SEMANTIC POOLING FOR IMAGE CATEGORIZATION USING MULTIPLE KERNEL LEARNING

Thibaut Durand[1,2], *David Picard*[1], *Nicolas Thome*[2], *Matthieu Cord*[2]

[1] ETIS, UMR 8051 / ENSEA, Université Cergy-Pontoise, CNRS, F-95000, Cergy,

[2] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France

ABSTRACT

In this paper, we propose a new method for taking into account the spatial information in image categorization. More specifically, we remove the loss of spatial information in Bag of Words related methods by computing the image signature over specific regions selected by object detectors. We propose to select the detectors using Multiple Kernel Learning techniques. We carry out experiments on the well known VOC 2007 dataset, and show our semantic pooling obtains promising results.

Index Terms— Image classification, Image representation, Object detection, Statistical learning

1. INTRODUCTION

In this paper, we are interested in image categorization and more specifically in taking spatial information into account in the categorization pipeline. Most image categorization systems rely on computing a vector signature from the content of the image, followed by a classification algorithm using machine learning techniques [1]. The most successful approaches are refinement of the Bag of Words (BoW) model [2]. In the BoW model, local visual descriptors are extracted from the image and aggregated in a single vector using a codebook of descriptor prototypes. The codebook is usually computed by clustering a large set of randomly sampled descriptors. For example, the authors in [3] propose to measure the deviation of second order moments between the descriptors of the image and the codebook.

As the aggregation mostly involves a statistical analysis of the set of extracted descriptors, the spatial information (*i.e.*, the localization of the descriptors) is usually lost in the process. Some methods have been proposed to retain some of the spatial information, like in Spatial Pyramid Matching (SPM) [4] or in Spatial Coordinate Coding (SCC) [5]. However, most of these methods are not invariant to the layout of the objects in the image (*e.g.*, an image with the object on bottom left barely matches with an image where the object is on the top right).

In this paper, we propose to incorporate spatial information in the image categorization pipeline by taking into account the semantic layout of the images. Our main contribution is a new image categorization method using semantic pooling regions on which a signature is computed. These regions are obtained using object detectors and are thus associated with a semantic concept. Building on a kernel framework, we propose to select the relevant regions using Multiple Kernel Learning [6].

The paper is organized as follows: in the next section we present recent approaches to incorporate spatial information in image categorization systems based on BoW. Then, we present our method and discuss its advantages over existing approaches. In Section 4, we present experiments on the well known Pascal VOC 2007 dataset [7] and show our approach is indeed very competitive, before we conclude in the last section.

2. RELATED WORK

The general principle of the BoW model is as follows: First, a set of local visual descriptors such as HOG [8] or SIFT [9] is extracted from the image, either using keypoints detectors or on a dense grid. Let $\mathbf{B}_i = \{\mathbf{d}_{ri}\}_r$ be the set of descriptors $\mathbf{d}_{ri}$ extracted from image i. Using a large sample of such descriptors extracted from a wide variety of images, a codebook of descriptor prototypes $\{\mu_c\}$ (called visual words) is computed, usually using the k-means clustering algorithm. To compute the signature, the descriptors of an image are *projected* on the codebook so as to obtain a single vector. In the case of the classic BoW, the histogram of occurrences of the visual words is computed, which amounts to computing the number of descriptors assigned to each visual word.

There are several popular extensions of Bow with the aim to improve the corresponding similarity. For example, the authors of BossaNova [10] propose to estimate the distribution of descriptors for each visual word by computing the histogram of distances to the center of the cluster. In [11], the authors propose to model the deviation between the codebook and the descriptors of the image by computing for each visual word c the difference between the visual word and the mean

of the descriptors assigned to it:

$$\mathbf{x}_{ci} = \sum_{\mathbf{d}_{ri} \in C_c} (\mathbf{d}_{ri} - \mu_c), \tag{1}$$

with $C_c = \{\mathbf{d}_{ri} | \mu_c = \arg\min_k \|\mathbf{d}_{ri} - \mu_k\|\}$. The resulting VLAD signature is then the concatenation of all deviations $\mathbf{x}_{ci}$ for all visual words c. In [12], the authors propose to extend the VLAD by computing the deviation of second order moments for each cluster c:

$$\mathbf{x}_{ci} = \sum_{\mathbf{d}_{ri} \in C_c} (\mathbf{d}_{ri} - \mu_c)(\mathbf{d}_{ri} - \mu_c)^\top - \tau_c, \tag{2}$$

with τ_c beeing the covariance matrix of cluster c. The resulting VLAT signature is the concatenation of all such matrices, where only the upper part is kept thanks to symmetry. Fisher Vectors [13] also consider second order deviation, albeit with respect to a Gaussian mixture model of the descriptors space.

