BossaNova at ImageCLEF 2012 Flickr Photo Annotation Task

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

[1] Pierre and Marie Curie University, UPMC-Sorbonne Universities, LIP6, France
[2] Federal University of Minas Gerais, NPDI Lab – DCC/UFMG, Brazil
[3] State University of Campinas, RECOD Lab – DCA/FEEC/UNICAMP, Brazil

sandra@dcc.ufmg.br, nicolas.thome@lip6.fr, matthieu.cord@lip6.fr, dovalle@dca.fee.unicamp.br, arnaldo@dcc.ufmg.br

Abstract. We present the BossaNova scheme for the ImageCLEF 2012 Flickr Photo Annotation Task. BossaNova is a mid-level image representation, recently developed by our team, that enriches the Bag-of-Words representation, by keeping a histogram of distances between the descriptors found in the image and those in the codebook. 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 Bag-of-Words model, it leads to much more compact representations. Furthermore, it complements well the cutting-edge Fisher Vector representations, showing even better results when employed in combination with them. In our participation, we submitted four purely visual runs. Our best result (MiAP = 34.37%) achieved the second rank by MiAP measure among the 28 purely visual submissions and the 18 teams.

Keywords: Image Classification, Image Representation, Bag-of-Words, Coding, Pooling, SVM

1 Introduction

The ImageCLEF 2012 Flickr Photo Annotation Task is a multi-label classification problem. The task can be solved by following three different approaches: i) automatic annotation with visual information only, ii) automatic annotation with textual information only, iii) multi-modal approaches that consider visual and textual information. We consider only the visual content for the feature extraction. The dataset consists of 25,000 Flickr images, splitting into training (15,000 images) and test (10,000 images) subsets.

The image set is annotated with 94 concepts that are very diverse and range across categories such as people (e.g., male, female), nature (e.g., lake, beach), weather (e.g., rainbow, fog) and even sentiments (e.g., unpleasant, euphoric). A detailed overview of the dataset and the task can be found in [1].

In our participation in the ImageCLEF 2012 Flickr Photo Annotation Task, we present our BossaNova scheme. Our aim is to emphasize the performance of

the BossaNova representation, using a single low-level feature (SIFT descriptors) and SVM classifiers. BossaNova is a mid-level image representation [2], recently developed by our team, that enriches the Bag-of-Words representation [3].

Bag-of-Words representations can be understood as the application of two critical steps [4]: *coding*, which quantizes the image local features according to a codebook or dictionary; and *pooling*, which summarizes the codes obtained into a single feature vector. Traditionally, the coding step simply associates the image local descriptors to the closest element in the codebook, and the pooling takes the average of those codes over the entire image.

Bossa Nova focus on the pooling step, by keeping a histogram of distances between the descriptors found in the image and those in the codebook. Our scheme has the advantage of being conceptually simple, nonparametric and easily adaptable. Additionally, it leads to much more compact representations, compared to other schemes to add information to the Bag-of-Words representation. Furthermore, it complements well the cutting-edge Fisher Vector representations [5], showing even better results when employed in combination with them.

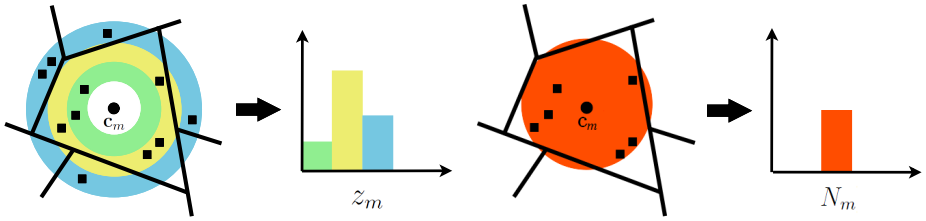
2 BossaNova Scheme

Our BossaNova scheme is composed of the following three steps: (i) extraction of local image features (by SIFT descriptors [6]), (ii) encoding of the local features in a global image representation (by a BossaNova representation [2]), and (iii) classification of the image representation (by SVM classifiers [7]). Here, we only provide a brief introduction to the BossaNova representation. More details can be found in [2][8].

BossaNova is a mid-level image representation which offers a more information-preserving pooling operation based on a distance-to-codeword distribution. In order to preserve a richer portrait of the information gathered during the coding step, the BossaNova pooling function produces a distance distribution, instead of compacting all information pertaining to a codeword into a single scalar, as performed by Bag-of-Words representations [3].

Figure 1 illustrates the BossaNova and the Bag-of-Words pooling functions. The BossaNova pooling (Figure 1a) represents the discrete (over B bins) density distribution of the distances between the codeword $\mathbf{c}_m$ and the local descriptors of an image. For each center $\mathbf{c}_m$, we obtain a local histogram z_m. The colors (green, yellow and blue) indicate the discretized distances from the center $\mathbf{c}_m$ to the local descriptors shown by the black dots. For each colored bin $z_{m,b}$, the height of the histogram is equal to the number of local descriptors, whose discretized distance to codeword $\mathbf{c}_m$ fall into the b^{th} bin. In Figure 1a, $B = 3$. We can note that if $B = 1$ (Figure 1b), the histogram z_m reduces to a single scalar value N_m counting the number of feature vectors falling into center $\mathbf{c}_m$.

