BOSSA: EXTENDED BOW FORMALISM FOR IMAGE CLASSIFICATION

S. Avila [(1,2)], *N. Thome* [(1)], *M. Cord* [(1)], *E. Valle* [(3)], *A. de A. Araújo*[(2)]

(1) Université Pierre et Marie Curie, UPMC-Sorbonne Universities, LIP6, 4 place Jussieu, 75005, Paris, France

(2) Federal University of Minas Gerais, NPDI Lab – DCC/UFMG, Belo Horizonte, MG, Brazil

(3) State University of Campinas, RECOD Lab – IC/UNICAMP, Campinas, SP, Brazil

ABSTRACT

In image classification, the most powerful statistical learning approaches are based on the Bag-of-Words paradigm. In this article, we propose an extension of this formalism. Considering the Bag-of-Features, dictionary coding and pooling steps, we propose to focus on the pooling step. Instead of using the classical sum or max pooling strategies, we introduced a density function-based pooling strategy. This flexible formalism allows us to better represent the links between dictionary codewords and local descriptors in the resulting image signature. We evaluate our approach in two very challenging tasks of video and image classification, involving very high level semantic categories with large and nuanced visual diversity.

***Index Terms*—** Image classification, pattern recognition, Bag-of-Features, Bag-of-Words, visual dictionary, max pooling, sum pooling, SVM

1. INTRODUCTION

For image retrieval and classification tasks, some methods use complex structured models [1] to represent specific types of object, *e.g.* humans. Nevertheless, other approaches represent images by orderless local descriptors, such as the Bag-of-Words (BoW) model [2]. BoW becomes popular due to its simplicity and good performance. Inspired by the Bag-of-Words model from text retrieval [3], where a document is represented by a set of words, the BoW representation describes an image as a histogram of the occurrence rate of "words" in a vocabulary induced by quantizing the space of a low-level local descriptor (*e.g.*, SIFT [4]).

The basic BoW representation has important limitations, and several improvements have been suggested. To overcome the loss of spatial information, separate BoWs can be computed in different sub-regions of the image, as in the Spatial Pyramid Matching (SPM) scheme [5]. To attenuate the effect of coding errors induced by the descriptor space quantization, one can rely on soft assignment [6] or explicitly minimize reconstruction errors, *e.g.* Local Linear Coding [7]. Finally, averaging local descriptor contributions (sum pooling) can be reconsidered by studying alternative (more biologically plausible) pooling schemes, *e.g.* max pooling [8].

Our approach follows the BoW formalism, but proposes a new representation of images which keeps more information than BoW during the pooling step is proposed. The introduction of that new pooling function is the main contribution of this work. The resulting image signature process, called BOSSA (Bag Of Statistical Sampling Analysis), is based on a statistical analysis of the contribution of the local features to each visual word. Like [9], we carefully parametrize and normalize each block contribution, that are then concatenated all together to form a super-vector signature, with a reasonable vector dimensionality.

Contact: Sandra.Avila@lip6.fr. This work was partially supported by CAPES/COFECUB 592/08, CNPq 14.1312/2009-2, ANR 07-MDCO-007-03 and FAPESP 2009/05951-8.

2. BOW FORMALISM

Let us denote the set of local descriptors, *i.e.* the Bag-of-Features (BoF), by $\mathbf{X} = \{x_j\}$, $j \in \{1; N\}$, where each local feature $x_j \in \mathbf{R}^d$ and N is the number of local regions of interests on the image. In the BoW model, let us denote the visual dictionary as $\mathbf{C} = \{C_m\}$, $m \in \{1; M\}$, where M is the number of visual words. $\mathbf{Z} \in \mathbf{R}^M$ is the final vectorial representation of the image used for classification. In all the improvements over the basic BoW model, the mapping from $\mathbf{X}$ to $\mathbf{Z}$ can be decomposed into three successive steps, as formalized in [10]. The first step is a coding phase, where each local descriptor is projected to the visual dictionary. This coding phase can be modeled by a function f:

$$\begin{array}{rcl} f : \mathbf{R}^d & \longrightarrow & \mathbf{R}^M \\ x_j & \longrightarrow & f(x_j) = \alpha_{\mathbf{j}} = \{\alpha_{m,j}\}, \quad m \in \{1; M\} \end{array} \quad (1)$$

