Revisiting the VLAD image representation

Jonathan Delhumeau, Philippe-Henri Gosselin, Hervé Jégou, Patrick Perez

Revisiting the VLAD image representation

Jonathan Delhumeau
INRIA Rennes

Philippe-Henri Gosselin
ETIS, UCP/ENSEA/CNRS

Hervé Jégou
INRIA Rennes

Patrick Pérez
Technicolor Rennes

Abstract

Recent works on image retrieval have proposed to index images by compact representations encoding powerful local descriptors, such as the closely related vector of aggregated local descriptors (VLAD) and Fisher vector (FV). By combining such a representation with a suitable coding technique, it is possible to encode an image in a few dozen bytes while achieving excellent retrieval results. This paper revisits some assumptions proposed in this context regarding the handling of "visual burstiness", and shows that ad-hoc choices are implicitly done which are not desirable. Focusing on VLAD without loss of generality, we propose to modify several steps of the original design. Albeit simple, these modifications significantly improve VLAD and make it compare favorably against the state of the art.

1. Introduction

Content-based image retrieval (CBIR) is a historical line of research in Multimedia. It receives a particular attention from this community because images are ubiquitous and a key modality in numerous applications. The problem usually consists in finding the images in a database that are most similar to a query image. In recent years, many solutions have improved the search quality. In particular, a sustained line of research has been initiated by the bag-of-words representation [18, 4] and shown effective to up to million-sized image sets [12]. It consists first in describing an image by a collection of local descriptors such as SIFTs [10], and then in aggregating these into a single vector that collects the statistics of so-called "visual words".

Recently, another step toward further more scalable CBIR was achieved with the VLAD [8, 9] and the Fisher vector [13, 14]. These image representations are also produced from local descriptors, yet they propose an alternative aggregation stage, which replaces bag-of-words histograms. They are both built as the concatenation of subvectors (SIFT-like in case of VLAD), one per visual-word. One of their main merits is that they can be reduced to very compact vectors by dimensionality reduction, while preserving high retrieval accuracy. This vector can then be compressed with binary codes [15] or product quantization (PQ) [7], both allowing the efficient search in the compressed domain, thereby reducing the memory requirement by orders of magnitude.

This paper shows that VLAD (likewise FV) makes undesirable assumptions that yield suboptimal results. This leads us to modify VLAD at two levels of the original design: The per-word aggregation step and the vector rotation prior to component-wise application of power law.

Firstly, the local descriptors of a given image do not contribute equally to the original VLAD representation. This is due to the encoding scheme, which accumulates residual vectors (vector difference between local descriptors and visual words) of arbitrary norms. It was argued [9] that this effect naturally down-weights the most common descriptors, which are closer to the centroids. In contrast, we show that it is not desirable: Enforcing equal norms for the residual vectors provides better results, which constitutes our first beneficial modification to VLAD.

Secondly, we consider the power-law normalization [14] that is applied to VLAD component-wise. The improvement provided by this post-processing in VLAD and FV is usually explained by its effect on the "visual bursts" [5], *i.e.*, the patterns that may massively recur in an image, like in repetitive structures, corrupting the comparison metric. However, power-law normalization is obviously not invariant by rotation and thus depends on the coordinate system in which VLAD's sub-vectors live. In [9], these blocks are rotated from natural SIFT coordinate system, which is arbitrary from the burstiness point of view, to a common prelearned coordinate system. We propose a more elaborate way to optimize the basis so that it better captures the bursts on some components, thereby magnifying the positive effect of power-law normalization on accuracy.

This paper is organized as follows. Section 2 briefly reviews VLAD, while Section 3 presents our evaluation protocol. Section 4 describes our revisited VLAD, which is experimentally validated in Section 5 with comparisons to both original design and other state-of-art representations. Our experiments are performed on the popular Oxford5k [16] and INRIA Holidays benchmarks, as well as on an image set comprising 1 million images.

2. Original VLAD pipeline

The vector of locally aggregated descriptors (VLAD) [8] is an encoding technique that produces a fixed-length vector representation $\boldsymbol{v}$ from a set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n local d-dimensional descriptors, *e.g.*, SIFTs, extracted from a given image. Similar to bag-of-word, a visual dictionary $\mathcal{C} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ is learned off-line. It is formally used as a quantization function assigning any input local descriptors to its closest centroid (visual word) as

$$q : \mathbb{R}^d \to \mathcal{C} \subset \mathbb{R}^d \tag{1}$$

$$\mathbf{x} \mapsto q(\mathbf{x}) = \arg\min_{\boldsymbol{\mu} \in \mathcal{C}} \|\mathbf{x} - \boldsymbol{\mu}\|^2, \tag{2}$$

where the norm operator $\|.\|$ refers to the L_2 norm.

