IMPROVING IMAGE SIMILARITY WITH VECTORS OF LOCALLY AGGREGATED TENSORS

David Picard and Philippe-Henri Gosselin

ETIS, CNRS, ENSEA, Univ Cergy-Pontoise
6 avenue du Ponceau, BP44
F95014 Cergy-Pontoise, France
{*picard,gosselin*}*@ensea.fr*

ABSTRACT

Within the Content Based Image Retrieval (CBIR) framework, three main points can be highlighted: visual descriptors extraction, image signatures and their associated similarity measures, and machine learning based relevance functions. While the first and the last points have vastly improved in recent years, this paper addresses the second point. We propose a novel approach to compute vector representations extending state of the art methods in the field. Furthermore, our method can be viewed as a linearization of efficient well known kernel methods. The evaluation shows that our representation significantly improve state of the art results on the difficult VOC2007 database by a fair margin.

Index Terms— Image classification, Image representation, Bag of Words

1. INTRODUCTION

In recent times, Content Based Image Retrieval (CBIR) benefited from significant advances in many of the fields it relies on. The first step is the introduction of powerful descriptors with high discriminative capacities like SIFT [1]. Either for similarity search or for image classification the local descriptors approach seems to be the most efficient for image matching. The second step is the introduction of efficient machine learning tools such as SVM with non-linear kernels [2]. Such techniques have great generalization capabilities that vastly improve results, and provide a unified framework for all image retrieval tasks.

In order to further improve the quality of the retrieval, recent work aims at filling the gap between descriptors and learning techniques. In this area, two steps can be investigated. The first consists in providing efficient image representation based on extracted descriptors. The main method for computing image representations is based on the "Bag of Features" scheme [3]. The latest works in this area show that results can be fairly improved by fine tuning image representations [4, 5]. The second step tackles the problem of similarity measures between image representations. Most advanced methods improve the similarity measure by combining several representations obtained from different features. Such combined similarities can be learned along with the classifier and thus further improve the results [6].

In this paper, we propose a new scheme of signatures aggregating a set of descriptors extracted from the images. It sums tensor product of the descriptors which we show is a good approximation of insightful similarity measures between descriptors. We find our scheme to outperform state of the art signatures on well known images.

This paper is organized as follows: the next section describes recent related work on image representation and similarities. Section 3 introduces our novel approach and gives insight on its soundness. We then present successful experiments on the very difficult VOC2007 dataset, before we conclude.

2. IMAGE REPRESENTATION AND SIMILARITIES

All recent approaches in image retrieval and classification use a set of local descriptors to encode the information contained in the image, such as SIFT [1].

Providing two images I_r and I_s, represented by two sets of local descriptors $B_r = \{\mathbf{b}_{ri} \in \mathbb{R}^N\}_i$ and $B_s = \{\mathbf{b}_{sj} \in \mathbb{R}^N\}_j$ (N being the dimension of descriptor space), the difficult point is to provide a relevant similarity measure between the sets B_r and B_s. Moreover, this similarity must conform the Mercer conditions in order to be used within the powerful Kernel Machines framework.

There are two main families of methods to obtain such similarities. The first category keeps the bag of descriptors in its entirety and tries to conduct similarities on bags of vectors. The second approach aims at producing a single vector representation for each image by aggregating its descriptors, and then using well known vector similarities (*i.e.* linear, rbf, polynomial, *etc*).

The remaining of this section details both families of methods.

2.1. Kernels on bags

Given two bags of vectors $B_r = \{\mathbf{b}_{ri} \in \mathbb{R}^N\}_i$ and $B_s = \{\mathbf{b}_{sj} \in \mathbb{R}^N\}_j$ representing the sets of descriptors (of dimension N) in two images, and a minor kernel function k on $\mathbb{R}^N \times \mathbb{R}^N$, the simplest kernel one can build to measure the similarity between these bags is the sum kernel [7]:

$$K(B_r, B_s) = \sum_{i,j} k(\mathbf{b}_{ri}, \mathbf{b}_{sj}) \quad (1)$$

However, this kernel tends to behave poorly when the size of the bags increase. In fact, the sum of noisy values (corresponding to non matching pairs $(\mathbf{b}_{ri}, \mathbf{b}_{sj})$) increases quadratically whereas the sum of relevant ones (*i.e.* matching pairs $(\mathbf{b}_{ri}, \mathbf{b}_{sj})$) increases linearly with the size of the bags.