As illustrated in Figure 1, if we represent a matrix $\mathbf{H}$ with columns $\mathbf{X}$ and rows $\mathbf{C}$, the coding function f for a given descriptor x_j corresponds to the j^{st} column. The second step is a pooling step, that can be modeled by the following function g:

$$\begin{array}{rcl} g : \mathbf{R}^N & \longrightarrow & \mathbf{R} \\ \alpha_{\mathbf{m}} = \{\alpha_{m,j}\}, j \in \{1; N\} & \longrightarrow & g(\alpha_m) = z_m \end{array} \quad (2)$$

The pooling function g for a given visual word c_m corresponds to the m^{st} row of the $\mathbf{H}$ matrix, as shown in Figure 1.

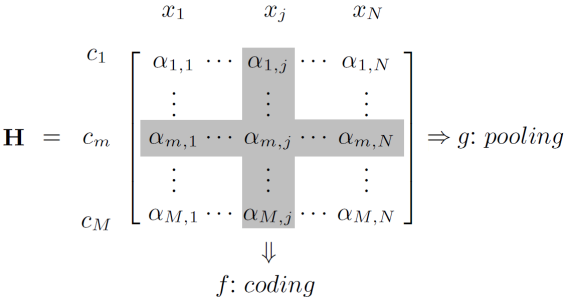

Fig. 1. BoW: $\mathbf{H}$ matrix representing coding and pooling functions.

For example, in the basic BoW representation:

- $f = f_Q$ assigns a constant weight to its closest center:

$$f_Q(x_j) = \begin{cases} 1 & \text{if } m = \underset{k\in\{1;M\}}{\operatorname{argmin}} \|x_j - c_k\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

- g computes the sum over the pooling region

$$z_m = \sum_{j=1}^{N} \alpha_{m,j} \quad (4)$$

The vector $\mathbf{Z}$, the final image representation, is given by sequentially coding, pooling and concatenating: $\mathbf{Z} = [z_1, z_2, \cdots, z_M]^T$. Regarding image categorization, the aim is to find out which operators f and g provide the best classification performance using z as input.

3. BOSSA: EXTENDING THE BOW POOLING

In this paper, we propose a new representation of images extending the BoW approach, called BOSSA (Bag Of Statistical Sampling Analysis). Basically, the idea is to keep more information than the BoW during the pooling step. Indeed, in BoW, the pooling step summarizes the vectorial information contained in $\alpha_{m,j}$ into a single scalar value (Equation 2): *e.g.* sum or max pooling. Instead, we propose here to estimate the distribution, *i.e.* the probability density function (pdf), of these $\alpha_{m,j}$.

3.1. Formalism

Regarding the coding function f defined in Section 2, each $\alpha_{m,j}$ coefficient traditionally quantifies a similarity between the descriptor x_j and the cluster c_m. In the following, however, $\alpha_{m,j}$ represents a dissimilarity (*i.e.* a distance) between c_m and x_j[1].

Therefore, keeping the same notations as in Section 2, the proposed modified g function aims at estimating the probability density function of α_m: $g(\alpha_m) = pdf(\alpha_{\mathbf{m}})$. We choose to estimate $pdf(\alpha_{\mathbf{m}})$ by computing the following histogram of distances:

$$\begin{aligned} g : \mathbf{R}^N &\longrightarrow \mathbf{R}^B \\ \alpha_{\mathbf{m}} &\longrightarrow g(\alpha_m) = z_m \\ z_{m,k} &= card\Big(x_j | \alpha_{m,j} \in \alpha_m^{max} \cdot \Big[\frac{k}{B}; \frac{k+1}{B}\Big]\Big) \end{aligned} \quad (5)$$

where B denotes the number of bins of each histogram z_m, and α_m^{max} is a parameter defined in Section 3.2.1. Thus, the g function represents the discrete (over B bins) density distribution of the distances $\alpha_{m,j}$ among the center c_m and the local descriptors of an image.

