Revisiting the VLAD image representation

Jonathan Delhumeau⋆, Philippe-Henri Gosselin⋆, Hervé Jégou and Patrick Pérez⋆
⋆: INRIA Rennes  ⋆: Technicolor R&I, 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 them 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.

Keywords
image search, multimedia retrieval, VLAD

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, 0] 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 sub-vectors (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 compressed with binary codes [15] or product quantization (PQ) [8], 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 existing [9], these blocks are rotated from natural SIFT coordinate system, which is arbitrary from the burstiness point of view, to a common pre-learned 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 \( v \) from a set \( \{ x_1, \ldots, x_n \} \) of \( n \) local \( d \)-dimensional descriptors, e.g., SIFTs, extracted from a given image. Similar to bag-of-word, a visual dictionary \( \mathcal{D} = \{ \mu_1, \ldots, \mu_k \} \) is learned off-line. It is formally used as a
quantization function assigning any input local descriptors to its closest centroid (visual word) as

\[ q : \mathbb{R}^d \rightarrow C \subset \mathbb{R}^d \]  
\[ x \mapsto q(x) = \arg \min_{\mu \in \mathbf{C}} \| x - \mu \|^2, \]  

where the norm operator \( \| \cdot \| \) refers to the L2 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 \( v_i \) is obtained by accumulating the residual vectors, i.e., the difference between the descriptor and the centroid it is assigned to:

\[ v_i = \sum_{x \in q(x) = \mu_i} (x - \mu_i). \]  

The concatenation \( v = [v_1 \ldots v_k] \) 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 \( v_j := \| v_j \| \times \text{sign}(v_j) \), where the quantity \( \alpha \) 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 [9]. The VLAD vector is finally \( L_2 \)-normalized as \( v := \| v \|^{-1} v \).

**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, Paragraph 12.1.4 in Bishop’s Book [3].
- 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. It is formally optimize a retrieval performance.

**3. EVALUATION PROTOCOL**

In order to evaluate our work, we adopt some datasets and corresponding evaluations protocol that 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 [12] but usually not considered for large-scale image retrieval.

![Figure 1: Norm’s distribution of residual vectors in the VLAD of an image: because of the subtraction \( x - q(x) \), the contribution of individual descriptors in VLAD is uneven.](image)

**4. IMPROVING VLAD**

This section introduces two complementary techniques that jointly improve the VLAD representation by revisiting 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...
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.

<table>
<thead>
<tr>
<th>Impact of our method: RN and LCS</th>
<th>VLAD+RN</th>
<th>VLAD+LCS</th>
<th>VLAD+LCS+RN</th>
</tr>
</thead>
<tbody>
<tr>
<td>C,R in LCS</td>
<td>54.3</td>
<td>63.1</td>
<td>65.8</td>
</tr>
<tr>
<td>C,R</td>
<td>55.0</td>
<td>65.0</td>
<td></td>
</tr>
<tr>
<td>C,R,D (64)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 EV) and with RootSIFT, which already achieves 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 and handled only by subsequent application of the power-law normalization. This is the first main modification we apply to original VLAD.

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 incorporated 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 ($\alpha \neq 1$) 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 captures 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, $i = 1 \ldots k$, we learn off-line (e.g., on Flickr60K for Holidays) a rotation matrix $Q_i$ 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

$$\tilde{\phi} = \sum_{x \sim q(i)=\mu_i} \frac{x - \mu_i}{\|x - \mu_i\|}$$

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 $\alpha = 0.2$. 

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 vari-
ant. 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).

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), (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, 17], 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 16% on Oxford5k and +5% on Holidays. We also compare our method to the recent intra-cell normalization [2] proposed by Arandjelović for VLAD as a replacement for the power-law. In our experiments, this VLAD_int variant is better than VLAD but not 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.

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 made 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>Oxford5k</th>
<th>Oxford105k</th>
<th>Holidays</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW [9]</td>
<td>20k</td>
<td>-</td>
<td>35.4</td>
<td>43.7</td>
</tr>
<tr>
<td>BoW [9]</td>
<td>200k</td>
<td>-</td>
<td>-</td>
<td>54.0</td>
</tr>
<tr>
<td>Fisher [7]</td>
<td>8192</td>
<td>37.8</td>
<td>-</td>
<td>55.8</td>
</tr>
<tr>
<td>VLAD [9]</td>
<td>8192</td>
<td>41.8</td>
<td>-</td>
<td>60.5</td>
</tr>
<tr>
<td>VLAD_int [2]</td>
<td>8192</td>
<td>44.8</td>
<td>-</td>
<td>56.5</td>
</tr>
<tr>
<td>LSC+RN</td>
<td>8192</td>
<td>51.7</td>
<td>45.6</td>
<td>65.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>size</th>
<th>Oxford5k</th>
<th>Oxford105k</th>
<th>Holidays</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAD [9]</td>
<td>D=128</td>
<td>28.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FV [9]</td>
<td>D=96</td>
<td>-</td>
<td>-</td>
<td>31.8</td>
</tr>
<tr>
<td>FV [9]</td>
<td>D=128</td>
<td>30.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VLAD* [9]</td>
<td>D=128</td>
<td>25.8</td>
<td>34.5</td>
<td>39.2</td>
</tr>
<tr>
<td>LSC+RN</td>
<td>D=128</td>
<td>32.2</td>
<td>26.2</td>
<td>39.7</td>
</tr>
</tbody>
</table>

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

Acknowledgments

This project was partly done as part of the Quaero project, supported by Oseo, French agency for innovation. It was also supported by the FIRE-ID project (ANR-12-CORD-0016).

7. REFERENCES