To handle the increasing noise in the sum, one can use a highly discriminative minor kernel, such as the Gaussian kernel $k(\mathbf{b}_{ri}, \mathbf{b}_{sj}) = e^{-\frac{||\mathbf{b}_{ri} - \mathbf{b}_{sj}||^2}{\sigma^2}}$, with a very small standard deviation. Several kernels on bags have been proposed to handle the noise introduced by non-matching pairs by removing them from the sum [2, 7]. For instance, taking the average of maximum matching pairs is shown to have good performances in [7]:

$$\begin{aligned} K(B_r, B_s) &= \frac{1}{|B_r|} \sum_i \max_j k(\mathbf{b}_{ri}, \mathbf{b}_{sj}) \\ &\quad + \frac{1}{|B_s|} \sum_j \max_i k(\mathbf{b}_{ri}, \mathbf{b}_{sj}) \end{aligned} \quad (2)$$

The main drawback of the kernel on bags methods is their high computational cost, which increase quadratically with the size of the bags. Recently, approaches have been introduced trying to reduce the overhead in computation [8].

2.2. Vector representations

To get rid of the computational cost of kernels on bags, one can pack the set of descriptors of each image into a single vector representation. The most famous method of this kind is the *Bag of Features* (BOF) approach [3]. Providing a visual codebook of k prototypes of descriptors (named "*visual words*"), the produced vectors can be seen as the histogram of visual words occurrences within the bag. Visual words are usually obtained by k-Means clustering. Practically, each descriptor in a bag is assigned to its closest visual word, and the corresponding bin in the signature is increased.

Novel approaches are proposed to better optimize the signatures [5]. They split the problem in two steps. Firstly, they compute an assignment of each descriptors of the image to each visual codeword. For each image, they thus obtain a set of vectors corresponding to a projection of each descriptor onto the codebook. This step is called the "*coding*" step. Then, the set of assignment vectors is aggregated into a single vector of dimension k during what is named the "*pooling*" step. Taking the maximum value over the assignment values has been shown to be a relevant pooling step in [5].

In order to be efficient, the BOF approach requires a large codebook of several thousands visual words. However clustering high dimensional feature spaces (typically 128 dimensions for SIFT) into that many clusters is not an easy task. To tackle this problem, Jégou et al. proposed signatures named VLAD that use smaller codebooks [9]. Given a small codebook of visual words $\{\mathbf{c}_n\}_n$, they compute the vector $\boldsymbol{v}_n$ for each visual word $\mathbf{c}_n$, sum of difference vectors $\mathbf{b}_{ri} - \mathbf{c}_n$ of all descriptors $\mathbf{b}_{ri}$ for which $\mathbf{c}_n$ is the closest visual word:

$$\boldsymbol{v}_n = \sum_{\mathbf{b}_{ri} \text{ such that NN}(\mathbf{b}_{ri}) = \mathbf{c}_n} \mathbf{b}_{ri} - \mathbf{c}_n \quad (3)$$

The signature is obtained by concatenating the v_n vectors. Hence, its dimension is $k \times N$, where k is the size of the codebook, and N the dimension of the descriptors.

VLAD approach may be seen as a simplification of the Fisher vectors proposed in [10]. In this method, a parametric generative model is estimated on the training set. The description vector is the sum of gradients of the descriptors likelihood with respect to the parameters of the model. It encodes the deviation in parameter space of the samples relatively to the model. Generally, a Gaussian Mixture Model (GMM) with diagonal variance matrices is used as model.

3. VLAT: VECTOR OF LOCALLY AGGREGATED TENSORS

We propose to expend the VLAD approach by adding an aggregation of the tensor product of descriptors.

As in VLAD, we first compute a visual codebook by clustering the descriptor space using k-Means. Let $\mathbf{c}_n$ be the center (*i.e.* "visual word") of cluster n. For each visual word $\mathbf{c}_n$, we then compute a signature over the descriptors $\mathbf{b}_{ri}$ of image I_r that fall into cluster n. This signature is composed of several terms. The first term is the sum of differences between $\mathbf{b}_{ri}$ and $\mathbf{c}_n$, like in VLAD:

$$\mathcal{T}_n^1 = \sum_{\mathbf{b}_{ri} \text{ such that NN}(\mathbf{b}_{ri}) = \mathbf{c}_n} \mathbf{b}_{ri} - \mathbf{c}_n \quad (4)$$

The second term is the sum of self tensor product of descriptors $\mathbf{b}_{ri}$ centered to $\mathbf{c}_n$:

$$\begin{aligned} \mathcal{T}_n^2 &= \sum_{\mathbf{b}_{ri} \text{ such that NN}(\mathbf{b}_{ri}) = \mathbf{c}_n} (\mathbf{b}_{ri} - \mathbf{c}_n) \otimes (\mathbf{b}_{ri} - \mathbf{c}_n) \quad (5) \\ &= \sum_{\mathbf{b}_{ri} \text{ such that NN}(\mathbf{b}_{ri}) = \mathbf{c}_n} (\mathbf{b}_{ri} - \mathbf{c}_n)(\mathbf{b}_{ri} - \mathbf{c}_n)^\top \quad (6) \end{aligned}$$

