IMAGE CLASSIFICATION USING OBJECT DETECTORS

Thibaut Durand (1), *Nicolas Thome* (1), *Matthieu Cord* (1), *Sandra Avila* (1,2)

(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

ABSTRACT

Image categorization is one of the most competitive topic in computer vision and image processing. In this paper, we propose to use trained object and region detectors to represent the visual content of each image. Compared to similar methods found in the literature, our method encompasses two main areas of novelty: introducing a new spatial pooling formalism and designing a late fusion strategy for combining our representation with state-of-the art methods based on low-level descriptors, *e.g.* Fisher Vectors and BossaNova.

Our experiments carried out in the challenging PASCAL VOC 2007 dataset reveal outstanding performances. When combined with low-level representations, we reach more than 67.6% in MAP, outperforming recently reported results in this dataset with a large margin.

Index Terms— Image Categorization, Object/Region Detectors, Spatial pooling, Combination with low-level Representations

1. CONTEXT

Image classification refers to the ability of predicting a semantic concept based on the visual content of the image. This topic is extensively studied due to its large number of applications in areas as diverse as Image Processing, Information Retrieval, Computer Vision, and Artificial Intelligence. This is also a very competitive research area and state-of-the art methods are currently evolving.

Bag-of-Words (BoW) representations using SIFT as low-level features proved to be the leading strategy in the last decade. Many attempts for improving the initial method [1] imported from text retrieval have been done, and an exhaustive review of the literature is far beyond the scope of the paper. However, we can highlight methods that enrich the BoW representation at the coding step, with sparse coding [2, 3], or with a vectorial representation, as done in Fisher Vectors [4]. It is also possible to improve the pooling step, as proposed with BossaNova [5, 6]. Using a spatially-preserving operator, *e.g.* Spatial Pyramidal Matching (SPM) [7], also proves to be crucial for reaching good classification performances.

Another strategy for image classification is to use deep networks with biologically inspired model [8, 9, 10]. Recently, deep learning has attracted lots of attention due to the large success of deep convolutional nets in the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)[1]. Using pixels as input, the network automatically learns useful image representations for the classification task. The results reveal that deep learning significantly outperforms competitors using BoW models with Fisher Vectors. However, although this trend is unquestionable for this large-scale context (1 million training examples), the feasibility of reaching state-of-the art performances in other complex datasets with fewer training examples, *e.g.* PASCAL VOC, remains unclear.

Methods using object detectors is another promising direction for building powerful image representations. Nowadays, the abundance of annotated images in Internet offers the opportunity to learn a wide range of object or regions classifiers. An increasing number of pre-trained detectors are even available, making their widespread use appealing. One option is to use the trained detectors using a sliding window strategy to produce detection maps for each concept. These maps then constitute mid-level representations or attributes that can further be processed to produce the final image representation. Pioneer works in that direction were proposed with Object Bank (OB) [11] and Classemes [12]. With OB [11], several regions and object detectors are used. These detectors are then spatially pooled to produce a compact image representation that is shown to give competitive results in various public datasets (15-Scenes, UCI-sports, MIT-indoor, *etc*).

2. PROPOSED IMAGE REPRESENTATION

In this paper, we propose a novel method using object detectors to create a discriminative and compact image signature. The whole pipeline of our approach is depicted in Figure 1.

As in Object Bank [11], we use object and region detectors as a first step in our process. However, we also suggest some improvements, both methodological and experimental. Our contribution can be summarized as follows. Firstly, we introduce a new method for spatially pooling response maps for object detectors. We enrich the signature compared to [11] by keeping n maximums, and propose a different method for pooling objects and region maps. We experimentally vali-

