Modeling the Spatial Layout of Images Beyond Spatial Pyramids

Jorge Sánchez[a,1], Florent Perronnin[b], Teófilo de Campos[c,*]

[a]CIEM-CONICET, FaMAF, Universidad Nacional de Córdoba,
X5000HUA, Córdoba, Argentine, Tel: +54 351 4334051 int. 309
[b]Xerox Research Centre Europe, 6 Chemin de Maupertuis,
38240 Meylan, France, Tel: +33 476 61 50 17, Fax: +33 476 61 50 99
[c]CVSSP, University of Surrey, Guildford, GU2 7XH, UK,
Tel: +44 1483 686032, Fax: +44 1483 300803

Abstract

Several state-of-the-art image representations consist in averaging local statistics computed from patch-level descriptors. It has been shown by Boureau *et al*. that such average statistics suffer from two sources of variance. The first one comes from the fact that a finite set of local statistics are averaged. The second one is due to the variation in the proportion of object-dependent information between different images of the same class. For the problem of object classification, these sources of variance affect negatively the accuracy since they increase the overlap between class-conditional probabilities.

Our goal is to include information about the spatial layout of images in image signatures based on average statistics. We show that the traditional approach to including the spatial layout – the Spatial Pyramid (SP) – increases the first source

[*] Corresponding author

Email addresses: jsanchez@famaf.unc.edu.ar (Jorge Sánchez), florent.perronnin@xrce.xerox.com (Florent Perronnin), t.decampos@st-annes.oxon.org (Teófilo de Campos)

[1] Most of this work was done while J. Sánchez was at CIII, Universidad Tecnológica Nacional, Factultad Regional Córdoba, X5000HUA, Córdoba, Argentine.

of variance while only weakly reducing the second one. We therefore propose two complementary approaches to account for the spatial layout which are compatible with our goal of variance reduction. The first one models the spatial layout in an image-independent manner (as is the case of the SP) while the second one adapts to the image content. A significant benefit of these approaches with respect to the SP is that they do not incur an increase of the image signature dimensionality. We show on PASCAL VOC 2007, 2008 and 2009 the benefits of our approach.

Keywords: image representation, spatial layout, image categorization, Fisher vectors, PASCAL VOC datasets, spatial pyramids

1. Introduction

One of the most successful approaches to describe the content of images is the bag-of-features (BOF). It consists in computing and aggregating statistics derived from local patch descriptors such as the SIFT [1]. The most popular variant of the BOF framework is certainly the bag-of-visual-words (BOV) which characterizes an image as a histogram of quantized local descriptors [2, 3]. In a nutshell, a codebook of prototypical descriptors is learned with k-means and each local descriptor is assigned to its closest centroid. These counts are then averaged over the image.

The BOV has been extended in several ways. For instance, the hard quantization can be replaced by a soft quantization to model the assignment uncertainty [4, 5] or by other coding strategies such as sparse coding [6, 7, 8]. Also the average pooling can be replaced by a max pooling [6, 7, 8, 9]. Another extension is to include higher order statistics. Indeed, while the BOV is only concerned with the number of descriptors assigned to each codeword, the Fisher Vector (FV) [10, 11] as well as the related Vector of Locally Aggregated Descriptors (VLAD) [12] and Super-Vector Coding (SVC) [13] also model the distribution of descriptors assigned to each codeword.

Obviously discarding all information about the location of patches incurs a loss of information. The dominant approach to include spatial information in the BOF framework is the Spatial Pyramid (SP). Inspired by the pyramid match kernel of Grauman and Darrell [14], Lazebnik *et al*. proposed to partition an image into a set of regions in a coarse-to-fine manner [15]. Each region is described independently and the region-level histograms are then concatenated into an image-level histogram. The SP enables to account for the fact that different regions can contain different visual information.

Several extensions of the SP have been proposed. Marszalek *et al*. suggested a different partitioning strategy [16]. Their system combined the full image with a 1x3 (top, middle and bottom) and a 2x2 (four quadrants) partitioning. Viitaniemi and Laaksonen proposed to assign patches to multiple regions in a soft manner [17]. The SP has also been extended beyond the BOV, for instance to the FV [11] or the SVC [13]. We note that all previous methods rely on a pre-defined partitioning of the image which is independent of its content. Uijlings *et al*. proposed a bi-partite image-dependent partioning in terms of object/non-object [18]. Two BOV histograms are computed per image: an object BOV and a context BOV. While the authors report a very significant increase of the classification accuracy on PASCAL VOC 2007, their method relies on the knowledge of the object bounding boxes which is unrealistic for most scenarios of practical value. We outline that the simple SP of Lazebnik *et al*. is still by far the most prevalent approach to account for spatial information in BOF-based methods. Recently, Krapac *et al*. [29] proposed to include a location prior per visual word and to derive a Fisher kernel from this model. They report similar results as with SP but using a more compact representation. In [30], the authors propose to include spatial and angular information directly at descriptor level. They used soft-BOV and sparse coding-based signatures, reporting promising results compared to SP[2].