In all the presented signatures, the spatial layout of the descriptors is lost during the aggregation. To overcome this drawback and retain some of the spatial information, the authors of [4] propose to split the image on a regular and recursive grid and then to compute a signature for each of such regions. Using the formalism of Mercer kernels, the method is called Spatial Pyramid Kernel (SPK). The full signature is simply the concatenation of the signatures obtained for all regions, or equivalently the sum of corresponding kernels. The SPK is shown to have good results on scene classification. The main drawback of this approach is the lack of invariance with respect to the layout of the image. In fact, the position of the object to be classified has a strong influence since differently located regions of the image are never compared. Moreover, the size of the signature is multiplied by the number of regions, which makes SPM prohibitive with large dictionaries.

To overcome the size problem, the authors of [5] propose to integrate the spatial coordinates of the descriptors into the codebook by performing Spatial Coordinate Coding (SCC). As a result, the visual words are localized within the image and thus encode some spatial information. However, the same criticism concerning the layout holds as images containing the same object will lead to different representations.

Finally, in [14] the authors propose to use the scores of a set of detectors to compute the signature. In this case, each component of the signature corresponds to the confidence in that a detected region of the image contains the object. Such signatures are indeed invariant to the position of the object in the image, since they consider the image as the result of a semantic composition (*i.e.*, the image is composed of the detected object). However, the regions corresponding to the object are never directly compared, which means that two very dissimilar regions of the same detected object will lead to a high similarity between their respective images.

3. PROPOSED METHOD

Let $\phi : \mathbf{B}_i \rightarrow \mathbf{x}_i$ be the function computing a signature $\mathbf{x}_i$ from the set of descriptors $\mathbf{B}_i$ extracted from image i. Examples of such functions are VLAD [15], VLAT [12], Fisher Vectors [13] or BossaNova [10]. Let now restrain the set $\mathbf{B}_i$ to $\mathbf{B}_{\mathcal{R}i}$ containing only the descriptors whose spatial coordinates belong to a region $\mathcal{R}$ of the image, and call $\phi_{\mathcal{R}} : \mathbf{B}_{\mathcal{R}i} \rightarrow \mathbf{x}_{\mathcal{R}i}$ the corresponding signature function. Let us call $\mathcal{R}$ a *pooling region*.

The pooling region $\mathcal{R}$ can be defined with a static layout as it is the case with SPK. In that case, only the descriptors whose coordinates are within the range defined by $\mathcal{R}$ are considered, independently of the content of the image. As such pooling regions do not encode any semantic information, we propose to use a detection step to select $\mathcal{R}$. We consider a detector providing the bounding boxes corresponding to all region of the image containing the specified concept. Popular examples of such detectors include person detector, face detector or the latent models of [16]. We propose to use only the bounding box corresponding to the maximum score of detection. In the case the specified concept is not present in the image, we propose to use the region with the maximum likelihood of containing the object. Examples of semantic regions are shown on Figure 1.

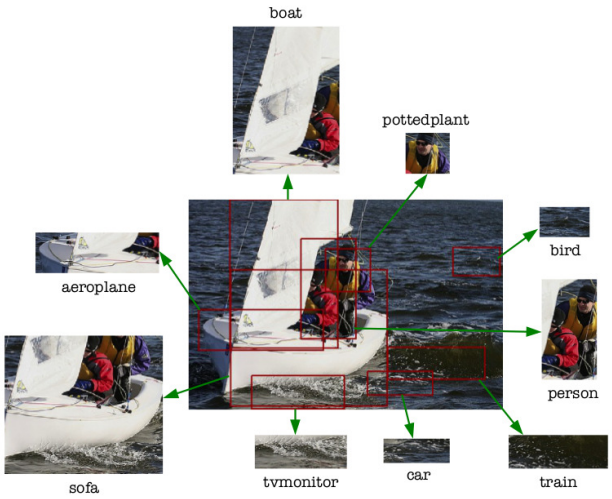

Fig. 1. Examples of semantic regions detected using [16]. Each selected region is corresponding to the maximum output of a specific object detector. For example, the region labeled *aeroplane* corresponds to the region of the image that is the most likely to host an aeroplane. In this example, only the objects *boat* and *person* are in the image.