More generally, we could add further tensor terms to the signature by computing higher order p of tensor products on centered descriptors:

Fig. 1. Images from PASCAL Visual Object Classes Challenge 2007.

$$\mathcal{T}_n^p = \sum_{\mathbf{b}_{ri} \text{ such that NN}(\mathbf{b}_{ri})=\mathbf{c}_n} \otimes_p (\mathbf{b}_{ri} - \mathbf{c}_n) \tag{7}$$

In practice, we limit the number of terms to the second order. The final image signature is the vector concatenating all the $\mathcal{T}_n^p$ flatten in vectors for all clusters n. A further ℓ_2 normalization of the signatures is done to ensure that images with different numbers of descriptors are still comparable.

3.1. Relation to kernels on bags

In order to prove the soundness of the VLAT approach, let us recall the sum kernel on bags for the Gaussian minor kernel:

$$K(B_r, B_s) = \sum_{i,j} e^{-\frac{||\mathbf{b}_{ri} - \mathbf{b}_{sj}||^2}{\sigma^2}} \tag{8}$$

If we assume that bag elements are normalized to unit ℓ_2-norm, we can simplify the above expression as:

$$K(B_r, B_s) = \sum_{i,j} e^{-\frac{2}{\sigma^2}} e^{\frac{2<\mathbf{b}_{ri}, \mathbf{b}_{sj}>}{\sigma^2}} \tag{9}$$

Let us now compute its Taylor expansion:

$$K(B_r, B_s) = \sum_{i,j} \sum_{p}^{\infty} \alpha_p (< \mathbf{b}_{ri}, \mathbf{b}_{sj} >)^p \tag{10}$$

With α_p being the coefficients of expansion. We now remark $(< \mathbf{b}_{ri}, \mathbf{b}_{sj} >)^p$ is the dot product between tensors of order p:

$$(< \mathbf{b}_{ri}, \mathbf{b}_{sj} >)^p = < \otimes_p \mathbf{b}_{ri}, \otimes_p \mathbf{b}_{sj} > \tag{11}$$

As the dot product is bilinear, we can further reduce the sum kernel to a sum of dot product between n-order tensor of elements of the bags:

$$K(B_r, B_s) = \sum_{p}^{\infty} \alpha_p < \sum_i \otimes_p \mathbf{b}_{ri}, \sum_j \otimes_p \mathbf{b}_{sj} > \tag{12}$$

When limiting the order of expansion to a finite number P, the sum can be interpreted as a dot product in a space concatenating flatten tensors. Furthermore, the higher the order of the expansion, the better is the approximation of the Gaussian kernel. However, this comes at the expense of the dimension D of the signatures, which increases exponentially with the order of expansion:

$$D = \sum_{p=1}^{P} k \times N^p \tag{13}$$

The main advantage of this approach is that it reduces a costly quadratic computation of exponentials to a simple dot product in a well chosen higher dimensional space obtained by non-linear mapping.

Using this interpretation, we can explain the underlying kernel on bags for the VLAT approach. It can be seen as the sum kernel on bags with the following minor kernel:

$$k(\mathbf{b}_{ri}, \mathbf{b}_{sj}) = \begin{cases} \sum_{p=1}^{P} < \otimes_p (\mathbf{b}_{ri} - \mathbf{c}), \otimes_p (\mathbf{b}_{sj} - \mathbf{c}) > \\ \qquad \text{if NN}(\mathbf{b}_{ri}) = \text{NN}(\mathbf{b}_{sj}) = \mathbf{c}, \\ 0 \quad \text{otherwise} \end{cases} \tag{14}$$

This kernel is a sum of dot products on tensors of centered vectors falling in the same cluster. We claim this kernel has two advantages: firstly, it reduces the noise of the sum kernel by removing pairs not falling in the same cluster. The higher the number of clusters, the less non matching pairs will be added to the sum. Secondly, the minor kernel measuring the similarity of a matching pair is non linear, and is indeed a good approximation of the Gaussian kernel.

4. EXPERIMENTS

We tested our method on the well known VOC2007 [11] dataset. This challenge is known as one of the most difficult image classification tasks. It consists in about 10000 images and 20 different categories. Fig. 1 shows random images taken from this dataset highlighting the variety of categories. We extracted SIFT features on a dense grid with 3 different scales in order to obtain bags of about 15000 SIFT for each image (about 150 millions of descriptors for the whole

dataset). All descriptors are reduced to only 32 dimension using Principal Component Analysis (PCA).