Our goal is to propose alternatives to the SP for object classification. We focus on the FV which is simple to implement, computationally efficient and which was shown to yield excellent results in a recent evaluation [31]. However, our work could be extended to other BOF-based techniques in a straightforward manner.

We build on the insights of Boureau et al. [8, 9]. If we have a two-class classification problem, linear classification requires the distributions of FVs for these two classes to be well-separated. However, there are two sources of variance which make the distributions of FVs overlap. The first one is due to the fact that the FV is computed from a finite set of descriptors. The second one comes from the fact that the proportion of object-dependent information may vary between two images of the same class. Reducing these sources of variance would increase the linear separability and therefore the classification accuracy. In this paper, we propose two different and complementary ways to include the spatial information into the image signature which target these two sources of variance.

The remainder of the article is organized as follows. In the next section, we briefly review the FV coding method. In section 3 we consider the variance due to the finite sampling of descriptors. We extend the analysis of [8, 9] to the case of correlated samples. We show that, because the SP reduces the size of the re-

[2] Some of our contributions are related to those of [29, 30], which have been developed in parallel to the work in this paper. As it will be clear in sec 5 our results in the VOC2007 dataset outperform theirs by a large margin.

gion over which statistics are averaged, it impacts negatively the variance of the distribution of FVs. We therefore propose a novel approach to include the spatial information by augmenting the descriptors with their location. In section 4 we analyze the second source of variance specifically in the case of the FV. We show that we could partially compensate for this source of variance if we had access to the object bounding boxes. However, as opposed to [18] we propose a practical solution to this problem based on the objectness measure of Alexe *et al*. [19]. In section 5, we provide experimental results on PASCAL VOC 2007, 2008 and 2009 showing the validity and the complementarity of the two proposed techniques. A major benefit is that, as opposed to the SP, they do not increase the feature dimensionality thus making the classifier learning more efficient.

2. The Fisher Vector

We only provide a brief introduction to the FV coding method. More details can be found in [10, 11]. Let $X = \{x_t, t = 1 \ldots T\}$ be the set of T local descriptors extracted from an image. Let $u_\lambda : \mathbb{R}^D \rightarrow \mathbb{R}_+$ be a probability density function with parameters λ which models the generation process of the local descriptors for any image. The Fisher vector $\mathcal{G}_\lambda^X$ is defined as:

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X. \tag{1}$$

L_λ is the Cholesky decomposition of the inverse of the Fisher information matrix F_λ of u_λ, *i.e.* $F_\lambda^{-1} = L'_\lambda L_\lambda$. G_λ^X denotes the gradient of the log-likelihood w.r.t. λ, *i.e.*:

$$G_\lambda^X = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t). \tag{2}$$

In our case $u_\lambda = \sum_{i=1}^{N} w_i u_i$ is a GMM with diagonal covariance matrices and parameters $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \dots N\}$ where w_i, μ_i and σ_i are respectively the mixture weight, mean vector and standard deviation vector of Gaussian u_i. Let $\gamma_t(i)$ be the soft assignment of descriptor x_t to Gaussian u_i. Following [10, 11] we discard the partial derivatives with respect to the mixture weights as they carry little discriminative information. We obtain the following formulas for the gradients with respect to μ_i and σ_i:[3]

$$\mathcal{G}_{\mu_i}^X = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right), \tag{3}$$

$$\mathcal{G}_{\sigma_i}^X = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]. \tag{4}$$

The image signature is defined as the concatenation of the vectors (3) and (4) for all Gaussians:

$$\mathcal{G}_\lambda^X = \left[\mathcal{G}_{\mu_1}^X, \cdots, \mathcal{G}_{\mu_N}^X, \mathcal{G}_{\sigma_1}^X, \cdots, \mathcal{G}_{\sigma_N}^X \right]^T. \tag{5}$$

As shown in [11], square-rooting and L2-normalizing the FV can greatly enhance the classification accuracy. Also, following the SP framework, one can split an image into several regions, compute one FV per region and concatenate the per-region FVs.

Let D be the dimensionality of the local descriptors, N be the number of Gaussians and R be the number of image regions. The resulting vector is $E = 2DNR$ dimensional.

[3] Vector divisions should be understood as term-by-term operations.

3. Average Pooling and Feature Augmentation

The FV, as given by eq (1) and (2), can be viewed as an average of patch-level statistics. Indeed, we can rewrite:

$$\mathcal{G}_\lambda^X = \frac{1}{T} \sum_{t=1}^{T} z_t \tag{6}$$

with:

$$z_t \equiv g(x_t) \equiv L_\lambda \nabla_\lambda \log u_\lambda(x_t). \tag{7}$$

If we assume the samples x_t to have been generated by a class-conditional distribution p_c (where the variable c indexes the class) and to be iid, then (6) can be seen as the sample estimate of a class-conditional expectation:

$$\lim_{T \to \infty} \mathcal{G}_\lambda^X = \mathbb{E}_{x \sim p_c}[g(x)]. \tag{8}$$

As noted in [8], there is an intrinsic variance in this estimation process which is caused by sampling from a finite pool of descriptors. Boureau *et al*. make a patch independence assumption and thus, in their analysis, the variance of this estimator decreases like $\frac{1}{T}$.