$\phi_{\mathcal{R}}$ defines an explicit kernel function $k_{\mathcal{R}}(\cdot, \cdot)$ measuring the similarity between any two images i and j, based on signatures computed on the corresponding pooling region $\mathcal{R}$ in i and j:

$$k_{\mathcal{R}}(i, j) = \langle \phi_{\mathcal{R}}(\mathbf{B}_{\mathcal{R}i}), \phi_{\mathcal{R}}(\mathbf{B}_{\mathcal{R}j}) \rangle \tag{3}$$

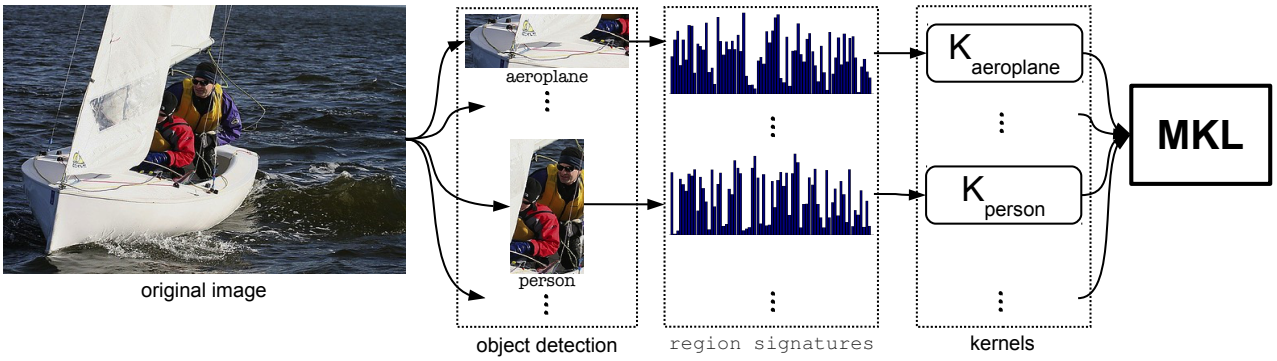

Fig. 2. Overview of our strategy: the 3 main steps are represented. First, the using of detectors in order to get a fixed number of regions. Second, the computation of the region signature using a VLAT strategy. And finally, the use of specific kernels for each region signatures. This representation is then combined with a classifier in order to get the labels for the input image.

When several detectors are available, we propose to define the similarity between two images as a linear combination of the kernel corresponding to the associated pooling regions:

$$k(i,j) = \sum_{\mathcal{R}} \beta_{\mathcal{R}} k_{\mathcal{R}}(i,j), \tag{4}$$

with $\beta_{\mathcal{R}}$ the weights associated with each pooling region $\mathcal{R}$. A baseline approach would be to assign uniform weights, leading to the average kernel. Remark the SPK is a special case of our formalism where the pooling regions are obtained by a regular and recursive grid splitting of the entire image domain.

In order to improve the results, we propose to learn the weights associated with each kernel (*i.e.*, each pooling region) using Multiple Kernel Learning (MKL) [6]. MKL consists in learning jointly the classifier and the kernel combination, by solving the following optimization problem:

$$\min_{\beta} \max_{\alpha} \quad \textstyle\sum_i \alpha_i - \frac{1}{2} \sum_{i,i} \alpha_i \alpha_i y_i y_i \sum_{\mathcal{R}} \beta_{\mathcal{R}} k_{\mathcal{R}}(i,j) \tag{5}$$

$$\text{s.t.} \qquad \forall \mathcal{R}, \beta_{\mathcal{R}} \geq 0 \tag{6}$$

$$\textstyle\sum_{\mathcal{R}} \beta_{\mathcal{R}} = 1 \tag{7}$$

$$\forall i, 0 \leq \alpha_i y_i \leq C \tag{8}$$

Such problem can be solved using off the shelf algorithms like [17]. Remark the ℓ_1 norm constraint on β enforces a sparsity pattern in the kernel weights which is akin to perform kernel selection. The whole process is shown in Figure 2.