[1] http://www.image-net.org/challenges/LSVRC/2012/

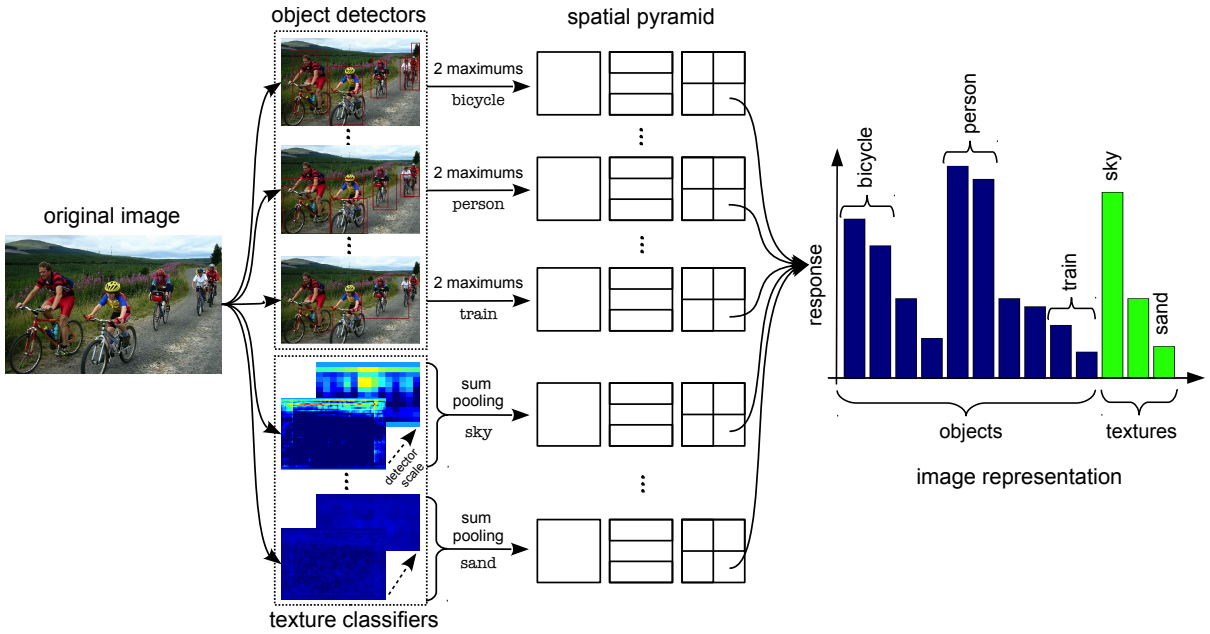

Fig. 1. Building Image Representation: the Proposed Pipeline

date that the proposed improvements favorably impact performances. Secondly, we propose to combine our representation with state-of-the art BoW-like representations, *i.e.*. [4, 6]. Last but not least, we carried out classification experiments on the challenging PASCAL VOC 2007 dataset. The proposed representation based on OB is by itself very competitive while being extremely compact. When combined to low-level representations, the results are shown to significantly outperform previously reported performances in this dataset.

2.1. Proposed Filter Bank

As in Object Bank (OB [11]), the main idea is using many object detectors as the basic representation of images. The word "object" is used in its very general form: an object can be a car, a person, a cat, sky, water, *etc*.

We use two well known object detectors: the latent SVM object detectors [13] for most of the blobby objects, and the texture classifier by Hoiem [14] for more texture- and material-based objects / regions. The latent SVM object detector uses a sliding window approach and works like a classifier: it determines whether or not there is an object of interest at the given position and scale of the image. For each scale, we obtain a response map. Then, there is a post-treatment for eliminating repeated detections via non-maximum suppression. The texture classifier by Hoiem has a functioning similar to the latent SVM object detector. But, there is not post-treatment. It returns a response map for each scale.

2.2. New spatial pooling strategy

The object and region filter banks output a set of region maps. As in [11], we aggregate the response maps in different spatial areas to extract a detection statistics. This pooling step has two interesting features: it helps to gain invariance and provides a compact signature for each image. If we consider N_c detectors and N_r spatial regions, the proposed representation Z concatenates the aggregation operator, denoted as $aggr(r, c)$ for each detector c and regions r, so that:

$$Z = [aggr(r, c)]_{(r,c)\in\{1;N_r\}\times\{1;N_c\}} \quad (1)$$

In our approach, however, the spatial pooling operator $aggr(r, c)$ differ from [11] in two important aspects.

