More About VLAD: A Leap from Euclidean to Riemannian Manifolds

Masoud Faraki, Mehrtash T. Harandi, Fatih Porikli
College of Engineering and Computer Science, Australian National University. NICTA, Canberra Research Laboratory, Australia

This paper takes a step forward in image and video coding by extending the well-known Vector of Locally Aggregated Descriptors (VLAD) [5] onto an extensive space of curved Riemannian manifolds. In particular, we consider structured descriptors from visual data, namely Region Covariance Descriptors (RCovD) and linear subspaces that reside on the manifold of Symmetric Positive Definite (SPD) matrices and the Grassmannian manifolds, respectively. We introduce the Riemannian version of the conventional VLAD, called R-VLAD, a new coding approach that enables fusing local descriptors on these curved spaces.

The motivation stems from the fact that, in $\mathbb{R}^n$, coding local image or video descriptors using VLAD has been shown to be exceptionally successful in addressing a variety of challenging problems with a negligible computational cost compared to more involved approaches like deep convolutional networks [3, 5]. On the other hand, structured representations such as RCovDs and linear subspaces have been shown to provide robust and efficient representations for a wide range of tasks [1, 2, 4, 6]. Therefore, we provide mathematical framework that helps us aggregate local descriptors on curved spaces in a fashion similar to the conventional VLAD. In the sequel, we use $\mathcal{S}_{++}^d$ to specify the space of $d \times d$ SPD matrices and $\mathcal{G}_d^p$ to denote the space of arbitrary $d \times p$, $0 < p < d$, matrices with orthogonal columns, i.e., Grassmann manifold.

In VLAD [5], the input space $\mathbb{R}^d$ is partitioned into K Voroni cells by means of a codebook $\mathcal{C}$ with centers $\{c_i\}_{i=1}^K$, $c_i \in \mathbb{R}^d$. For a query set $\mathcal{X} = \{x_t\}_{t=1}^T, x_t \in \mathbb{R}^d$ extracted from an image or a video, the VLAD code $V \in \mathbb{R}^{Kd}$ is obtained by concatenating K Local Difference Vectors (LDV) v_i storing the differences $c_i - x_t$ in each cell, i.e.,

$$v_i = \sum_{x_t \in c_i} c_i - x_t \ , \tag{1}$$

where $x \in c_i$ means that the local descriptor x belongs to the Vornoi defined by c_i, i.e. the closest codeword to x is c_i.

Now assume that $\mathcal{X} = \{x_t\}_{t=1}^T$, $x_t \in \mathcal{M}$ and $\mathcal{C} = \{c_k\}_{k=1}^K$, $c_k \in \mathcal{M}$, are a set of local descriptors and codewords on a Riemannian manifold $\mathcal{M}$, respectively. The R-VLAD descriptor on $\mathcal{M}$ is obtained once we have a metric $\delta(x,y) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ to determine how the local descriptors should be assigned to the codewords and operators to perform the role of vector addition or subtraction on $\mathcal{M}$. We formulate our R-VLAD descriptor to support any metric on $\mathcal{M}$. As for the second requirement, we utilize the logarithm map, $\log_c(\cdot) : \mathcal{M} \to T_c\mathcal{M}$ as it is related to the gradient of the geodesic distance function $\delta_g : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ via the following equation:

$$\nabla_c \delta_g^2(c,x) = -2\log_c(x). \tag{2}$$

However, here the difficulty raises when the chosen metric is not the geodesic distance. As a result, choosing $\nabla_{c_i}\delta^2(c_i,x_t)$ for LDV will not work in practice. The main reason being that for δ_g, the norm of $\nabla_c\delta_g^2(c,x)$ is related directly to the metric, i.e.,

$$\|\nabla_c \delta_g^2(c,x)\|^2 = 4\|\log_c(x)\|^2 = 4\delta_g^2(c,x).$$

As an example, Fig. 1 shows the behavior of $\nabla_X\delta^2(X,Y)$ by varying $\delta^2(X,Y)$ for the projection metric on the Grassmann manifold $\mathcal{G}_3^2$. Interestingly, the norm of the gradient will start decreasing while point Y gets farther away from X. This means, during encoding, a point which should contribute significantly to the output, can act as an insignificant point, hence deteriorating the discriminatory power of the descriptor.

Here, the important message is that for a new metric $\delta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$, the length of the LDV should represent the metric considered on $\mathcal{M}$.

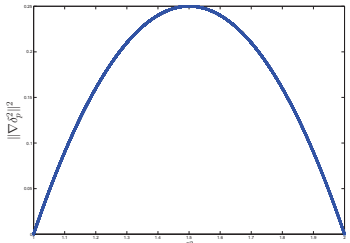

Figure 1: Illustration of the squared norm of the gradients vs distance for the projection distance on $\mathcal{G}_3^2$.

Table 1: Metrics and associated gradients on the Grassmann and SPD manifold.

Manifold	Metric	$\delta^2(X,Y)$	$\nabla_X\delta^2$
$\mathcal{G}_d^p$	geodesic	$\|\Theta\|^2$	obtained numerically
$\mathcal{G}_d^p$	projection	$2p - 2\|X^TY\|_F^2$	$-4(\mathbf{I}_d - XX^T)YY^TX$
$\mathcal{S}_{++}^d$	geodesic	$\|\log(X^{-1/2}YX^{-1/2})\|_F^2$	$2X^{1/2}\log(X^{-1/2}YX^{-1/2})X^{1/2}$
$\mathcal{S}_{++}^d$	Stein	$\ln\det\left(\frac{X+Y}{2}\right) - \frac{1}{2}\ln\det(XY)$	$X(X+Y)^{-1}X - \frac{1}{2}X$
$\mathcal{S}_{++}^d$	Jeffrey	$\frac{1}{2}\mathrm{Tr}(X^{-1}Y) + \frac{1}{2}\mathrm{Tr}(Y^{-1}X) - d$	$\frac{1}{2}X\left(Y^{-1} - X^{-1}YX^{-1}\right)X$

As such, we propose the following form of LDV for our general R-VLAD descriptor.

$$v_i = \sum_{x_t \in c_i} \delta(c_i,x_t)\frac{\nabla_{c_i}\delta^2(c_i,x_t)}{\|\nabla_{c_i}\delta^2(c_i,x_t)\|} \ . \tag{3}$$

The gradients used in R-VLAD for the studied metrics are depicted in Table 1.

Through rigorous experimental validation, we demonstrate the superiority of this novel Riemannian VLAD descriptor on several visual classification tasks including video-based face recognition, dynamic scene recognition, and head pose classification. To this end, we assess and contrast the performance of R-VLAD against Riemannian Bag of Words (BoW) model using different metrics and their Log-Euclidean counterparts as well as the state-of-the-art.

[1] Masoud Faraki, Mehrtash T Harandi, Arnold Wiliem, and Brian C Lovell. Fisher tensors for classifying human epithelial cells. *Pattern Recognition (PR)*, 47(7):2348–2359, 2014.

[2] Masoud Faraki, Mehrtash T. Harandi, and Fatih Porikli. Material classification on symmetric positive definite manifolds. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 749–756, 2015.

[3] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proc. European Conference on Computer Vision (ECCV)*, pages 392–407. Springer, 2014.

[4] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian Lovell. Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 3120–3127. IEEE, 2013.

[5] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(9):1704–1716, 2012.

[6] Raviteja Vemulapalli, Jaishanker K Pillai, and Rama Chellappa. Kernel learning for extrinsic classification of manifold features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1782–1789. IEEE, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.