In the rest of the section we extend this variance analysis by relaxing the independence assumption. Although our focus is on FVs, the analysis we present applies to the broader class of image descriptors which average statistics computed from local descriptors. We also outline the shortcomings of the SP framework in the light of the previous analysis. Our conclusion is that partitioning the image into a set of regions increases the variance of the estimator. We finally present a new image representation which encodes the spatial layout while alleviating the partitioning.

3.1. Variance analysis of average pooling

In what follows we assume that image patches are extracted from the nodes of a regular grid[4] and described by D-dimensional vectors, *e.g.* SIFT [1] descriptors. To facilitate the analysis, let us consider a simplified model where all variables z_t have equal variances *i.e.* $\mathrm{Var}(z_t) = \sigma^2$. The variance of the sample mean estimator is, in this case:

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} z_t\right) = \frac{\sigma^2}{T} + \frac{\sigma^2}{T^2}\sum_{t=1}^{T}\sum_{\substack{s=1\\ s\neq t}}^{T} \rho(z_t, z_s) \tag{9}$$

with $\rho(z_t, z_s)$ the correlation coefficient between variables z_t and z_s. If we define the average of these cross-term correlations as $\bar{\rho} = \frac{1}{T(T-1)} \sum_{t=1}^{T} \sum_{s=1, s\neq t}^{T} \rho(z_t, z_s)$, eq. (9) can be rewritten as

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} z_t\right) = \frac{\sigma^2}{T} + \sigma^2 \frac{T-1}{T}\bar{\rho}. \tag{10}$$

Note that the value $\bar{\rho}$ is a function of several factors including the sampling step or the size of the pooling window. We now analyze the impact of these two factors[5]. Figure 1 shows estimates of the average cross-term correlation $\bar{\rho}$ as a function of the grid sampling step for two pooling window sizes: 128×128 and 96×96 pixels respectively (see sec. 5.2 for details about the feature extraction procedure). As expected, $\bar{\rho}$ increases when the sampling step or the window size decrease.

We now study the implications on the SP framework. Based on the previous analysis, we can see that partitioning the image incurs an increase in the variance

[4] Other sampling strategies can be analyzed as well, *e.g.*, random sampling or sampling based on interest point/region detection. The conclusions that follow remain the same.

[5] We note that $\bar{\rho}$ might depend on many other factors including the semantic content of the image.

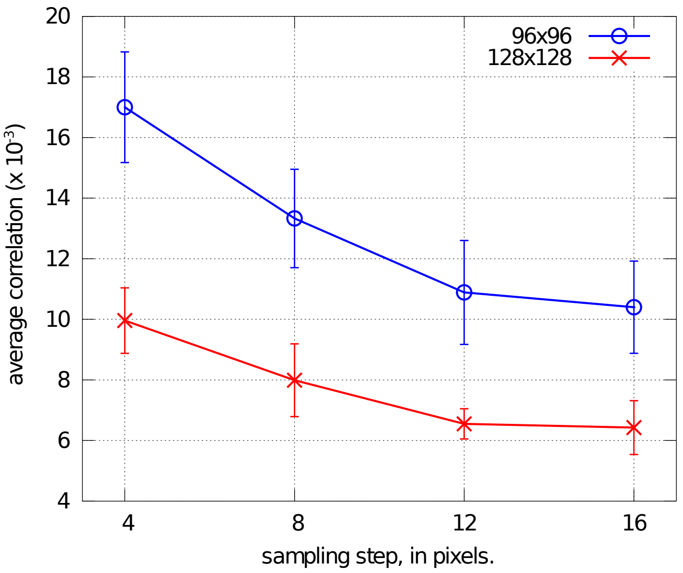

Figure 1: Average correlation $\bar{\rho}$ as a function of the sampling step for pooling windows of 128x128 and 96x96 pixels. This analysis was performed on the *train* set of the PASCAL VOC 2007 dataset.

of the estimator when compared to the case where the patch-level statistics are pooled over the whole image. Indeed, *for a fixed sampling step*, when the size of the pooling region decreases, we have the two following effects: i) the number of patches T decreases and ii) the average correlation $\bar{\rho}$ increases.

We would like to point out that using as many patches as possible (e.g. by sampling patches at each pixel location) might not be optimal for the average pooling strategy contrary to what is claimed in [9]. Indeed, on one hand decreasing the step size will increase the sample cardinality, as desired. On the other hand, increasing T will also increase the patch overlap and thereby the average correlation. From (10) in the limit:

$$\lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} z_t\right) = \sigma^2\bar{\rho}. \tag{11}$$

Therefore, the benefits brought by a greater sample cardinality might be compensated by an increase of $\bar{\rho}$.