Aggregation. The VLAD departs from bag-of-words on how this visual dictionary is used. For each quantization index $i \in [1, \ldots, k]$, a d-dimensional sub-vector $\boldsymbol{v}^i$ is obtained by accumulating the *residual vectors*, *i.e.*, the difference between the descriptor and the centroid it is assigned to:

$$\boldsymbol{v}^i = \sum_{\mathbf{x}: q(\mathbf{x}) = \boldsymbol{\mu}_i} \mathbf{x} - \boldsymbol{\mu}_i. \tag{3}$$

The concatenation $\boldsymbol{v} = [\boldsymbol{v}^1 \ldots \boldsymbol{v}^k]$ is a D-dimensional vector, where $D = k \times d$.

Normalization. The VLAD is then obtained by applying two normalization stages. First, a component-wise non-linearity operation is applied: Each component v_j, $j = 1$ to D, is modified as $v_j := |v_j|^\alpha \times \mathrm{sign}(v_j)$, where the quantity α is a parameter such that $\alpha \leq 1$. It's the "power-law normalization" [14], which is motivated in [9] by the presence of bursts in natural images [5]. The VLAD vector is finally L_2-normalized as $\boldsymbol{v} := \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$.

SIFT processing. Prior to aggregation, it is optionally proposed in [8] to project all local descriptors of the image on the 64 first principal directions of a Principal Component Analysis (PCA) basis learned off-line.

Compact codes. The VLAD memory footprint is significantly reduced by performing a jointly optimized succession of dimension reduction [8, 9] and compression with product quantization [7]:

- The dimensionality of VLAD is reduced to $D' < D$ components by PCA, which is typically computed with the Gram dual method. See, for instance, [3] (paragraph 12.1.4).
- After a random rotation that balances the components of reduced vector, product quantization splits it into m sub-vectors, which are separately vector quantized with a k-means quantizer. This compression scheme allows the computation of distances between a query and a set of vectors in the compressed domain. It does not require the explicit decompression of the database vectors and is therefore very fast, see [7] for details.

The choice of D' and m is tuned thanks to an optimization procedure [7] that solely relies on a reconstruction criterion.

3. Evaluation protocol

In order to evaluate our work, we adopt some datasets and corresponding evaluation protocols that are usually considered in this context.

Local descriptors. They have been extracted with the Hessian-affine detector [11] and described by SIFT [10]. We use the RootSIFT variant [1], in all our experiments. As for SIFT, RootSIFT descriptors are normalized w.r.t. L_2 norm. We also report some results with a dense detector. This choice is common in classification [14] but usually not considered for large-scale image retrieval.

Oxford building datasets. *Oxford5k* dataset [16] consists of 5062 images of buildings and 55 query images corresponding to 11 distinct buildings in Oxford. The search quality is measured by the mean average precision (mAP) computed over the 55 queries. Images are annotated as either relevant, not relevant, or *junk*, which indicates that it is unclear whether a user would consider the image as relevant or not. These *junk* images are removed from the ranking before computing the mAP.

The *Oxford105k* dataset [16] is the combination of Oxford5k with a set of 100k negative images, in order to evaluate the search quality on a large scale.

The *Paris6k* dataset [17] consists of 6412 images collected from Flickr by searching for particular Paris landmarks. As standardly done in the literature, it is used for unsupervised learning of the parameters when evaluating the results on Oxford5k and Oxford105k.

INRIA Holidays and Flickr. INRIA Holidays [6] is a dataset comprising 1491 high resolution personal photos of different locations and objects, 500 of them being used as queries. The search quality is measured by mAP, with the query removed from the ranked list. To determine the parameters, we have used the independent dataset Flickr60K provided with Holidays. Large scale evaluation is performed by adding 1 million images collected from Flickr referred to as Flickr1M and used in [6] for large scale evaluation.

Parameters. For the power-law normalization, we will only consider two values: $\alpha = 1$ for no power-law normalization and $\alpha = 0.2$. This last choice is reasonable and often close to optimum for the regular VLAD. We fix it to ensure a fair comparison between the methods.