We compared our approach with VLAD and state of the art BOF approach like in [5]. The same set of descriptors were used for all methods. For the BOF approach, we generated a codebook of k =4000 visual words, which is known to have good performances. For the VLAD and VLAT approach, we had the parameter k varied from 16 to 64. All vector representations are normalized to unit ℓ_2-norm. We used a standard SVM classifier with a triangular kernel, trained on the *trainval* set, and evaluated on the *test* set. The hyperparameter C of the SVM is set to 10, regardless of the category. No spatial pyramids ([4]) were used for any of the methods, but all of them can be improved with it.

	BOF	VLAD			VLAT		
Category	4k	16	32	64	16	32	64
aeroplane	65.3	60,2	61,0	62,0	64,6	65,9	**66,1**
bicycle	44.5	35,3	38,6	40,3	45,3	47,2	**49,2**
bird	38.7	33,7	34,6	35,1	38,1	**39,6**	39,5
boat	49.4	53,7	55,8	55,3	57,4	**59,5**	58,7
bottle	18.7	13,9	16,0	17,2	17,2	18,5	**19,1**
bus	37.1	41,0	44,8	47,6	46,8	48,3	**48,9**
car	63.3	65,8	68,3	69,2	70,5	**71,1**	71,0
cat	33.9	35,2	37,9	40,9	40,8	41,1	**43,1**
chair	39.0	38,2	40,2	40,4	41,3	**43,1**	42,8
cow	24.8	22,7	22,7	25,3	25,2	26,9	**27,3**
diningtable	22.3	23,3	25,9	28,3	25,7	**30,4**	30,4
dog	26.9	34,1	32,4	34,0	36,6	37,0	**37,3**
horse	58.4	65,8	65,2	66,7	70,7	70,6	**70,8**
motorbike	36.2	45,2	47,2	49,2	49,6	51,1	**51,3**
person	75.6	75,6	77,4	78,0	79,4	**80,1**	79,9
pottedplant	11.9	11,9	14,4	**17,0**	14,9	15,8	16,7
sheep	24.2	20,2	26,1	**28,6**	23,1	28,2	28,2
sofa	33.9	34,3	36,2	35,0	39,4	**39,7**	38,9
train	57.7	60,7	64,6	64,3	66,1	**66,3**	65,3
tvmonitor	39.8	37,3	34,9	38,3	38,1	42,3	**42,4**
all	40.1	40,4	42,2	43,6	44,5	46,1	**46,4**

Table 1. Comparison of mAP between VLAD and VLAT for different size of dictionaries on VOC2007 dataset.

Tab. 1 sums up the *mean Average Precision* (mAP) for all methods. At first, we remark that the BOF approach is outperformed by VLAD, even with a much simpler codebook. This is consistent with recently reported results.

For same sized dictionaries, our method improve the results over VLAD by a fair margin of about **4%** of mAP. On the category *bicycle*, the improvement is even as high as **9%**. Furthermore, we find by comparing VLAD for $k = 64$ and VLAT for $k = 16$ that our method can lead to an improvement even with a smaller codebook, which highlight the soundness of the approach.

5. CONCLUSION

In this paper, we introduced a new vector image representation based on the aggregation of tensor products of local descriptors. Our representation can be seen as a generalization of state of the art representation methods, and outperforms them on the very difficult dataset VOC2007 by about 4%. When using more restrictive parameters such as a smaller codebook, our approach can still provide better results than state of the art with less restrictive parameters.

Moreover, we provide insightful clues showing that our method can be seen as an approximation of kernels on bags. This further explains the good performances of our signatures, and is encouraging to further developments.

6. REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, 2003, vol. 20, pp. 91–110.

[2] P.-H. Gosselin, M. Cord, and S. Philipp-Foliguet, "Kernel on bags of fuzzy regions for fast object retrieval.," in *IEEE International Conference on Image Processing (ICIP 07)*, San Antonio, Texas, USA, September 2007.

[3] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, Oct. 2003, vol. 2, pp. 1470–1477.

[4] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR '06*, Washington, DC, USA, 2006, pp. 2169–2178, IEEE Computer Society.

[5] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[6] David Picard, Nicolas Thome, and Matthieu Cord, "An efficient system for combining complementary kernels in complex visual categorization tasks," in *ICIP*, 2010, pp. 3877–3880.

[7] P.-H. Gosselin, M. Cord, and S. Philipp-Foliguet, "Kernel on bags for multi-object database retrieval.," in *ACM International Conference on Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands, July 2007.

[8] D. Gorisse, M. Cord, and F. Precioso, "Scalable active learning strategy for object category retrieval," in *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP10)*, 2010.

[9] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

[10] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV (4)*, 2010, pp. 143–156.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/.