3.2. Feature augmentation

We now propose to model the layout of an image without partitioning it. We consider the joint distribution of low-level descriptors *and* patch locations. As we will see, our approach results in a very simple solution that competes favorably in performance with SPs.

Let $m_t = [m_{x,t}, m_{y,t}]^T$ denote the 2D-coordinates of an image patch associated to a low-level descriptor x_t and σ_t the patch scale. Let H and W represent the image height and width respectively. We define the following augmented feature vector $\hat{x}_t \in \mathbb{R}^{D+3}$:

$$\hat{x}_t = \begin{pmatrix} x_t \\ m_{x,t}/W - 0.5 \\ m_{y,t}/H - 0.5 \\ \log \sigma_t - \log \sqrt{WH} \end{pmatrix}. \qquad (12)$$

By using (12) instead of the raw descriptors, the underlying distribution u_λ now reflects not only the generation process of local descriptors but also the location and scale at which they are likely to be generated.

This augmented representation offers several benefits with respect to the SP. First, it does not rely on a partitioning of the image and therefore does not lead to an increase of the variance. Second, it leads to only a very small increase in the dimensionality of the FV: $2N(D+3)$ dimensions compared to $2DNR$ dimensions for a SP with R regions[6]. This makes the learning of classifiers significantly more

[6] Actually in our experiments with augmented features we keep the feature dimensionality con-

efficient and helps scaling to larger datasets. Third, it does not require to choose, a priori, a given spatial layout. Indeed, the optimal layout of a SP may depend on the dataset.

Note that, as we consider diagonal covariances for the generative model of eq. (2), the components of the mixture (single Gaussians) can be decomposed as $u_i = u_i^{(app)} u_i^{(loc)}$. Here, $u_i^{(app)}$ and $u_i^{(loc)}$ denote the appearance and location/scale part of the augmented representation, respectively. This is equivalent to explicitly including a (Gaussian) location prior per visual word, as proposed by Krapac *et al*. (*c.f*. eq. (18) of [29]). In our case, the model remains the same and we only change the low level feature representation, making it possible to extend the model to other encoding methods.

4. Within-Class Variance and Objectness

In the previous section, we showed that the FV can be understood as the sample estimate of a class-conditional expectation and that there is an intrinsic variance in this estimation process which is caused by sampling from a finite pool of descriptors. We now show that there is a second source of variance which is inherent to the model and we propose another approach to take into account the spatial layout to remediate this issue.

4.1. Within-class variance

We follow the same line of thought as Boureau *et al*. [8] and Perronnin *et al*. [11] and assume that the local descriptors in a given image of class c are generated by a mixture of two distributions: a class-dependent distribution q_c and a

stant by selecting a subset of $(D-3)$ original features (c.f. sec. 5.1).

background class-independent distribution. Furthermore, as is the case in [11], we make the assumption that the class-independent distribution can be approximated by u_λ. Therefore, the generative model of patches in an image of class c can be written as:

$$p_c(x) = \omega q_c(x) + (1-\omega)u_\lambda(x) \tag{13}$$

with $0 \leq \omega \leq 1$ reflecting the proportion of class-specific information. As shown in [11], if the parameters characterizing the background distribution u_λ were estimated to maximize (at least locally) the likelihood function, then we have approximately:

$$\lim_{T\to\infty} G_\lambda^X = \omega \nabla_\lambda \mathbb{E}_{x\sim q_c}[\log u_\lambda(x)] \tag{14}$$

and consequently we can rewrite (8) as follows:

$$\lim_{T\to\infty} \mathcal{G}_\lambda^X = \omega \mathbb{E}_{x\sim q_c}[g(x)]. \tag{15}$$

Following [8], we further assume that ω is drawn form a distribution (*e.g.* a beta distribution) and that, while it may vary from one image to another, it is sampled only once per image. We underline that the distribution from which ω is sampled might be class-dependent (c.f. Figure 2). In such a case, the quantity $\omega\mathbb{E}_{x\sim q_c}[g(x)]$ is a random variable. Therefore, even if we had access to an infinite number of iid patches T in each image (perfect estimation of the class-conditional expectation) there would be some variance *between images* as we have:

$$\mathrm{Var}\left(\lim_{T\to\infty} \mathcal{G}_\lambda^X\right) = \mathrm{Var}(\omega)\left(\mathbb{E}_{x\sim q_c}[g(x)]\right)^2 \tag{16}$$

where the variance has been taken with respect to ω. Therefore, we can decrease the variance, and therefore increase the class separability, by cancelling the effect of ω. We propose an approximate method to do so in the next subsection.

