Fusion of Deep Features and Weighted VLAD Vectors based on Multiple Features for Image Retrieval

Yanhong Wang[1,2], *Yigang* Cen[1,2] , *Liequan* Liang[3,*], *Linna* Zhang[4], *Viacheslav* Voronin[5], *Vladimir* Mladenovic[6]

[1]School of Computer and Information Technology, Beijing Jiaotong University, 100081, Beijing, China
[2]Key Laboratory of Advanced Information Science and Network Technology of Beijing, 100081, Beijing, China
[3]Institute of Electronic Commerce, Guangdong University of Finance & Economics, 510320, Guangzhou, China
[4]College of Mechanical Engineering, Guizhou University, 550025, Guiyang, China
[5]Department of Radio-electronic systems, Don State Technical University, 346500, Rostov-on-Don, Russia
[6]Faculty of Technical Sciences University of Kragujevac, 32000, Cacak, Serbia

Abstract. In traditional vector of locally aggregated descriptors (VLAD) method, the final VLAD vector is reshaped by summing up the residuals between each descriptor and its corresponding visual word. The norm of the residuals varies significantly, and it can make "visual burst". This is caused by a fact that the contribution of each descriptor to VLAD vector is not the same. To address this problem, we add a different weight to each residual such that the contribution of each descriptor to the VLAD vector becomes even to a certain degree. Also, traditional VLAD method only uses the local gradient features of images. Thus it has a low discrimination. In this paper, local color features are extracted and used to the VLAD method. Moreover, we fuse deep features and the multiple VLAD vectors based on local gradient and color information. Also, in order to reduce running time and improve retrieval accuracy, PCA and whitening operations are used for VLAD vectors. Our proposed method is evaluated on three benchmark datasets, i.e., Holidays, Ukbench and Oxford5k. Experimental results show that our proposed method achieves good performance.

1 Introduction

In this paper we consider the task of large-scale image retrieval. In the past few years, Bag-of-Visual-Words (BOW) [1] [2] method has achieved great effect in image retrieval area. Generally, in order to ensure retrieval recall, a relatively large vocabulary will be required. Thus, it will lead to a low efficiency of retrieval time and high memory consumption.

Recently, Jégou et al [3] proposed vector of locally aggregated descriptors (VLAD) model, which aggregates descriptors based on a locality criterion in feature space. In fact, VLAD is a kind of representation of Fisher vector without probability. Its implementation is very similar to the BOW model. Also, VLAD is very cheap in consumptions of time and memory. In traditional VLAD method, the final VLAD vector is reshaped by summing up the residuals between each descriptor and its corresponding visual word. The norm of the residuals varies significantly, thus it can make "visual burst" [4]. To address this problem, we add a weight to each residual such that the contribution of each descriptor to the VLAD vector becomes even to a certain degree.

Originally, the SIFT [5] descriptors are adopted in VLAD method, and has shown good performance. As we all known, the SURF descriptor [6] is faster than the SIFT descriptor. Moreover, the performance of SURF and SIFT is comparative in most cases. [7] verified that the SURF descriptor was not only more efficient but also leading to higher accuracy than SIFT and rootSIFT descriptors. However, both the SIFT and SURF descriptors represent only local gradient information, which miss important color information. In order to solve this problem, many works combine the gradient and color information. For examples, in [7] CSURF feature was proposed, which are SURF-based color information; In [8], the author fused the CLOG [9] features and the SURF features at the stage of similarity measurement; In [10] the author proposed "color-SURF" descriptors which combined SURF with the approximate color local kernel histograms. In this paper, Color names (CN) [11] and SURF features are used in VLAD method.

In recent years, deep features are popular for image processing, such as image classification [12], object detection [13] and speech recognition [14] etc. In this paper, we adopt the pre-trained networks to obtain the deep features of images. Also, in order to improve retrieval accuracy, multiple VLAD vectors and image representation based on deep features are fused.