To form the whole BossaNova image representation, all local histograms z_m are then concatenated. In addition, since the occurrence rate of each codeword $\mathbf{c}_m$ in the image is lost, BossaNova representation incorporates an additional scalar value N_m for each codeword, counting the number of local descrip-



(a) BossaNova pooling (b) Bag-of-Words pooling

Fig. 1. Illustration of BossaNova and Bag-of-Words pooling functions.

tors close to that codeword. That value corresponds to a Bag-of-Words term, accounting for a raw measure of the presence of the codeword $\mathbf{c}_m$ in the image. Thus, BossaNova image representation $\mathbf{z}$ can be written as [2]:

$$\mathbf{z} = \left[\left[z_{m,b}\right], sN_m\right]^{\mathrm{T}}, \quad (m,b) \in \{1,\ldots,M\} \times \{1,\ldots,B\}$$

where $\mathbf{z}$ is a vector of size $M \times (B+1)$, M is the number of codewords, and s is a weighted term learned via cross-validation.

BossaNova Parameters

The key parameters in our BossaNova representation are the number of codewords M, the number of bins B in each histogram z_m, and the range of distances $[\alpha_m^{min}, \alpha_m^{max}]$ – the minimum distance α_m^{min} and the maximum distance α_m^{max} in the $\mathbb{R}^d$ descriptor space that define the bounds of the histogram.

The bounds α_m^{min} and α_m^{max} define the range of distances for the histogram computation. Local descriptors outside those bounds are ignored. For α_m^{max}, the idea is to consider only descriptors that are "close enough" to the center, and to discard the remaining ones. For α_m^{min}, the idea is to avoid the empty regions that appear around each codeword, in order to avoid wasting space in the final descriptor.

In BossaNova, α_m^{min} and α_m^{max} are set up differently for each codeword $\mathbf{c}_m$. Since our codebook is created using k-means, we take advantage of the knowledge about the "size" of the clusters, given by the standard deviations σ_m. We set up the bounds as $\alpha_m^{min} = \lambda_{min} \cdot \sigma_m$ and $\alpha_m^{max} = \lambda_{max} \cdot \sigma_m$. In practice, the three parameters of the BossaNova become B (M being fixed), λ_{min} and λ_{max}.

3 Experimental Results

We first describe our experimental setup (Section 3.1). We then detail our submitted runs (Section 3.2). Finally, we analyze our results at ImageCLEF 2012 Flickr Photo Annotation Task (Section 3.3).

3.1 Experimental Setup

As low-level descriptors, we have extracted SIFT and Opponent SIFT descriptors [9] on a dense spatial grid, with the step-size corresponding to half of the patch-size, over 10 scales (SIFT) and 5 scales (Opponent SIFT) separated by a factor of 1.2, and the smallest patch-size set to 16 pixels. As a result, roughly $9,000$ SIFT and $7,000$ Opponent SIFT descriptors are extracted from each image of ImageCLEF 2012 Flickr Photo Annotation dataset. The dimensionalities of the descriptors are reduced by using principal component analysis (PCA), resulting in a 64-dimensional SIFT and a 128-dimensional Opponent SIFT.

To learn the codebooks, we apply the k-means clustering algorithm with Euclidean distance over one million randomly sampled descriptors. For Fisher Vectors [5], the descriptor distribution is modeled using a Gaussian mixture model (GMM), whose parameters (w, μ, Σ) are also trained over one million randomly sampled descriptors, using an expectation maximization algorithm. For all mid-level representations, we incorporate spatial information using the standard spatial pyramidal matching (SPM) scheme [10]. In total, we extracted 8 spatial cells $(1 \times 1, 2 \times 2, 3 \times 1)$.

One-versus-all classification is performed by support vector machine (SVM) classifiers. We use a linear SVM for Fisher Vectors, since it is well known that nonlinear kernels do not improve performances for those representations, see [5]. For BossaNova, we use a nonlinear Gauss-ℓ_2 kernel. Kernel matrices are computed as $exp(-\gamma d(x, x'))$ with d being the distance and γ being set to the inverse of the pairwise mean distances. For the combination of BossaNova and Fisher Vector representations, we apply a weighted sum of kernel functions. To map the SVM scores to probabilities we used a sigmoid function, $f(x) = (1 + exp(Ax + B))^{-1}$.

3.2 Submitted Runs

We have submitted four runs in total. All runs use only visual information.

Run 1 – ID 1341070721262: Combination of BossaNova and Fisher Vector representations. We use only SIFT descriptors. BossaNova parameters values are: 4096 codewords, 2 bins, 5-nearest codewords in semi-soft coding, $[0.4 \cdot \sigma_m, 2.0 \cdot \sigma_m]$ (range of distances for the histogram computation), see [2] for more details. Fisher Vectors are obtained with 384 Gaussians. We apply a sigmoid function to map the SVM scores to probabilities, where $A = 10$ and $B = 1$. This run achieved our best MiAP result and the second score by MiAP measure among all visual submissions.