In the case of SPK, the kernel selection dos not make a lot of sense unless the dataset has a bias in the position of some objects (*i.e.*, cars are always on the bottom of the images). However, with our semantic pooling regions, we argue the kernel selection is able to filter out the objects which are uncorrelated with the considered category. Since many detectors are irrelevant for a given category, the kernel selection allows to get rid of the corresponding noisy features. This has the mechanical effect that the global similarity is computed mainly with signatures corresponding to regions of the image whose contents are highly correlated with the considered category. Moreover, the semantic pooling regions and the associated kernel selection allows invariance with respect to the localization of relevant objects within the image, which was not the case with SPK or SCC. Finally, although MKL introduces an overhead in computational complexity during the learning step, it produces less pooling regions than SPK and thus leads to less computational complexity in inference.

4. EXPERIMENTS

We evaluated our method on the well known VOC 2007 dataset [7], consisting of about 10k images with 20 classes. To compute the signatures, we used the method proposed in [12] with HOG descriptors sampled every 3 pixels at 4 scales. The visual codebook was set to 64 visual words, which led to high dimensional signatures. To ease the computation, we compressed the VLAT using the method of [19]. The MKL was trained using JKernelMachines [20].

For the spatial pooling, we performed a baseline SPM with the $1 \times 1, 2 \times 2, 3 \times 1$ configuration (which we denote p thereafter). For our semantic pooling, we use the detectors of [16] trained on the *train* set only, leading to 20 pooling regions.

We show the results in Table 2. The Baseline VLAT without spatial information performs reasonably well at 57.9% of mAP. Adding spatial information by concatenating the signatures obtained by the different regions leads to a small improvement both for the pyramid pooling (pVLAT: 59.0% of mAP) and the semantic pooling (sVLAT: 58.4% of mAP). Using MKL to learn the combination of the pooling regions for the pyramid also leads to slight improvements (pMKL: 59.7% of mAP). On the contrary, using MKL to select the semantic pooling regions (sMKL) leads to a significant improvement at 63.2% of mAP. When combining both pyramid and seman-

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	
spMKL	78.0	73.5	45.6	70.0	45.2	71.1	83.0	62.5	55.2	47.8	
FV [1]	79.0	67.4	51.9	70.9	30.8	72.2	79.9	61.4	56.0	49.6	
VLAT [18]	80.3	72.2	51.4	71.4	28.1	72.1	81.6	63.1	54.4	47.5	
Detector [14]	64.7	75.6	32.1	62.7	48.2	70.3	83.8	49.3	57.2	48.4	
Det+BNFV [14]	80.8	78.8	55.4	73.8	52.2	76.6	86.4	64.1	62.1	55.2	
	d-table	dog	horse	m-bike	person	p-plant	sheep	sofa	train	tv	mAP
spMKL	62.5	47.5	79.7	71.5	86.7	41.6	51.4	61.9	81.2	63.7	64.0
FV [1]	58.4	44.8	78.8	70.8	85.0	31.7	51.0	56.4	80.2	57.5	61.7
VLAT [18]	57.8	46.5	81.1	70.3	86.8	30.8	41.2	54.0	84.1	54.9	61.5
Detectors [14]	52.4	34.2	76.9	68.1	87.7	36.6	44.4	51.8	71.3	63.8	59.0
Det+BNFV [14]	66.6	49.7	83.4	74.7	89.8	37.9	50.7	64.1	80.9	68.9	67.6

Table 1. Detailed results on VOC 2007 and comparison with recent methods, in terms of average precision.

tic regions to the selection performed by MKL (spMKL), we obtain a mAP of 64.0%.

	VLAT	pVLAT	sVLAT
mAP ($\%$)	57.9	59.0	58.4
	spMKL	pMKL	sMKL
mAP($\%$)	**64.0**	59.7	63.2

Table 2. Results on VOC 2007 in mean Average Precision (mAP).

Table 1 shows the comparison of the approach with recent results from the literature, and shows that our method performs often better than existing systems that do not take into account the spatial information. This is especially the case for difficult categories like *bottle* or *potted-plant*, for which the improvement is very significant. Compared to the method of [14] also based on detectors, our method is able to perform much better than the detectors alone, although it does not reach the results obtained by combining detectors with Fisher Vectors and Bossa Nova signatures.