* Corresponding author: lianglq@gdufe.edu.cn

2 Methodology

2.1 Framework of our proposed method

The framework of our proposed method is shown in Fig.1. Vocabulary based on local features (SURF or CN) are trained from an independent training dataset. For an image, SURF and CN features are extracted and quantized on corresponding vocabulary respectively. Here, we improve the traditional VLAD method by adding a weight for each residual, called as "weighted VLAD". The CN and SURF features are adopted to weighted VLAD method, respectively. Then, weighted VLAD vectors based on the two features are obtained. Moreover, the deep features are extracted from the image, and image representation based on deep features is computed. Then the vectors are fused into a vector to represent the image. Finally, similarity scores of the query image and dataset images are measured, and the retrieval results are returned.

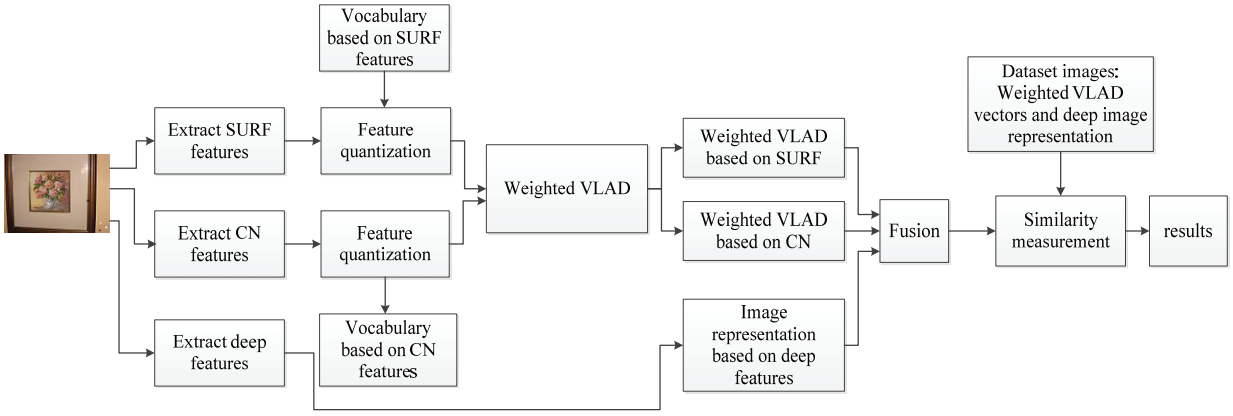


Fig. 1. The framework of our proposed retrieval method

2.2 Weighted VLAD

For an image, the generation process of VLAD vector is as follows: (1) Detect interest regions of the image and extract local descriptors, denoted as $x=(x_1,x_2,\ldots,x_n)\in\mathbb{R}^{d\times n}$. (2) The local descriptors are quantized on the vocabulary $C=(c_1,c_2,\ldots,c_K)$ with K visual words $(L=K)$, denoted as $NN(x_i)=\arg\min_k\|x_i-c_k\|$. (3) The residuals of a visual word and the descriptors that are quantized to this word are computed, formulated as $V_k=\sum_{i:NN(i)=k}(x_i-c_k)$. Then, the residuals are summed. A vector of length $L=K*d$ is obtained, which is called as the VLAD vector. (4) Power-law normalization is adopted for the vector obtained at the step (3). It contains two steps: firstly, there is the square root with symbol, formula as $V_i=|V_i|^{\alpha}\times sign(V_i),0\le\alpha<1$; secondly, the vector is normalized by the L_2-norm, denoted as $V=V/\|V\|_2$.