This step is illustrated on Figure 2. For each center c_m, we obtain a local histogram z_m. The histogram colors indicate the discretized spatial distances from the center c_m to the local descriptors shown by the black dots. For each colored bin $z_{m,k}$, the height of the histogram is equal to the number of local descriptors x_j which discretized distance with respect to cluster c_m fall into the k^{th} bin. We can note that if $B = 1$, the histogram z_m reduces to a single scalar value counting the number of points x_j falling into center c_m. Therefore, the proposed histogram representation can be considered as a consistent generalization of the BoW pooling step.

[1]This is performed without loss of generality, since estimating a similarity pdf for $\alpha_{m,j}$ from our model is straightforward. However, using rather a distance pdf makes illustrations clearer and more intuitive.

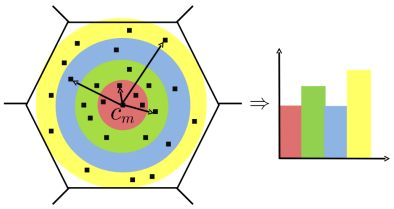

Fig. 2. Illustration of local histogram z_m.

After computing a local histogram z_m for all the c_m centers, we concatenate them to form the whole image representation. In addition, since we choose to ℓ_1 normalize each histogram z_m (see Section 3.2.2), the occurrence rate of each visual word c_m in the image is lost. To overcome this shortcoming, we propose to incorporate in our image representation $\mathbf{Z}$ an additional scalar value, which we denote as N_m, counting the number of points x_j falling at each center c_m. Thus, our final image representation $\mathbf{Z}$ can be rewritten as:

$$\mathbf{Z} = [[z_{m,k}], N_m]^T \qquad (m,k) \in \{1;M\} \times \{1;B\} \quad (6)$$

$\mathbf{Z}$ is a vector of size $D = M \times (B+1)$, as illustrated in Figure 3.

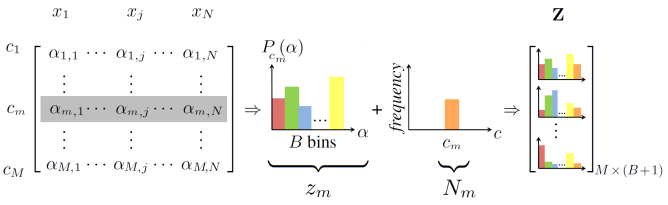

Fig. 3. Visual representation of BOSSA vector construction.

The idea enriching the BoW representation with extra knowledge from the set of local descriptors has been explored in [11, 12]. It is interesting to note, however, that those proposals rely in sophisticated statistical models, leading to very high-dimensional image representations. By using a simple histogram of distances to capture the relevant information, our approach remains very flexible and keeps the representation compact.

3.2. Implementation

3.2.1. Parameters

Our representation is defined by the three followings parameters: the number of visual words M, the number of bins B in each histogram z_m, and the maximum distance α_m^{max} in the $\mathbf{R}^d$ feature space to which z_m is computed (Equation 5).

M has a similar meaning than in standard BoW approaches. B defines the granularity to which $pdf(\alpha_{\mathbf{m}})$ is estimated. The choices of M and B are co-dependent, and $M \cdot B$ determines the compromise between accuracy and robustness. The smaller $M \cdot B$, the less the representation is accurate, the larger $M \cdot B$, the less the statistical estimate of the underlying distribution is confident (too large $M \cdot B$ values may lead to sparse vectorial representations). In our experiments, we use $M \sim 500$ and B in the range $[2; 10]$.

Finally, α_m^{max} is set up differently in each cluster c_m. Since our visual dictionary is built from a clustering algorithm (*e.g.* k-means),

we take advantage of the size (*i.e.* standard deviation σ_m) of the m^{th}, so that $\alpha_m^{max} = \lambda \cdot \sigma_m$, as shown in Figure 4. In practice, the two parameters of the BOSSA approach are B (M being fixed) and λ. In our experiments, we consider λ from 1 to 3.

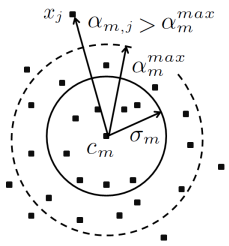

Fig. 4. Illustration of α_m^{max} parameter.