Run 2 – ID 1341070953984: Combination of BossaNova and Fisher Vector representations. We use only SIFT descriptors. BossaNova parameters values are: 4096 codewords, 2 bins, 5-nearest codewords in semi-soft coding, $[0.4 \cdot \sigma_m, 2.0 \cdot \sigma_m]$ (range of distances for the histogram computation). Fisher Vectors are obtained with 384 Gaussians. We apply a sigmoid function to map the SVM scores to probabilities, where $A = 20$ and $B = 1$.

Run 3 – ID 1341348153832: BossaNova representation. We use only SIFT descriptors. BossaNova parameters values are: 4096 codewords, 2 bins, 5-nearest codewords in semi-soft coding, $[0.4 \cdot \sigma_m, 2.0 \cdot \sigma_m]$ (range of distances for the histogram computation). We apply a sigmoid function to map the SVM scores to probabilities, where $A = 10$ and $B = 1$. This run achieved the third score by MiAP measure among all visual submissions.

Run 4 – ID 1341348523492: Combination of BossaNova and Fisher Vector representations. We use SIFT and Opponent SIFT (only for Fisher Vector) descriptors. BossaNova parameters values are: 4096 codewords, 2 bins, 5-nearest codewords in semi-soft coding, $[0.4 \cdot \sigma_m, 2.0 \cdot \sigma_m]$ (range of distances for the histogram computation). Fisher Vectors are obtained with 384 Gaussians (for SIFT descriptors) and 128 Gaussians (for Opponent SIFT descriptors). We apply a sigmoid function to map the SVM scores to probabilities, where $A = 10$ and $B = 1$.

3.3 Results

In Table 1, we list the performance of our submitted runs. As detailed in [1], the following three quality metrics were evaluated to compare the submitted results: Mean interpolated Average Precision (MiAP), Geometric Mean interpolated Average Precision (GMiAP) and F-measure (F-ex).

Regarding the MiAP metric, we can notice that our best run reached 34.37% by combining the BossaNova and Fisher Vector representations (Run 1), achieving thus the second rank among the 28 purely visual submissions and the 18 teams. It is worthwhile to point out that, according to [2], the combination is performed by concatenating the vectors of BossaNova and Fisher Vector representations. Here, we opted to combine the two representations by a weighted sum of kernel functions, which is less time-consuming. The former combination, however, presents results slightly better over the latter. Therefore, we can improve our results even further.

Also, our BossaNova scheme (Run 3) achieved the third rank reporting 33.64% MiAP. Moreover, from Table 1, we can observe that using opponent SIFT (Run 4) as supplementary features does not bring any improvement. However, we consider that result is particularly affected by the severe dimensionality reduction of Opponent SIFT, from 392 to 128 dimensions.

4 Conclusion

In this paper, we presented our BossaNova scheme for the ImageCLEF 2012 Flickr Photo Annotation Task. Our method has the advantage of being conceptually simple, non-parametric and easily adaptable.

In our participation, we submitted four purely visual runs. Our best result (MiAP = 34.37%), which applied the combination of BossaNova and Fisher Vector representations, achieved the second rank by MiAP measure among the 28 purely visual submissions, while our BossaNova method achieved the third

Table 1. Overview of the different submissions.

Runs	MiAP (%)	GMiAP (%)	F-ex (%)	Type
Run 1	**34.37**	28.15	41.99	Visual
Run 2	33.56	27.75	37.86	Visual
Run 3	33.64	27.65	40.09	Visual
Run 4	33.56	26.88	42.28	Visual

rank (MiAP = 33.64%). The absolute difference between the first MiAP and our best MiAP is only 0.44%. We consider that those results are particularly noteworthy considering the fact we have not yet exploited the use of complex combinations of different low-level local descriptors.

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

Acknowledgements

This work is partially supported by CAPES/COFECUB 592/08/10, CNPq 14.13-12/2009-2, ANR 07-MDCO-007-03, FAPESP and FAPEMIG.

References

1. Thomee, B., Popescu, A.: Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: CLEF 2012 working notes, Rome, Italy (2012)
2. Avila, S., Thome, N., Cord, M., Valle, E., de A. Araújo, A.: Pooling in image representation: the visual codeword point of view. CVIU, Special Issue on Visual Concept Detection (under review)
3. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. Volume 2. (2003)
4. Boureau, Y., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. (2010) 2559–2566
5. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: ECCV. (2010) 143–156
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
7. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc. (1995)
8. Avila, S., Thome, N., Cord, M., Valle, E., de A. Araújo, A.: BOSSA: extended BoW formalism for image classification. In: ICIP. (2011) 2909–2912
9. Vedaldi, A., Fulkerson, B.: VLFeat – An open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia. (2010)
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178