According to above step (3), it may cause "visual burst" phenomenon because contribution of each descriptor for the VLAD vector is not same, i.e., the closer from center of the cluster, the greater the contribution is, and vice versa. In the similarity measurement stage, this residual will be reflected in the contribution since the Euclidean distance is used. To address the problem, we add a different weight for each residual. Here, the weight is set to be the normalized distance of the descriptor and its nearest visual word, denoted as Eq. (1) and Eq. (2), i.e., the smaller (greater) is the distance between the descriptor and its nearest visual word, the smaller (greater) is the weight. The validity is verified in the experimental section (Section 3). It should be noted that a same weight is added to the residuals corresponding to each visual word in [4]. But in our algorithm, different weights are added to the residuals respectively. This makes our algorithm become more flexible and adaptive for image retrieval.

$$V_k=\sum_{i:NN(i)=k}\omega_i(x_i-c_k) \tag{1}$$

$$\omega_i=\frac{d(x_i,c_k)}{\sum_{i:NN(i)=k}\|x_i-c_k\|} \tag{2}$$

where ω_i is a weight coefficient. Also, $d(x_i,c_k)$ represents the distance of the descriptor x_i and the visual word c_k.-

For a given image, SURF descriptors are extracted, denoted as $x^{SURF}=(x_1^{SURF},x_2^{SURF},\ldots,x_{n_1}^{SURF})$. For the SURF-based vocabulary $C^{SURF}=(c_1^{SRUF},c_2^{SURF},\cdots,c_{K_1}^{SURF})$, according to Section 2.2, a SURF-based VLAD vector can be obtained, denoted as V^{SURF}. Its length is $L^{SURF}=K_1*d_1$. In addition, each CN descriptor is extracted from an image patch of size $p\times p$, denoted as x_i^{CN}. Specifically, for each pixel of a patch, an 11-D CN descriptor is extracted. Then, the average of all descriptors in the patch is regarded as the color descriptor of the patch. Thus, a set of CN descriptors for the image are obtained, denoted as $x^{CN}=(x_1^{CN},x_2^{CN},\cdots,x_{n_2}^{CN})$. For the CN-based

vocabulary $C^{CN}=\left(c_1^{CN},c_2^{CN},\cdots,c_{K_2}^{CN}\right)$, similarly, the CN-based VLAD vector is computed, denoted as V^{CN}. Also, the length is $L^{CN}=K_2*d_2$. In order to improve retrieval accuracy, deep features are extracted by using pre-trained deep convolutional neural networks (CNN) models, and the length is L_3.

2.3 Feature fusion and similarity measurement

For an image, the three image representations are fused to a vector, denoted as:

$$V=[\lambda_1*V^{SURF},\lambda_2*V^{CN},\lambda_3*V^{CNN}] \quad (3)$$

where λ_1, λ_2, λ_3 are the weight parameters, and $\lambda_1+\lambda_2+\lambda_3=1$. The best values fitting λ_1, λ_2, λ_3 are selected by loop iteration.

Euclidean distance is used to compute similarities between the query image and the dataset images. To reduce running time, we adopt PCA and whitening method which suppressed the co-occurrence problem with the dimensionality reduction [15].

Our proposed algorithm is summarized as follows:

Algorithm I	Fusion of deep features and weighted VLAD vectors based on multiple features

1) Off-line

- Train vocabularies C^{SURF},C^{CN} on the training dataset.
- Extract the dense SURF and CN descriptors from each image of the dataset.
- Compute weighted VLAD vectors V_I^{SURF},V_I^{CN} and extract deep feature V_I^{CNN} for each image I in the dataset.
- Fuse V_I^{SURF}, V_I^{CN} and V_I^{CNN} by Eq. (3), denoted as V_I.
- Reduce dimensionality of V_I, and the result is denoted as V_{IR}.

2) On-line

- Extract the dense SURF and CN descriptors from a query Q.
- Compute weighted VLAD vectors V_Q^{SURF},V_Q^{CN} and extract deep feature V_Q^{CNN} for the query image.
- Fuse V_Q^{SURF}, V_Q^{CN} and V_Q^{CNN} by Eq. (3), denoted as V_Q.
- Reduce dimensionality of V_Q, and the result is denoted as V_{QR}.
- Compute similarity score.
- Return images of the dataset with the high similarity scores.

