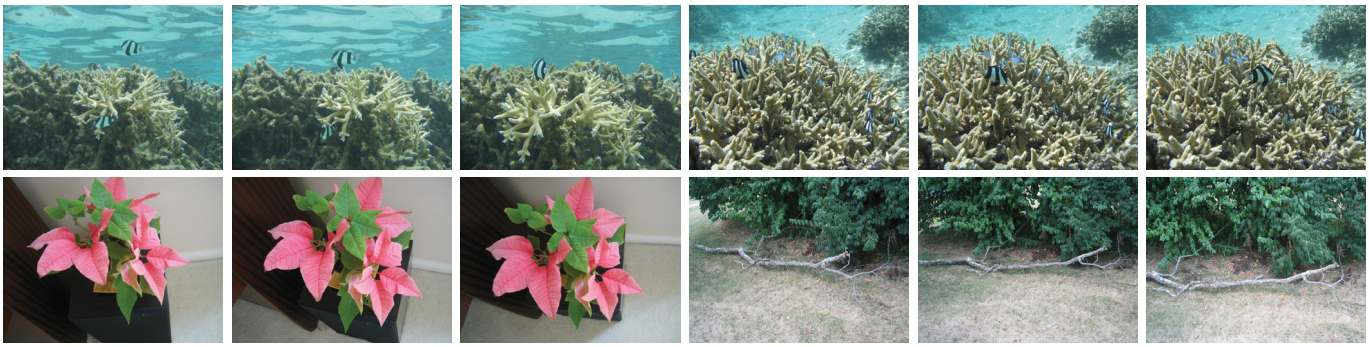

Fig. 2. Images from Holidays and UKbench. **Top**: for Holidays dataset; **bottom**: for UKbench dataset.

better than original VLAD with global PCA under different vocabulary size.

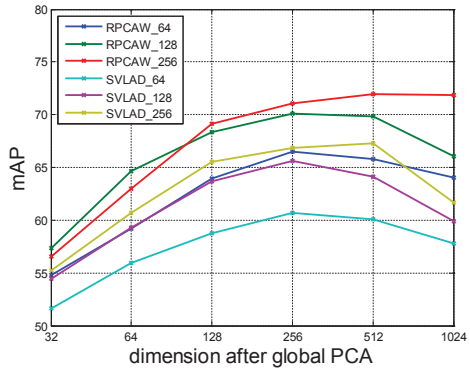

Fig. 3. Global PCA performed on VLAD with RPCAW and SVLAD for Holiday dataset under different vocabulary size (64, 128, 256).

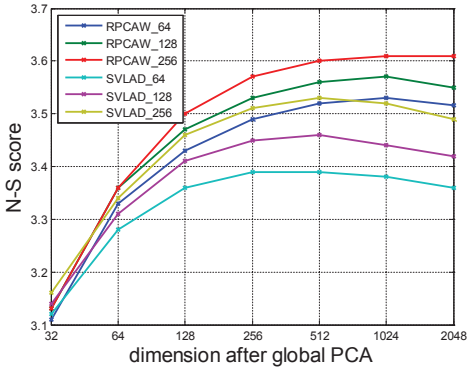

Fig. 4. Global PCA performed on VLAD with RPCAW and SVLAD for UKbench dataset under different vocabulary size (64, 128, 256).

IV. Conclusion

In this paper, a regional PCA Whitening method is proposed to further improve the VLAD. Compared with global PCA method, the RPCAW method can not only reduce the dimensionality of VLAD vector, but also decorrelate the features assigned to the same cluster which is proved to be effective in our experiments. In addition, we use the explicit feature mapping method to map VLAD into a high dimensional space. As presented in the experiments on both Holidays and UKbench datasets, our proposed method significantly enhances the VLAD.

V. Acknowledgement

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

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.
[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.
[3] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.
[4] 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.
[5] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *CVPR*, 2009.
[6] F. Perronnin, J. Snchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.
[7] R. Arandjelovic and A. Zisserman, "All about vlad," in *CVPR*, 2013.
[8] H. Jgou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *ECCV*, 2012.
[9] W.-L. Zhao, H. Jégou, and G. Gravier, "Oriented pooling for dense and non-dense rotation-invariant features," in *BMVC - 24th British Machine Vision Conference*, 2013.
[10] 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*.
[11] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.
[12] X. Zhao, Y. Yu, Y. Huang, K. Huang, and T. Tan, "Feature coding via vector difference for image classification," in *ICIP*, 2012.
[13] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "A comprehensive study over vlad and product quantization in large-scale image retrieval," *Multimedia*, 2014.
[14] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.
[15] D. Nistr and H. Stewnius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.
[16] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, 2004.
[17] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012.
[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching." in *CVPR*, 2007.