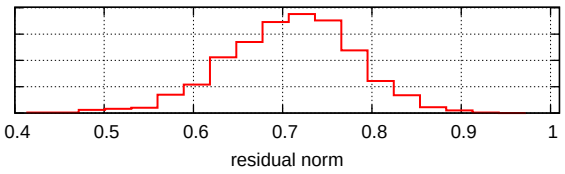

Figure 1. Norm's distribution of residual vectors in the VLAD of an image: because of the subtraction $\mathbf{x} - q(\mathbf{x})$, the contribution of individual descriptors in VLAD is uneven.

4. Improving VLAD

This section introduces two complementary techniques that jointly improve the VLAD representation by re-visiting some choices which are implicitly done in the initial design.

4.1. Residual normalization (RN)

The standard VLAD method sums up all residuals to shape the final representation (Eq. 3). Although the SIFT descriptors are L_2-normalized, it is not the case of the residual vectors, whose norm varies significantly, as shown in Figure 1 for a representative image. As a result, the individual local descriptors contribute unequally to the VLAD representation. This fact was underlined by the authors of VLAD [9], who argue that it provides some sort of natural *inverse document frequency*, but without evaluating its actual merit. As we will demonstrate later in our experiments, this is not a desirable behavior. More specifically, we propose to normalize the residuals so that all descriptors contribute equally (at least at this stage) to the summation. This amounts to changing (3) as

$$\dot{\boldsymbol{v}}^i = \sum_{\mathbf{x}:q(\mathbf{x})=\boldsymbol{\mu}_i} \frac{\mathbf{x}-\boldsymbol{\mu}_i}{\|\mathbf{x}-\boldsymbol{\mu}_i\|}. \qquad (4)$$

Note that the denominator is always greater than 0, because the SIFT descriptors lie on the unit sphere, while the centroids in $\mathcal{C}$ have norms strictly lower than 1.

Table 1 shows the interest of this modification. For the sake of comparison, we report the results both with regular SIFT (as results reported so far on VLAD and FV) and with RootSIFT, which already gives an improvement at no cost. Our discussion focuses on the relative improvement.

First, observe that the residual normalization (RN) tends to decrease the performance when $\alpha = 1$, *i.e.*, when no specific treatment is done to handle the bursts. Our interpretation is that descriptors coming into bursts are, on average, closer to their centroids, thus yielding lower normed residuals. However, one should not conclude that this implicit down-weighting of bursty patterns is beneficial: indeed, the power-law ($\alpha = 0.2$) appears as a better way to handle burstiness. When using it, RN consistently improves the results by almost 1 point of mAP at no cost. This suggests that all descriptors should equally contribute in the first place, while bursts are handled only by subsequent application of the power-law normalization. This is the first main modification we apply to original VLAD.

	PCA steps ↓	mAP on Holidays $\alpha = 1$	$\alpha = 0.2$
VLAD results with regular SIFT [9]			
VLAD	-	51.8	54.9
VLAD	C	51.8	54.0
VLAD	C,R	51.9	57.5
VLAD	C,R,D (64)	52.2	54.4
VLAD results with RootSIFT			
VLAD	-	53.9	57.3
VLAD	C,R	55.0	62.2
Impact of our method: RN and LCS			
VLAD+RN	C,R in LCS	54.3	63.1
VLAD+LCS	C,R in LCS	55.0	65.0
VLAD+LCS+RN	C,R in LCS	54.3	**65.8**

Table 1. From standard VLAD to its new version. (Top part): impact of (C)entering, (R)otation and (D)imensionality reduction by a factor 2 of regular SIFTs in original VLAD; (Middle part): impact of trading regular SIFT for RootSIFT; (Bottom part): impact of the two proposed modifications, RN and LCS. The dictionary size is $k = 64$ in all experiments.