3.2.2. Normalization

The spatial information provided by bin counts in local histogram z_m is independent of spatial information provided by local histogram z_k for $m \neq k$. Therefore, we opted for histogram-wise normalization instead of globally normalizing z. We obtained our best results when using ℓ_1-normalization on each z_m, i.e. $z_m = z_m / \|z_m\|_1$.

Figure 5 illustrates the proposed image representations. We can notice the relevance of the improved BoW scheme, since intra-class variability is small whereas inter-class variability is large.

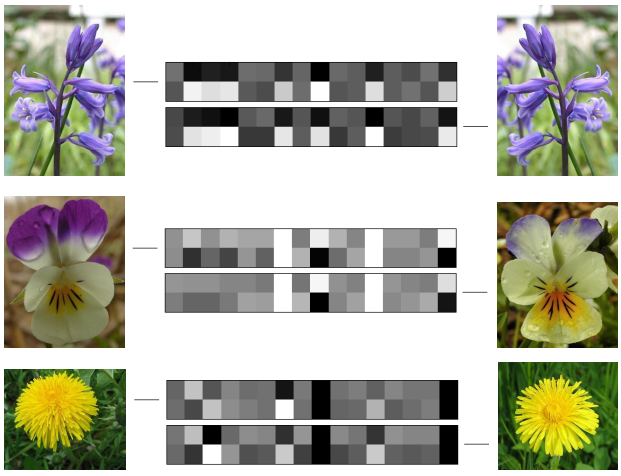

Fig. 5. Extracts of BOSSA signature on Oxford Flowers dataset. Images (left or right side) and the corresponding extracts BOSSA descriptors (middle grey level images).

4. EXPERIMENTAL RESULTS

We apply our proposal on two challenging datasets: Oxford Flowers [13] (image classification) and Pornography (video classification). As a low-level local descriptor, we have employed HueSIFT [14], a SIFT variant including color information, which is particularly relevant for our datasets. The 165-dimensional HueSIFT descriptors are extracted densely every 6 pixels.

We create a vocabulary by k-means clustering algorithm with Euclidean distance, fixing on 10% the number of sampling HueSIFT points. The vocabulary sizes we consider are $\{128, 256, 512\}$.

For classification, we apply the popular maximum-margin SVM classifier, specifically a non-linear χ kernel and the one-versus-all approach for multi-class approach. Kernel matrices are computed as $exp(-\gamma d(x, x'))$ with d being the distance and γ being fixed to the inverse of the pairwise distances mean.

We compare the performance of the classic BoW with the proposed BOSSA approach, which is an extension of the former.

4.1. Oxford Flowers

The Oxford Flowers dataset [13] contains 17 different flower categories with 80 images per category. Example images are shown in Figure 6.

Fig. 6. Example images from Oxford Flowers (one per category). The difference among categories is often subtle, even for humans.

The dataset comes already separated into three different folds, each with its own training (17×40 images), validation (17×20 images) and test sets (17×20 images). The accuracy rate is reported by the average scores of the three folds.

We use of the validation set to cross valid the parameter B (number of bins) with several values of dictionary size (from 128 to 512).

Table 1 presents the results for BOSSA and BOW using their best tested configuration parameters, namely $M = 512$, $B = 6$, $\lambda = 2$, C-SVM$_{BOSSA} = 10$ and C-SVM$_{BoW} = 1$.

Table 1. BOSSA and BoW classification performances on the Oxford Flowers.

	BOSSA	BoW
Acc.(%)	64	59
std (%)	± 2	± 1

Our BOSSA approach gives the best accuracy results comparing to the classical BoW approach on this dataset. In order to really appreciate the improvement coming from the difference between BOSSA and BoW, we do not have considered in our experiments extended representations of the BoW as the spatial pyramid representation of Lazebnik *et al.* [5] or any others. It will be interesting to consider feature combination expansions as we know that much higher scores may be obtained in classification when this combination is learnt [15].