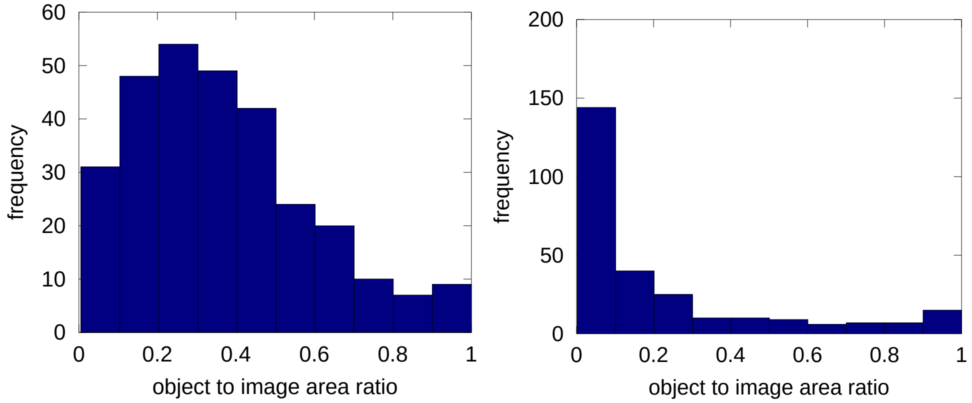

Figure 2: We show the histogram of the ω values for two VOC 2007 classes: horse (left) and potted-plant (right). Since we do not have access to omega directly, we use as a proxy the ratio between the object bounding-box area and the image area.

4.2. Leveraging the objectness measure

Let us use as a proxy for ω the proportion of the image which is covered by a given object. We note that if we worked only with images of cropped objects then we would have $\omega \approx 1$ and the variance effect described in the previous section would be canceled out. Uijlings *et al.* indeed showed that the recognition accuracy of a BOV-based image classifier could be greatly increased by assuming the knowledge of the object locations in images [18]. However, in their scenario, the object bounding boxes were provided manually which is unrealistic for most applications of practical value[7]. The above has also been observed by De Campos *et al.* [20], who explored the use of human feedback to provide the approx-

[7] We note that the SP could somewhat compensate for this source of variance for a given class if the location of the considered object was fixed and matched a given region of the SP. However, such stringent conditions would rarely hold in practice.

imate location of objects in images (given in terms of “soft” bounding boxes). The authors showed significant improvements compared to other alternatives, *e.g.* methods based on the “saliency” of local image patches.

Recently, Alexe *et al*. [19] proposed a method to measure how likely an image window is to contain an object of any class. The method relies on the combination of different cues designed to reflect generic properties of objects, *i.e.* global saliency, local contrast and boundary closeness. This measure, used as a prior over object locations was successfully employed to speed-up object detectors [19]. We propose to use this objectness measure to approximately estimate the location of objects in images. More precisely, we combine the objectness measure of [19] with the *locally-weighted patches* approach of De Campos *et al*. [20]. In the weighted-patches representation, we have a weight ϕ_t associated with each descriptor x_t and we have the following weighted representation of the image:

$$\tilde{\mathcal{G}}_\lambda^X = \frac{\sum_{t=1}^T \phi_t z_t}{\sum_{t=1}^T \phi_t}. \tag{17}$$

In our case, the weights ϕ_t are computed as follows. For a given image, we draw a set of windows from the objectness distribution with the sampling procedure described in [19]. Let $\Omega_j, j = 1, \ldots, M$ represent the spatial support of the j-th window (defined by *e.g.* its top-left and bottom-right corners). The weight ϕ_t for the descriptor x_t located at position m_t is computed as follows:

$$\phi_t = \sum_{j=1}^{M} \delta_j(m_t) \tag{18}$$

where $\delta_j(m_t)$ has been defined as:

$$\delta_j(m_t) = \begin{cases} 1 & \text{if } m_t \in \Omega_l, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

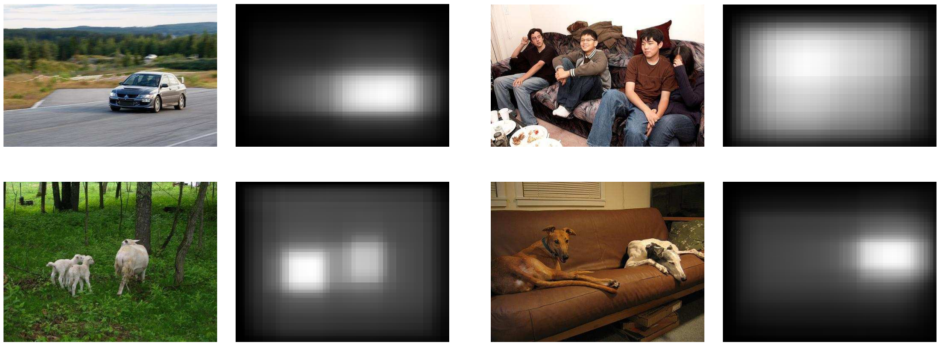

Figure 3: Sample images from the PASCAL VOC 2007 dataset and the corresponding objectness maps obtained by sampling 1000 random windows.

Figure 3 shows the objectness maps obtained by this procedure for some example images of the PACAL VOC 2007 dataset [21]. Note that, obviously, even if the objectness measure of [19] provided a perfect prediction of the presence/absence of an object, the proposed approach would only partially cancel the effect of ω for several reasons. First, some object patches might have been emitted by the background distribution, *e.g*. the uniform patches of an untextured object. Second, some background patches might have been emitted by the class-specific distribution, *e.g*. when the background strongly correlates with the presence of the object. Third, realistic images contain multiple objects and the objectness measure does not distinguish between different objects. Therefore, multiple objects might contribute to the weighted image signature.