In order to analyze the region selection, we show in Figure 3 the weights associated with each semantic region obtained by the MKL. Each row corresponds to the category being recognized, while the columns stand for the weights of the corresponding pooling region. For most of the categories, the full image (all) and the considered category pooling region obtain most of the weight. Some categories obtain significant weights in pooling regions that seem consistent from the semantic point of view (*e.g.*, pooling *dinningtable* for the categories *bottle* and *chair*).

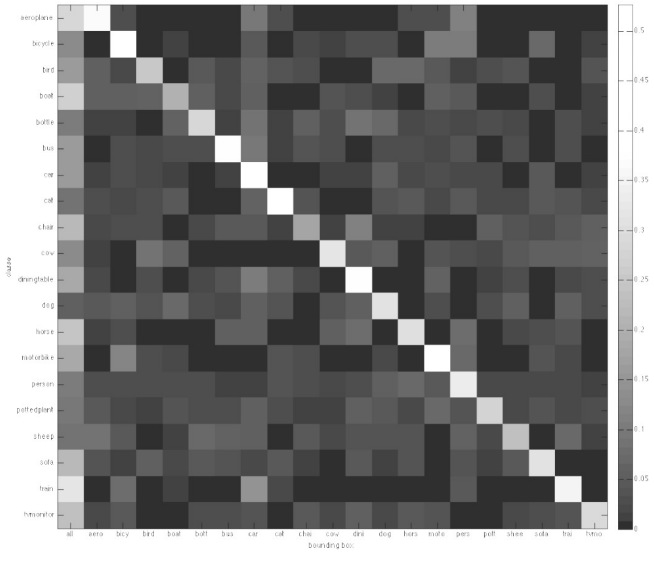

Fig. 3. MKL weights for each pooling region with respect to the learned category. Each row corresponds to the category being classified, while the columns are the pooling regions. The first column corresponds to the whole image.

5. CONCLUSION

In this paper, we proposed a new image categorization system based on a semantic pooling. Contrary to most image categorization systems, our proposal takes into account the layout of the images and allows to compare different regions of the images based on their semantic content. In order to meet this goal, our method uses regions extracted by a set of object detectors and computes well known signatures inside these regions.

In addition, to select the relevant detectors with respect to a specific category, we also proposed a kernel framework that allows the use of Multiple Kernel Learning algorithms. We made experiments on the well known VOC 2007 dataset and showed the soundness of the approach.

6. REFERENCES

[1] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent

feature encoding methods," 2011, vol. 76, pp. 1–12.

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

[3] David Picard and Philippe-Henri Gosselin, "Improving image similarity with vectors of locally aggregated tensors," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 665–669.

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

[5] Piotr Koniusz and Krystian Mikolajczyk, "Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match," 2011.

[6] Francis R. Bach and Gert R. G. Lanckriet, "Multiple kernel learning, conic duality, and the smo algorithm," 2004.

[7] 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/workshop/index.html.

[8] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005, vol. 2, pp. 886–893.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," vol. 2, no. 60, pp. 91–110, 2004.

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

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

[12] D. Picard and P.H. Gosselin, "Efficient image signatures and similarities using tensor products of local descriptors," vol. 117, pp. 680–687, 2013.

[13] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," 2010, pp. 143–156.

[14] Thibaut Durand, Nicolas Thome, Matthieu Cord, and Sandra Avila, "Image classification using object detectors," in *20th IEEE International Conference on Image Processing (ICIP)*, September 2013.

[15] Hervé Jégou, Florent Perronnin, Matthijs Douze, Cordelia Schmid, et al., "Aggregating local image descriptors into compact codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1704–1716, 2012.

[16] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[17] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, Yves Grandvalet, et al., "Simplemkl," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[18] R. Negrel, D. Picard, and P.H. Gosselin, "Using spatial pyramids with compacted vlat for image categorization," Tsukuba Science City, Japan, November 2012.

[19] R. Negrel, D. Picard, and P.H. Gosselin, "Web scale image retrieval using compact tensor aggregation of visual descriptors," vol. 20, no. 3, pp. 24–33, 2013.

[20] David Picard, Nicolas Thome, and Matthieu Cord, "Jkernelmachines: A simple framework for kernel machines," *Journal of Machine Learning Research*, vol. 14, no. May, pp. 1417–1421, 2013.