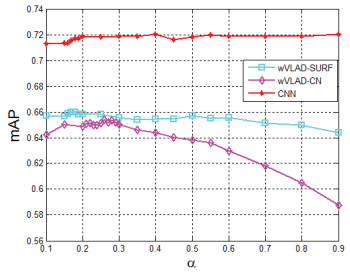
3 Experimental results

In this section we verified our proposed method on three benchmark datasets, i.e., Holidays [16], Ukbench [17] and oxford5k [18]. In addition, Paris60k [19] is used to train vocabularies for Oxford5k. Vocabularies are trained from Mirflickr25k [20] for other datasets. All experiments are implemented on a computer with 8GB memory and 3.3GHz CPU (Intel(R) Core(TM) i5-4590).

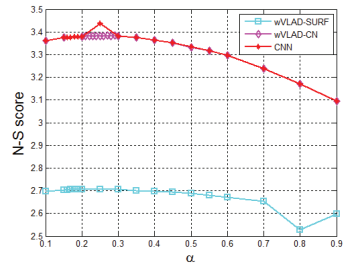
3.1 Selection of parameters

In our experiments, the dense SURF descriptors are extracted for each image. Moreover, each CN descriptor is obtained in an image patch of size 4×4. Also, the CN-based and SURF-based vocabularies of size 64 are used. Moreover, CNN features of images are obtained by the VGG-f model [21]. Here, a CNN-based representation is obtained from the second fully-connected layer of convolutional networks for each image, it is a 4096-D vector.

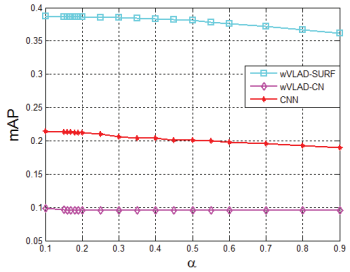
α is a power of the absolute value of VLAD vectors matrix. However, we find that the best value of α is between 0.1 and 0.6. The accuracies with different α by using weighted VLAD based on different features on three datasets are shown in Fig. 2.



(a)Holidays



(b) UKbench



(C) Oxford5k

Fig. 2. The comparison with different values of α by using CNN-based vectors and VLAD vectors based on SURF features and CN features. In (a) and (c), the mAPs are

shown on the Holidays and Oxford5k dataset. In (b), the N-S scores on the Ukbench dataset are shown.

Here, weighted VLAD vectors based on SURF and CN features are denoted as wVLAD-SURF, wVLAD-CN, respectively. On Holidays, we select α=0.19 for wVLAD-SURF, α=0.25 for wVLAD-CN, and α=0.2 for CNN-based vectors. On UKbench datasets, α is set to be 0.19, 0.25, 0.25 for wVLAD-SURF, wVLAD-CN and CNN-based vectors, respectively. Moreover, α is set to be 0.1, 0.15, 0.1 on Oxford5k , respectively.

In Eq. (3), λ_1 , λ_2 , λ_3 are the weight parameters of SURF-based and CN-based VLAD vectors and CNN-based vectors for feature fusion. On Holidays they are respectively set to be 0.3, 0.3, 0.4. Also, they are set to be 0.2, 0.35, 0.45 and 0.6, 0.05, 0.35 on the UKbench and Oxford5k.

3.2 Effectiveness of weighted VLAD

In this subsection, we verified the effectiveness of proposed weighted VLAD model. In Fig.3, two examples on UKbench dataset are shown. It can be seen that the results of wVLAD-SURF are better than traditional VLAD-SURF. In addition, we compare our weighted VLAD with [4] (VLAD-LCR-RN) in Table 1. On Holidays, it can be seen that the results are the same, but the length of our vector is only about a half of VLAD-LCR-RN. On Oxford5k when vectors are reduced to 128D, wVLAD-SURF achieves a better result.