2.2.1. *Spatial pooling with n maximums*

We propose to define a vectorial pooling operator for object detectors, denoted as $aggr(r, c) = n\text{-}max(r, c)$. Instead of keeping a scalar value (*e.g.* single max as done in [11]), we extract a vector of size n by taking the n bounding boxes with the n largest detection scores (after non maxima suppression). Spatial pooling with n maximums permits reduce the dimension while keeping information about the number of objects of class c present in region r.

2.2.2. *Different pooling for objects and textures*

On the contrary to [11], we propose to use a different pooling policy for objects and textures. In Object Bank [11], max-pooling is used for all detectors $c \in \{1; N_c\}$. Max-pooling

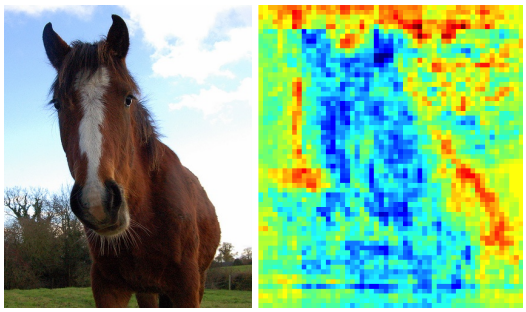

Fig. 2. Original image (left) and response map (one scale) for sky classifier (right). For this kind of texture classifier, sum-pooling is more appropriate than max-pooling. See Text.

seems to be appropriate for object detectors because we want detect the presence (or the absence) of an object in an image. However, applying the same strategy for region classifiers characterized by texture is more questionable. Indeed, texture features are intrinsically lower-level than object detector, so that false positives are supposed to occur more frequently. This is illustrated in Figure 2. Although the sky detector shown in Figure 2b) fits sky areas pretty well, there are some outliers, *e.g.* high values appear in the grass area.

Max-pooling exacerbates the impact of false positives done by texture classifiers, whereas sum-pooling is supposed to attenuate it. Therefore, we propose to use max-pooling for object detectors and sum-pooling for region classifiers, so that our spatial aggregation operator can be written as:

$$aggr(r,c) = \begin{cases} sum(r,c) & \text{if } c \text{ is a textured object} \\ n\text{-}max(r,c) & \text{otherwise} \end{cases}$$

If we denote N_{obj} and N_{text} to be the number of object and texture detectors, respectively, the final dimension of our image representation is $N_r \cdot (n \cdot N_{obj} + N_{text})$.

2.3. Combination with low-level representations

In this paper, we also explore the combination between our signature and low-level representations : BossaNova (BN) [6] and Fisher Vectors (FV) [4]. We aim at studying the complementarity of the information stemming from object detectors and information stemming from low-level representation. We propose a late fusion strategy: we first learn individual classifiers on our representation, that we denote as f_{ours} and on BossaNovaFisher, that we denote as f_{BNFV}. for each image x, the classification score $f(x)$ is then computed as a linear combination between f_{ours} and f_{BNFV}.

$$f(x) = \alpha \cdot f_{ours}(x) + (1-\alpha) \cdot f_{BNFV}(x) \qquad (2)$$

The weighting coefficient α represents the relative importance given to the two representations. It can be fixed heuristically, or learned by cross-validation.

3. EXPERIMENTS

In this section, we first describe our experimental setup and we then show our results on the PASCAL VOC 2007 dataset.

3.1. Experimental setup

The PASCAL VOC challenge 2007 [15] consists in recognizing 20 visual object classes in realistic scenes. It contains two main tasks: classification and detection. We evaluate our representation for the classification task. The dataset consists of 9,963 images, splitting into three subsets: *train* (2,501), *val* (2,510) and *test* (4,952). Our experimental results are obtained on *train+val* and *test* sets.