We note that Perronnin *et al*. [11] proposed the L2 normalization of the FV to cancel the effect of ω. In our experiments, we always found that the combination of the L2 normalization and the objectness measure improved classification which seems to indicate that there is a complementarity between these two approaches.

We also note that Uijlings *et al*. partitioned the image into object/background and consequently computed two representations per image: one for the object and one for the background. The two representations were subsequently combined. We also tried to compute two FVs: one using the objectness measure and one using the complement to focus on context information. We observed experimentally that adding the context information had little impact in our case. This might be because the FV weighted by the objectness measure already contains a fair amount of background information (c.f. Figure 3). Therefore, we discarded the context FV. Consequently, using the objectness measure does not increase the dimensionality of the FV representation.

We note that using the objectness measure to compute patch weights can be regarded as a saliency estimation process. However, traditional approaches for saliency detection (i.e. bottom-up methods) rely on scoring small regions according to their rarity w.r.t. to their local surroundings. As such, salient regions detectors show difficulties in dealing with cluttered or textured backgrounds (as observed, e.g. in [28]). Although the method of Alexe *et al*. includes a multi-scale saliency detector as a basic cue, it also considers other measures related to the presence of whole objects besides of simple local characteristics.

It has been observed that highlighting whole objects may not always be best strategy. For instance, if the goal is to distinguish between cats and dogs, it is better to highlight their heads than give equal importance to their whole body [20, 32]. In that context, it is possible that novel top-down saliency estimators may lead to better performance with the proposed representation. Such an evaluation is a suggestion for future work.

Finally we point out that, in the case of the BOV representation, the max-pooling strategy was shown to be more resilient to the variance of ω than the average pooling strategy. However, extending the max-pooling strategy to the FV – *i.e.* beyond count statistics – is non-obvious in our opinion and would be an interesting topic of future research.

5. Experiments

We first present the experimental setup. We then provide more details about the computation of the average correlation in sec 3. We finally report our results.

5.1. Experimental setup

Datasets. We ran experiments on three challenging datasets: PASCAL VOC 2007 [21], 2008 [22] and 2009 [23]. These datasets contain images of 20 object categories: *aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorbike, person, pottedplant, sheep, sofa, train* and *tv-monitor*. The set of images for each class exhibits a large degree of intra-class variation, including changes in viewpoint, illumination, scale, partial occlusions, etc.. Images from these datasets are split into three groups: *train* for training, *val* for validation and *test* for testing. We followed the recommended procedure of selecting parameters by training on the *train* while using the *val* set for testing. The system was re-trained using the *train+val* sets once the best choice for the parameters have been selected. Classification performance is measured using the mean Average Precision (mAP).

Low-level features. In all our experiments we used *only* 128-dimensional SIFT descriptors, computed over image patches of 32×32 pixels uniformly distributed over the image, *i.e.* extracted from the nodes of a regular grid with a step size of 8 pixels (we used the "flat" implementation of [24]). We did not perform any normalization on the image patches before computing SIFT descriptors. The dimensionality of these descriptors were further reduced to 80 by Principal Components Analysis (PCA). To account for variations in scale, we extracted patches at 7 levels with a scale factor of $\sqrt{2}$ between them. Images were first upsampled at twice their original resolution as in [1].

Feature augmentation. For the experiments based on the feature augmentation approach (sec. 3.2) we kept the same dimensionality of low level features by replacing the 3 "least-significant" dimensions of the PCA-reduced SIFT with the 3 location and scale dimensions. This ensures a fair comparison with the original 80-dimensional PCA features.

Objectness measure. To compute the objectness measure, we used the default pre-trained system provided by the authors of [19]. We sampled 1,000 windows per image.

Generative model. We trained a GMM with the Maximum Likelihood (ML) criterion using the Expectation-Maximization (EM) algorithm. We used 1M random samples from the training set and the EM algorithm initialized by running standard k-means and using the statistics of points assigned to each Voronoi partition (relative count, mean and variance vectors) as initial estimates for the mixing coefficients, mean and variance vector respectively.

Classifiers. We learnt a linear Support Vector Machine (SVM) independently for each class (one-vs-all classification) using Stochastic Gradient Descent (SGD) in the primal. We used the code made available by Bottou [25].