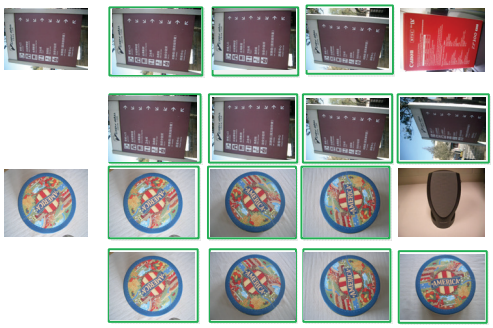


Fig. 3. Two examples by using the VLAD vectors based on SURF features on the UKbench dataset. The left images are queries, and the first row represents the results obtained by the traditional VLAD based on SURF descriptors, where those with green boxes are the correct results. Also, the second row is the results obtained by using our proposed weighted VLAD based on the SURF.

Table 1. Comparison between VLAD-LCR-RN and weighted VLAD-SURF.

Methods	VLAD-LCR-RN[4]	wVLAD-SURF
Holidays (mAP/d)	0.658/8192	0.6581/4096
Oxford5k (mAP/d)	0.517/8192	0.3877/4096
Oxford5k (mAP/d)	0.322/128	0.4050/128

3.3 Fusion of multiple features

In experiments, wVLAD-SURF and wVLAD-CN are fused, denoted as wVLAD-SURF+wVLAD-CN. The multiple weighted VLAD vectors and deep features are fused into a vector, denoted as wVLAD-SURF+wVLAD-CN+CNN. In Table 2, the retrieval accuracies on different datasets are listed, where $L = 128$ denotes the length of VLAD vectors obtained by PCA and whitening operations. When $L = 128$, on Holidays, it can be seen that the mAP obtained by wVLAD-SURF + wVLAD-CN increased by nearly 13% compared to wVLAD-SURF and wVLAD-CN respectively. Also, the mAP achieves 0.8306 for wVLAD-SURF+wVLAD-CN+CNN. Also, On UKbench and Oxford5k, the N-S score and the mAPs reach respectively 3.6916 and 0.4322 for wVLAD-SURF+wVLAD-CN+CNN. Thus, feature fusion make retrieval accuracies obviously improved.

Table 2. Comparison of results obtained by using different features and by reduction on Holidays, Oxford5k and UKbench datasets.

Methods	L	Holidays (mAP)	Ukbench (N-S score)	Oxford 5k (mAP)
wVLAD-SURF	4096	0.6581	2.7074	0.3877
	128	0.6902	2.6268	0.4050
wVLAD-CN	704	0.6515	3.3825	0.0962
	128	0.6698	3.3766	0.0964
CNN[21]	4096	0.7179	3.4375	0.2146
	128	0.7402	3.4199	0.2146
wVLAD-SURF+ wVLAD-CN	4800	0.7702	3.5572	0.3907
	128	0.7954	3.4774	0.4066
wVLAD-SURF+ wVLAD-CN+CNN	8896	0.8220	3.7421	0.4311
	128	0.8306	3.6916	0.4322

Table 3. Comparison between various methods and our proposed method.

Methods	L	Holidays (mAP)	Oxford 5k (mAP)	Ukbench (N-S score)
VLAD-CSURF [6]	128	0.738	0.293	3.50
Triangulation embedding [22]	128	0.617	--	3.40
wVLAD-SURF+wVLAD-CN	128	0.7954	0.4066	3.4774
Deep fully conneted+VLAD [23]	512	0.783	--	--
OUR	128	0.8306	0.4322	3.6916
OUR	256	0.8435	0.4455	3.7215

We compare our method with various methods in Table 3. Specially, in reference [22], the authors constituted "hand-crafted" features like SIFT which is called triangular embedding. It can be seen that the mAP of wVLAD-SURF+wVLAD-CN is higher than the mAPs of other methods on Holidays. We fuse VLAD vectors and deep features. In reference [23], they proposed the multi-scale orderless pooling (MOP-CNN) scheme which combined the deep features and VLAD. The results compared with [23] are listed in Table 3. It can be seen that our method achieves the better results on three datasets.