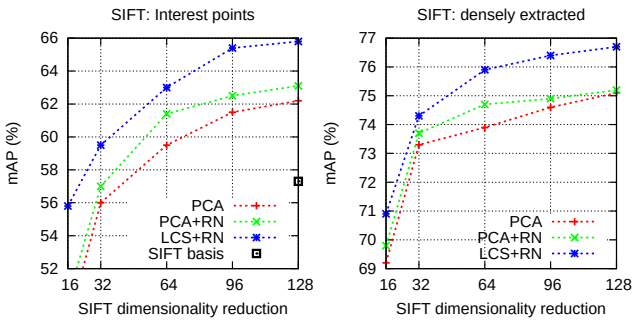

Figure 2. Performance on Holidays as a function of dimension reduction for RootSIFT descriptors (extracted from interest points or from a dense grid with 3 pixels shifts), using shared PCA (red), shared PCA and RN (green), LCS and RN (blue). Baseline is with no change on descriptors (black box). Parameters: $k = 64$ and $\alpha = 0.2$.

4.2. Local coordinate system (LCS)

Preliminary analysis of SIFT processing. It was shown that performing a PCA of SIFT descriptors improves Fisher and VLAD [9]. This processing actually encompasses three distinct operations: centering (C), rotation with PCA basis (R) and dimensionality reduction by a factor of 2 (D).

In present work, using RootSIFTs instead of such transformed SIFTs already yields a performance gain, achieving mAP=53.9 for $\alpha = 1$ and 57.3 for $\alpha = 0.2$ (Table 1). Nonetheless, in an attempt to get better insight into the original processing of regular SIFTs, we progressively incorpo-

rated the three above mentioned steps C, R and D (first part of Table 1) and made the following observations:

- For $\alpha = 1$, PCA rotation (R) has virtually no effect, which is to be expected: in the absence of power-law, this change of descriptor coordinate system has no impact on similarity measures between resulting VLADs.

- However, power-law normalization ($\alpha < 1$) introduces a subsequent non-linearity that makes final pipeline dependent on the basis. Using the basis obtained by PCA (on the Flickr60K dataset) gives a large improvement.

- Combined with power-law, dimension reduction (D) from 128 to 64 components decreases the performance (from mAP=63.1 to 61.4).

Therefore, it appears that it is not the dimensionality reduction that improves the results. On the contrary it is harmful, as also shown in Figure 2. The fundamental reason why PCA-based processing of local descriptors improves VLAD lies in its interplay with power-law normalization. In other terms, the impact of the power-law normalization (*e.g.*, for $\alpha = 0.2$) is magnified by a proper choice of the basis in which it is performed. This observation suggests to define a new and even better basis for building VLAD.

Proposed LCS. Our interpretation for PCA being a beneficial processing of SIFTs in conjunction with power-law normalization is that the first eigenvectors capture the main bursty patterns. However, this is so far a global operation applied to the whole descriptor space. It is therefore not likely to capture a large variety of bursty patterns. We argue that a better handling of burstiness should be achieved by adapting independently the coordinate system for each visual word. This is simply obtained by learning a "local" PCA per Voronoi cell of the partitioned feature space.

More precisely, for i-th visual word, $i = 1 \dots k$, we learn off-line (*e.g.*, on Flickr60K for Holidays) a rotation matrix $\mathbf{Q}_i$ from training descriptors mapped to this word. The k rotation matrices are then applied to the normalized residual vectors (or residual vectors if RN is not used) before their aggregation into VLAD. Equation 4 is thus replaced by

$$\tilde{\boldsymbol{v}}^i = \sum_{\mathbf{x}:q(\mathbf{x})=\boldsymbol{\mu}_i} \mathbf{Q}_i \frac{\mathbf{x}-\boldsymbol{\mu}_i}{\|\mathbf{x}-\boldsymbol{\mu}_i\|}. \quad (5)$$

Table 1 and Figure 2 show that this new LCS method, when combined with power-law, significantly improves the results. Also, it is complementary with proposed RN variant. The results are reported as a function of the dimensionality reduction in Figure 2 (blue curve), which clearly shows the improvement of our technique compared with original shared PCA (red curve) and with no PCA (black box).

Method	Size	Oxford5k	Oxford105k	Holidays
BoW [9]	20k	35.4	-	43.7
BoW [9]	200k	-	-	54.0
VLAD [9]	8192	37.8	-	55.6
Fisher [9]	8192	41.8	-	60.5
$\text{VLAD}_{\text{Intra}}$ [2]	8192	44.8	-	56.5
VLAD^*	8192	50.0	44.5	62.2
LCS+RN	8192	**51.7**	**45.6**	**65.8**

Table 2. Comparison of proposed image representation with state of the art (mAP performance).