5.2. Estimation of the sample correlation $\bar{\rho}$

We now give a detailed explanation of the estimation procedure outlined in section 3. We generated a set of 100 fixed-size images by randomly sampling windows (of 128×128 and 96×96 pixels respectively) from the *train* set of the

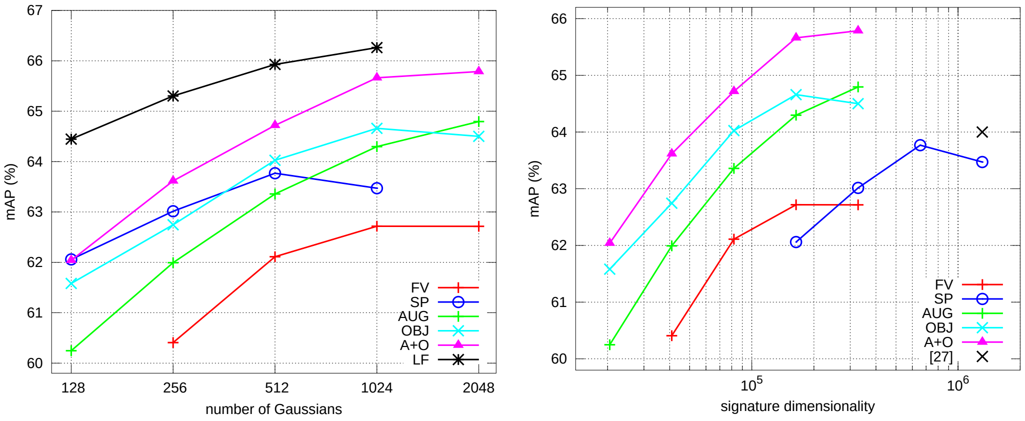

Figure 4: Classification performance vs number of Gaussians (top) and vs the image signature dimensionality (bottom) on VOC 2007.

PASCAL VOC 2007 dataset. For each such window we extracted SIFT descriptors as described above but considering only one level and no up-sampling was performed. We computed a “single-feature” FV for each extracted sample by using a model with 128 Gaussians. We did not perform any further normalization, neither L_2 nor square rooting. We repeated the experiment 5 times with a different subset on each run. In Figure 1, we show the mean over all runs and error bars at 1 standard deviation.

5.3. Results

VOC 2007. Figure 4 (top) shows the classification performance as a function of the number of Gaussians (from 128 to 2,048) on the PASCAL VOC 2007 dataset. We consider two baseline systems: one which does not include any kind of spatial information (FV) and one based on a FV with a 1x1+2x2+1x3 partitioning (SP). Note that the signatures of the SP system are 8 times larger than those of the FV for the same number of Gaussians. Compared to the state-of-the-art, our SP system achieves a performance comparable to the best published results for systems using only SIFT descriptors (63.8% vs 64.0% of Zhou *et al*. [13]).

We also evaluate the following systems: a FV system based on feature augmentation (AUG), a FV system employing the objectness prior on top of non-augmented PCA-reduced vectors (OBJ) and a FV system based on the combination of both of the above (A+O), *i.e*. by using the objectness measure to weight the contribution of augmented low-level features. We also evaluate a system based on a late-fusion approach (LF): averaging the outputs of the classifiers from the A+O and SP systems.

Let us first compare our two baseline systems: FV and SP. It can be seen that, besides the notorious benefit of including the spatial information into the representation, these two systems behave differently as the size of the vocabulary increases. In the case of FV, it reaches a plateau at 1024 Gaussians while SP does reach a maximum at 512. We can explain this behavior by noting that the variance of the FV depends not only on the number of patches but also on the number Gaussians, since the larger the number of Gaussians the fewer "per Gaussian" statistics are pooled together (higher sparsity). This also applies to other image-level representations and especially to the BOV as noted in [8, 9]. Therefore, *by partitioning the image we are not only reducing the number of samples contributing to the representation but also limiting the capacity of the system to benefit from richer vocabularies*.

Let us now consider the performance of the proposed systems (AUG and OBJ) alone. In both cases, we observe a consistent improvement w.r.t. the FV-baseline for all vocabulary sizes. Compared to SP, the OBJ system shows a slightly worse performance for models having up to 512 visual words. It reaches its maximum accuracy of 64.7% mAP at N=1,024 and outperforms our best SP system while using 4 times smaller signatures. We observe a similar behavior on the AUG system. It shows a lower performance for small vocabularies but the gain brought by using a more complex model becomes even more pronounced. It reaches a value of 64.8% mAP at 2,048 Gaussians and, contrary to the SP and OBJ systems, it does no show a decrease in classification performance with larger vocabularies.

If we now consider the combination of the two, *i.e*. our A+O system, we observe some complementarity between these two approaches: while the augmentation approach models the location information in an *image-independent* manner, the objectness prior *adapts to the image content.* The combined system achieves 65.8% mAP. This is +2.0% better than our SP baseline.

Finally, let us consider the system obtained by averaging the outputs of the SP and A+O classifiers. Note the great complementarity that exists between the system for small vocabularies. The combined system achieves a state-of-the-art accuracy of 64.4% mAP with barely $N = 128$ Gaussians. For larger values of N the effect becomes less noticeable: +2.4% absolute points (+3.8% relative) at 128 Gaussians vs. +0.6% (+0.9%) at 1,024 Gaussians.

We also show in Figure 4 (bottom) the classification accuracy as a function of the dimensionality of the image signatures. When compared to the SP baseline or to [13] the advantages of our representation are clear: we can achieve the same accuracy with much smaller dimensional image representations. Again, this is an important advantage when scaling to large datasets.