4 Conclusion

Since the contribution of each descriptor to the VLAD vector is not the same in traditional VLAD method, it will result in visual burst phenomenon. To address the problem, we added a different weight for every residual to balance the contribution of each descriptor for the VLAD vector. The SURF features describe local gradient information of an image, while the CN features represent local color information. Thus, to improve the image retrieval accuracy, we proposed a simple and effective method that fused our proposed weighted VLAD vectors based on local texture features and local color features. Moreover, in order to improve the accuracy further, deep features are extracted and fused with the multiple weighted VLAD vectors. In order to reduce running time, the PCA and whitening operations were adopted in this paper. Finally, our experiments obtain the better results when compared with other methods.

This work was supported by National Natural Science Foundation of China (61572067); International (Regional) Project Cooperation and Exchanges of National Nature Science Foundation of China (61611530710); Beijing Municipal Natural Science Foundation (4162050); The Natural Science Foundation of Guangdong Province (2016A030313708) and the Fundamental Research Funds for the Central Universities (2017JBZ108).

References

1. J. Sivic, A. Zisserman, *Proceedings Ninth IEEE International Conference on Computer Vision*, **2**,1470-1477 (2003)
2. J. Philbin, M. Isard, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 1-8 (2007)
3. H. Jégou, M. Douze, C. Schmid, *IEEE In Proc. CVPR*, 3304-3311(2010)
4. J. Delhumeau, P. H. Gosselin, *ACM International Conference on Multimedia*,653-656 (2013)
5. D. Lowe, International Journal of Computer Vision, IJCV, no.2, 91-110 (2004)
6. H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, Computer vision and image understanding, no.3, 346-359 (2008)
7. E. Spyromitros-Xioufis, S. Papadopoulos, I.Y. Kompatsiaris, IEEE Transactions on Multimedia, no.6, 1713-1728(2014)
8. Q. S. Chen, Y. Y. Ding, H. Li, J. Wang, X. Deng, *IEEE International Conference on System, Man, and Cybernetics,* 2391-2396(2014)
9. D. A. R. Vigo, F. S. Khan, J. V. D. Weijer, *20th International Conference on Pattern Recognition,* 1549-1553(2010)
10. P. Fan, A. Men, M. Chen, *IEEE International Conference on Network Infrastructure and Digital Content*, 726-730(2009)
11. J. V. D. Weijer, C. Schmid, J. Verbeek, IEEE Transactions on Image Processing a Publication of the IEEE Signal Processing Society, no.**7**, 1512-23(2009)
12. A. Krizhevsky, I. Sutskever, G. E. Hinton, Advances in Neural Information Processing Systems, 25, 1097-1105 (2012)
13. M. D. Zeiler, R. Fergus, *In Proc. ECCV*, 8689, 818-833(2014)
14. A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, *In Proc*. *CVPR Workshops*, 806-813 (2014)
15. H. Jégou, O. Chum, *In Proc . ECCV*, 774-787(2012)
16. H. Jégou, M. Douze, C. Schmid, *In Proc. ECCV,* 304-317 (2008)
17. D. Nister, H. Stewenius, *In Proc*. *CVPR'06,* 2161-2168 (2006)
18. J. Philbin, O. Chum, M. Isard, *In Proc*. *CVPR* (2007)
19. J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman*, In Proc*. *CVPR*(2008)
20. M.J. Huiskes, M.S. Lew. *Proceeding MIR '08 Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39-43 (2008)
21. K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, *In proc. BMVC*(2014)
22. H. Jégou and A. Zisserman, *In Proc. CVPR*, 3310–3317 (2014)
23. Y. Gong, L. Wang, R. Guo, and S, Lazebnik. *In Proc. ECCV*, 392–407(2014).