We choose to use 20 latent SVM object detectors which correspond to 20 object categories of the VOC 2007: *aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorbike, person, pottedplant, sheep, sofa, train* and *tvmonitor*. We employ the models available at [13]. We also utilize 6 texture classifiers of Object Bank [11]: *rock-stone, sand, water, sky, grass* and *building-edifice*. Those object and region filters are applied in each image, resulting in detection maps which are spatially pooled using the SPM strategy [7]. We use the standard SPM layout of the VOC 2007, a three-level pyramid ($1 \times 1, 2 \times 2, 3 \times 1$).

The aggregation method proposed in this paper outputs a vector Z for each image, as defined in Equation 1. An "one-versus-all" SVM classifier for each object class is then trained on the vectors obtained. The classification performance is evaluated using the Mean Average Precision (MAP) across all classes. Here, we use the LIBSVM library [16] to train the classifiers. We apply the C-SVC SVM type, with the constant value of $C = 1$. Radial basis function (RBF) 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.

To evaluate the complementarity of our signature and BNFV representation [6], we keep the BNFV parameters values the same as in [6]. The BossaNova (BN) parameters values are: 4,096 visual words, 2 bins, $\lambda_{min} = 0.4$ and $\lambda_{min} = 2.0$ (range of distances for the histogram computation), and $s = 10^{-3}$ (weighted factor); and the Fisher Vector (FV) is computed with 256 Gaussians. Furthermore, a linear and a Gauss-ℓ_2 kernel matrices are computed for FV and BN, respectively. The BNFV representation is obtained by combining the BN and FV kernel matrices. The MAP on PASCAL VOC 2007 with those parameters is 60.3%.

3.2. Classification results

Our method is an improved version of the OB [11] with respect to the pooling policy: 1) it keeps n maximums for objects (Section 2.2.1) and 2) it applies the sum-pooling for the textured objects (Section 2.2.2). In order to quantify the performance gain brought out by those contributions, we perform the following experiments.

	MAP	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
OURS	59.0	64.7	75.6	32.1	62.7	48.2	70.3	83.8	49.3	57.2	48.4
BNFV	60.3	79.5	65.6	53.6	72.1	32.7	66.0	79.0	59.7	54.5	43.0
LF = OURS+BNFV	**67.6**	80.8	78.8	55.4	73.8	52.2	76.6	86.4	64.1	62.1	55.2
		diningtable	dog	horse	motorbike	person	plant	sheep	sofa	train	tvmonitor
OURS		52.4	34.2	76.9	68.1	87.7	36.6	44.4	51.8	71.3	63.8
BNFV		60.0	46.8	78.6	64.8	84.5	31.2	45.3	54.6	78.5	55.1
LF = OURS+BNFV		66.6	49.7	83.4	74.7	89.8	37.9	50.7	64.1	80.9	68.9

Table 1. Image classification MAP (%) on PASCAL VOC 2007 dataset (LF: Late Fusion, OURS: our signature)

First, we compute our signature using object detectors only. We then vary the number of maximum n used for the n-max pooling method. Therefore, for $n = 1$ (our baseline, $\sim$ OB [11]), the MAP is 56.7%. For $n = 2$, the MAP is 57.3% (+0.6%). For $n > 2$, however, the performances start to drop.

Next, to validate our assumption that the sum-pooling for textured object is better than the max-pooling, we investigate both pooling strategies with a "texture signature", which is build with the 6 textured objects only. Thus, for the sum-pooling strategy, the MAP is 23.8%, while for the max-pooling approach is 14.9%. That confirms the relevance of the improvement introduced Section 2.2.2. Moreover, for some categories, its improvement reached up to 22.6% (sum *vs* max): *aeroplane* (48.0% *vs* 25.4%), *car* (42.7% *vs* 30.4%) *person* (61.7% *vs* 54.9%). Our final image representation consists in concatenating the "texture signature" with the n maximums signature. The results obtained with sum-pooling for textured objects is 59.0% (+2.3%), whereas only 58.0% (+1.3%) with max-pooling. In brief, our representation shows very good performance (59.0%), besides its extremely compact dimension: $(n \times 20 + 6) \times 8$. The detailed results are shown in Table 1.