4.2. Pornography database

We have also evaluated our approach after a real-world application, pornographic detection. The Pornography dataset contains nearly 80 hours of 400 pornographic and 400 non-pornographic videos. For the pornographic class, we have browsed websites which only

host this kind of material. For the non-pornographic class, we have browsed general-public purpose video network and selected two samples: 200 videos chosen at random (which we called "easy") and 200 videos selected from textual search queries like "beach", "wrestling", "swimming", which we knew would be particularly challenging for the detector ("difficult"). Figure 7 shows selected frames from the dataset.

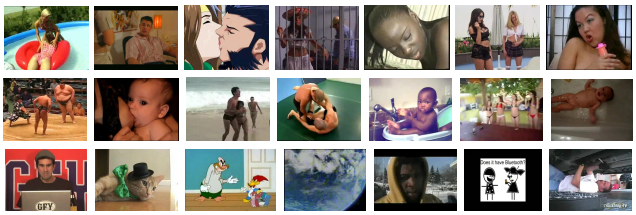

Fig. 7. Illustration of the diversity of the pornographic videos (top row) and the challenges of the "difficult" non-pornographic ones (middle row). The easy cases are shown at bottom row. The huge diversity of cases in both pornographic and non pornographic videos makes this task very challenging.

We preprocess this dataset by segmenting videos into shots. An industry-standard segmentation software[2] has been used. As it is often done in video analysis, a key-frame is selected to summarize the content of the shot into a static image. In our case, we have just selected the middle frame of each shot.

Both BoW and BOSSA image signatures are computed and used to train SVM and classify images. The image classification rate is reported by the mean average precision (MAP). For SVM, we use a 5-fold cross-validation to tune the best C parameter. Each method has been optimized considering its parameters (codebook size, normalization, etc.).

Table 2 shows the results for each method to their best tested configuration parameters ($M = 256$, $B = 10$, $\lambda = 3$, C-SVM$_{BOSSA} = 10$ and C-SVM$_{BoW} = 1$).

Table 2. BOSSA and BoW classification performances on the pornographic database. MAP are computed at image classification level, and Accuracy rate are reported for video shot classification.

	MAP (frames)	Acc. rate (videos)
BOSSA (%)	95 ± 1	87 ± 2
BoW (%)	91 ± 1	83 ± 3

In both Oxford Flowers and Pornography datasets, BOSSA outperforms the BoW approach, with a 4%-5% of improvement.

5. CONCLUSION

We proposed in this paper a new representation of images for classification tasks. Analyzing the popular Bag-of-Words scheme, we pointed out weakness in the standard pooling operation used in the BoW signature generation. The BOSSA scheme presented here offers a more information-preserving pooling operation based on a distance-to-codeword distribution. With this improvement to the basic pooling scheme, we carried out our final super-vector image signature used in SVM framework for classification.

Our scheme has the advantage of being conceptually simple, non-parametric and easily adaptable. Compared to other schemes existing in the literature to add information to the BoW model, it leads to much more compact representations.

We experimentally compared the performances of our BOSSA algorithm with the classic BoW on a standard image flower dataset as well as on a realistic application. In both cases, BOSSA performed better than BoW.

Feature combinations in a kernel learning framework is currently investigated in order to take advantages of all the features together.

[2]http://www.stoik.com/products/svc/

6. REFERENCES

[1] N. Thome, D. Merad, and S. Miguet, "Learning articulated appearance models for tracking humans: A spectral graph matching approach," *Image Communication*, vol. 23, pp. 769–787, 2008.

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

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1st edition, 1999.

[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.

[6] J. van Gemert, C. Veenman, A. Smeulders, and J-M. Geusebroek, "Visual word ambiguity," *PAMI*, vol. 32, pp. 1271–1283, 2010.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[8] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009, pp. 1794–1801.

[9] X. Zhou, K. Yu, T. Zhang, and T. Huang, "Image classification using super-vector coding of local image descriptors," in *ECCV*, 2010, pp. 141–154.

[10] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*, 2010, pp. 2559–2566.

[11] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007, pp. 1–8.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

[13] M-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *CVPR*, 2006, pp. 1447–1454.

[14] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, vol. 32, no. 9, pp. 1582–1596, 2010.

[15] D. Picard, N. Thome, and M. Cord, "An efficient system for combining complementary kernels in complex visual categorization tasks," in *ICIP*, 2010, pp. 3877–3880.