Oriented pooling for dense and non-dense rotation-invariant features

Wan-Lei Zhao[1]
http://www.cs.cityu.edu.hk/~wzhao2
Hervé Jégou[1]
http://people.rennes.inria.fr/Herve.Jegou
Guillaume Gravier[2]
http://people.irisa.fr/Guillaume.Gravier

[1] INRIA
Rennes, France
[2] CNRS/IRISA
Rennes, France

Bag-of-words has been extended in various ways. This paper enriches the representation of each descriptor with the (quantized) characteristic scale and dominant orientation [2] associated with the region of interest. This additional information is exploited by a Hough-like voting procedure [1] to favor the images that have been scaled and rotated consistently.

In this paper, we depart from most existing works by using *rotation-invariant dense features*. In image search, most systems rely on keypoints or region detectors. A description relying on regular dense features achieves good performance but at the cost of loosing invariance to orientation, which is not desirable in many applications. Pooling on dominant orientations of the local features have been explored in the context of object classification [5], however they do not enforce the relative orientation to be preserved. In our work, we aim at obtaining both the discriminative power conveyed by dense descriptors and invariance to orientation.

The vector of locally aggregated descriptors (VLAD) [4] is an encoding technique that produces a fixed-length vector representation v from a set $\mathcal{X} = \{x_1, ..., x_m\}$ of m local d-dimensional descriptors (*e.g.*, SIFT, $d = 128$), which have been extracted from a given image. The VLAD computation procedure relies on a visual vocabulary $\mathcal{C} = \{c_1, \cdots, c_k\}$ where the dictionary (size k) is trained offline with k-means algorithm. It is used by a quantization function $q : \mathbb{R}^d \rightarrow \mathcal{C}$ that associates x_i to its Euclidean nearest neighbor in the vocabulary $\mathcal{C}$, as $q(x) = \arg\min_{c \in \mathcal{C}} \|x - c\|$. VLAD is a $d \times k$ vector, where each component is indexed by both the indices i and j associated to the quantization indexes and sift components, respectively. A component of VLAD vector $v = [v_{1,1}, \ldots, v_{i,j}, \ldots, v_{k,d}]$ associated with $\mathcal{X}$ is obtained as

$$v_{i,j} = \sum_{x \in \mathcal{X} : q(x) = c_i} x_j - c_{i,j}, \tag{1}$$

where x_j and $c_{i,j}$ are the j^{th} components of descriptor x and visual word c_i, respectively. As a post-processing, the vector v is ℓ_2-normalized.

To boost the performance of VLAD, several pre- and post-processing operations have been proposed, such as pairwise square-root on SIFT, rotating SIFT with PCA and applying pairwise power law on VLAD. VLAD undergone these series operations is denoted as VLAD*.

In the regular VLAD, features with different orientation are aggregated, loosing the possibility to estimate any geometrical transformation between the aggregated features. To alleviate this problem, we propose to pool features according to some characteristic geometrical quantities, more specifically the characteristic scales and dominant orientations [6] obtained as a byproduct of the descriptor computation stage. We aggregate features having similar characteristic scales or dominant orientations, to obtain a new pooling strategy termed Covariant-VLAD (CVLAD).

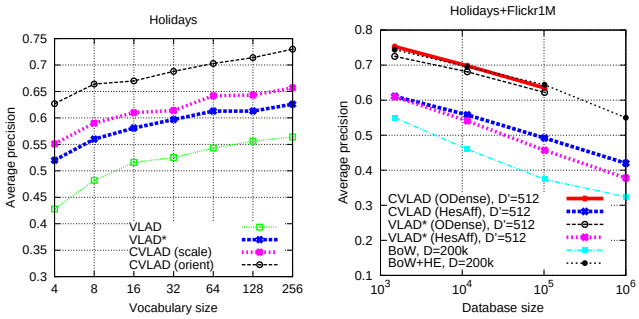
Take pooling on dominant orientation as an example. Let denote by θ the dominant orientation associated with a given feature x, and let

$$b_B(\theta) = \left\lfloor B \frac{\theta}{2\pi} \right\rfloor \tag{2}$$

be the quantization function used to quantize angles with B equally sized bins. Our pooling strategy modifies Equation 1 as

$$p_{b,i,j} = \sum_{x \in \mathcal{X} : q(x) = c_i \wedge b_B(\theta) = b} x_j - c_{i,j}. \tag{3}$$

In Eqn. 3, the pooling of the feature x is controlled by both its quantization index $q(x)$ and its quantized dominant angle $b_B(\theta)$. Another way to see the CVLAD construction procedure is to consider that $\mathbf{P} = [\mathbf{P}_1, \ldots, \mathbf{P}_B]$ is a concatenation of B VLAD $k \times d$-dimensional vectors, each of which encodes the features having the same quantized dominant orientation. This produces a vector B times longer than VLAD. Similar pooling is feasible with the characteristic scale. Note that the series pre- and post-processing adopted in VLAD* is also applied on each sub-vector of CVLAD.



(a) CVLAD *vs* VLAD and VLAD* (b) Search in large-scale

Figure 1: Performances of CVLAD in Holiday datasets. (a) Performances of pooling with dominant orientation (orient) and characteristics scale (scale) in comparison with conventional VLAD and VLAD*. (b) Performance of CVLAD in large-scale image search as a function of database size. The performance of VLAD has been compared with CVLAD*, BoW and BoW+HE [3]. Hessian-Affine+SIFT and dense SIFT have been adopted in experiment (a) and (b) respectively.

The similarity $s(.,.)$ between two CVLAD vectors $\mathbf{P}^i$ and $\mathbf{P}^j$ is defined on the basis of VLAD sub-vectors as

$$s(P^i, P^j) = \underset{\Delta t \in 0 \ldots B-1}{\operatorname{argmax}} \sum_{t=0}^{B-1} \cos\left(\mathbf{P}^i_t, \mathbf{P}^j_{\ \mathrm{mod}\,(t+\Delta t, B)}\right) \tag{4}$$

which amounts to selecting the orientation maximizing the similarity between the two vectors. This process is comparable to estimating the dominant rotation transformation between two feature sets in WGC [3], however here it is done directly on the aggregated vectors.

Eqn. 4 performs a circulant matching between two sets of VLAD subvectors. The matching shifts the VLAD sub-vector in P^j circularly to search for the best match between two groups of VLAD.

We presented CVLAD which pools on dominant orientations of local features. This approach builds upon the recent VLAD descriptor, but offers a new trade-off with respect to geometrical invariance by implicitly selecting the rotation (likewise scale) to maximize the similarity for a given image pair.

[1] Paul V. C. Hough. Method and means for recognizing complex patterns. US patent 3069654, Dec. 1962.

[2] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, Oct. 2008.

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

[4] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local descriptors into compact codes. In *Trans. PAMI*, Sep. 2012.

[5] Piotr Koniusz, Fei Yan, and Krystian Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 17 (5):479–492, 2013.

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