5. Comparison to state of the art

The results presented in Table 1 and Figure 2 have shown the relative improvement provided by the residual normalization (RN) and by LCS on the INRIA Holidays benchmark. The relative gain is about $+4\%$ with respect to using only our SIFT processing (RootSIFT, (C) and (R)) and of $+10\%$ compared with the initial VLAD [9]. In this section, we provide a comparison with the state of the art and report our performance on other, larger datasets, namely the Oxford5k, Oxford105k and Holidays merged with 1 million images from Flickr. We use $k = 64$ in all experiments for the sake of consistency. We use the same SIFT descriptors as in [9] for a fair comparison.

Performance of the proposed methods - full vectors. Table 2 compares our technique with the results of the literature [9, 2], which includes different vector representations, in particular VLAD and FV, and a bag-of-words baseline. The improved VLAD obtained by our processing of SIFTs (RootSIFT, (C) and (R)) is referred to as VLAD^* in order to demonstrate the own merits of the methods proposed in Section 4. As one can see from this table, the improvement provided by VLAD^* is very large and further improved by LCS+RN, leading to outperform the best baseline by about 10% on Oxford5k and +5% on Holidays. We also compare our method to the intra-cell VLAD normalization recently proposed in [2] as a replacement for the power-law. In our experiments, this $\text{VLAD}_{\text{Intra}}$ variant is better than VLAD but not as good as the power-law when the input vectors are rotated by PCA. Our scheme significantly outperforms this choice in a similar setup. Note that we have not used multiple vocabularies in any of these comparisons.

Performance with projected and coded representations. Table 3 shows that the relative improvement of our technique is comparatively reduced when applying dimensionality reduction and using compact codes obtained by product quantization [7]. The gain remains very significant on the largest Holidays+Flickr1M dataset after dimensionality reduction, but not on Oxford105k. Overall, one should observe that compressing the data tends to reduce the gap between the different methods.

Method	size	Oxf5k	Oxf105k	Hol.+Flickr1M
After final dimensionality reduction to D'=96/128 components				
VLAD [9]	D'=128	28.7		
FV [9]	D'=96	-	-	31.8
FV [9]	D'=128	30.1	-	-
VLAD*	D'=128	**32.5**	**26.6**	33.5
LCS+RN	D'=128	32.2	26.2	**39.2**
Encoded into compact codes with product quantization [7]				
FV [9]	16 bytes	-	-	28.7
VLAD*	16 bytes	28.9	**22.2**	29.9
LCS+RN	16 bytes	27.0	21.0	**32.3**

Table 3. Large scale comparison of compacted and encoded image representations (k =64).

Dense. We also carried out experiments but with dense SIFT descriptors, reported in Fig. 2 (right). For better readability, we do not report in this case the result with no rotation of local descriptors (65.8%). As one can see, the same conclusions can be drawn when using dense or non-dense SIFT descriptors. One should note the very large improvement provided by dense descriptors.

6. Conclusion

This paper has analyzed the VLAD representation, and shows that it leads to sub-optimal results due to some undesirable properties: The descriptors do not contribute equally and the coordinate system used to apply the power-law normalization is arbitrary. We proposed two simple solutions to address these problems, at no cost. As a byproduct of our analysis, we have given some insight on the impact of the different steps of PCA on the accuracy.

Acknowledgments

Jonathan Delhumeau supported by the Quaero project (Oseo/French Agency for Innovation), Hervé Jégou and Philippe-Henri Gosselin supported by the FIRE-ID project (ANR-12-CORD-0016) and Patrick Pérez by the AXES project (FP7).

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, Jun. 2012.

[2] R. Arandjelovic and A. Zisserman. All about VLAD. In *CVPR*, Jun. 2013.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop Statistical Learning in Computer Vision*, 2004.

[5] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, Jun. 2009.

[6] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, Feb. 2010.

[7] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *Trans. PAMI*, 33(1):117–128, Jan. 2011.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, Jun. 2010.

[9] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local descriptors into compact codes. In *Trans. PAMI*, 34(9):1704–1714, Sep. 2012.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

[11] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, Oct. 2004.

[12] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, Jun. 2006.

[13] F. Perronnin and C. R. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, Jun. 2007.

[14] F. Perronnin, J.Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, Sep. 2010.

[15] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed Fisher vectors. In *CVPR*, Jun. 2010.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, Jun. 2007.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, Jun. 2008.

[18] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, Oct. 2003.