Finally, we evaluate the complementarity of BNFV and our signature, with n maximums for objects and sum-pooling for textures. In order to assign equal importance to the both signatures, we choose $\alpha = 0.5$. From the Table 2, we can notice the considerable improvement obtained by late fusion, reaching a MAP of 67.6% (+8.6%/our signature). This corresponds to a remarkable success of the complementariness of BNFV and our signature. Furthermore, we can observe that the combination surpasses the state-of-the-art results (+7.3%/BNFV, +6.4%/FV[17]). We also outperform method that combine object detectors and low-level representations (+1,3%/[18], +1,0%/[19]), but [19] use more external data than us. To the best of our knowledge, *these is the best result reported so far on PASCAL VOC 2007*. Figure 3 gives an example of a Recall/Precision curve, for the bus category, comparing our approach with BNFV and the late fusion between the two methods.

Method	BNFV	[17]	[18]	[19]	OURS	LF
MAP (%)	60.3	61.7	66,3	66,6	59.0	**67.6**

Table 2. State-of-the-art results (MAP %) on PASCAL VOC 2007 dataset. We outperform best results with the proposed Late Fusion approach.

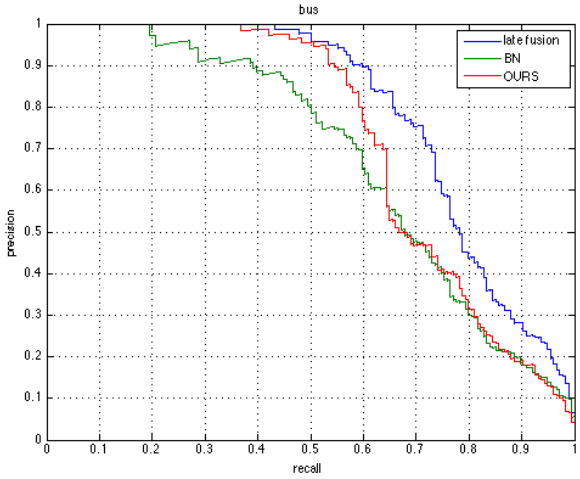

Fig. 3. Recall/Precision curve for the bus category

4. CONCLUSION

We present an approach for image classification based on object detectors, where a novel approach for pooling filter maps is proposed. The evaluation on the challenging VOC 2007 dataset show very good results with a very compact image representation. When combined with low-level representations, we outperform state-of-the-art performances. Directions for future works include a more proper representation learning from the detection maps using deep learning.

5. REFERENCES

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

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

[3] Hanlin Goh, Nicolas Thome, Matthieu Cord, and Joo-Hwee Lim, "Unsupervised and supervised visual codes with restricted boltzmann machines," in *Proceedings of the 12th European conference on Computer Vision - Volume Part V*, Berlin, Heidelberg, 2012, ECCV'12, pp. 298–311, Springer-Verlag.

[4] Florent Perronnin and Christopher R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[5] Sandra Eliza Fontes de Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo, "Bossa: Extended bow formalism for image classification," in *ICIP*, 2011, pp. 2909–2912.

[6] Sandra Eliza Fontes de Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

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

[8] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[9] Christian Theriault, Nicolas Thome, and Matthieu Cord, "Extended coding and pooling in the hmax model," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 764–777, 2013.

[10] Christian Theriault, Nicolas Thome, and Matthieu Cord, "Hmax-s: Deep scale representation for biologically inspired image categorization," in *ICIP*, 2011, pp. 1261–1264.

[11] L.-J. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010.

[12] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.

[13] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," `http://cs.brown.edu/~pff/latent-release4/`, 2010.

[14] D. Hoiem, A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, no. 3, 2005.

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

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, 2011.

[17] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011.

[18] J Sanchez, F Perronnin, and T E deCampos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, 2012.

[19] Yu Su and Frédéric Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision*, vol. 100, no. 1, 2012.