Table 1 shows the classification accuracy obtained for the best of each system in figure 4. We also compare with the supper vector coding (SVC) approach of Zhou *et al*. [13].

Table 1: Classification performance for each class of the PASCAL VOC 2007 dataset for the systems shown in figure 4.

Class	[13]	FV	SP	AUG	OBJ	A+O	LF
aeroplane	79.4	80.2	81.7	81.6	82.9	83.1	83.8
bicycle	72.5	69.1	69.5	71.0	69.8	71.0	72.0
bird	55.6	52.8	55.9	58.0	57.4	61.0	59.7
boat	73.8	72.9	73.0	74.6	72.4	73.4	74.6
bottle	34.0	37.6	34.9	37.2	38.6	38.5	37.8
bus	72.4	69.5	71.7	71.3	69.8	70.7	72.9
car	83.4	81.8	81.7	82.1	82.2	83.2	82.9
cat	63.6	61.8	63.4	65.0	66.2	68.1	67.7
chair	56.6	54.9	57.0	58.2	53.9	57.2	57.8
cow	52.8	47.2	50.4	50.7	52.1	54.4	55.1
diningtable	63.2	61.5	63.8	64.9	62.4	64.5	66.7
dog	49.5	50.5	49.5	52.7	56.3	57.9	54.9
horse	80.9	79.1	80.3	80.8	79.8	80.7	81.6
motorbike	71.9	67.1	68.8	68.1	69.3	70.7	71.2
person	85.1	85.8	86.0	86.7	86.4	86.6	87.0
pottedplant	36.4	37.6	37.7	38.3	37.7	37.5	37.6
sheep	46.5	46.6	49.7	50.9	56.8	53.6	53.5
sofa	59.8	57.0	59.1	59.5	60.7	60.8	63.0
train	83.3	82.3	82.6	83.9	81.8	82.7	84.0
tvmonitor	58.9	59.0	58.7	60.0	59.5	59.7	60.7
average	64.0	62.7	63.8	64.8	64.5	65.8	66.3

VOC 2008 and VOC 2009. Next, we evaluate the performance of our system on both PASCAL VOC 2008 and PASCAL VOC 2009. We compare the performance of our A+O approach against the winning teams of these challenges: the “SurreyUVA_SKRDA” system on VOC 2008 [26] and the “NECUIUC_CLS-DTCT” system on VOC 2009 [27]. The first one is based on the combination of several types of detector/descriptor channels, the use of SPs and costly non-linear classifiers. The second one combines several encoding techniques with class-specific object detectors. We believe these two methods to be significantly more computationally intensive than ours. We also show results obtained with our baseline SP system for further comparisons. In the case of VOC 2009 results, we also include those obtained by Zhou *et al*. [13] with SVC. Table 2 shows the performance for each of the above systems. As a complementary note, table 3 compares the average performance of the LF system with the best results of table 2, showing that on these datasets the late fusion of SP and A+O classifiers brings little improvement (+0.4% absolute).

6. Conclusions

We addressed the problem of representing the spatial layout of images with two different and complementary approaches. Both originated from a theoretical well founded analysis. We showed on three of the challenging PASCAL VOC benchmarks the benefits of our approach: a higher accuracy without increasing the image signature dimensionality. Although our focus was on FVs, the generality of the approach makes it applicable to other BOF-based representations.

[1] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Computer Vision, 60 (2) 91–110, 2004

Table 2: Comparison with the state-of-the-art on the PASCAL VOC 2008 and PASCAL VOC 2009 datasets.

	VOC2008			VOC2009			
Class	[26]	SP	A+O	[27]	[13]	SP	A+O
aeroplane	79.5	83.5	85.0	88.0	87.1	86.4	87.1
bicycle	54.3	57.8	60.3	68.6	67.4	63.1	65.9
bird	61.4	62.7	67.4	67.9	65.8	62.5	68.1
boat	64.8	71.4	71.9	72.9	72.3	71.1	72.5
bottle	30.0	33.2	37.5	44.2	40.9	40.5	45.8
bus	52.1	56.4	57.9	79.5	78.3	75.7	76.4
car	59.5	64.6	67.9	72.5	69.7	66.2	69.3
cat	59.4	64.7	69.3	70.8	69.7	66.1	70.4
chair	48.9	49.2	47.5	59.5	58.5	56.3	56.0
cow	33.6	36.7	40.3	53.6	50.1	47.8	51.5
diningtable	37.8	40.2	46.2	57.5	55.1	52.8	55.8
dog	46.0	52.2	57.6	59.0	56.3	56.5	62.5
horse	66.1	68.1	72.0	72.6	71.8	68.7	73.1
motorbike	64.0	66.2	68.1	72.3	70.8	69.2	72.0
person	86.8	87.4	89.0	85.3	84.1	84.6	85.3
pottedplant	29.2	24.2	25.0	36.6	31.4	31.1	31.7
sheep	42.3	35.4	45.9	56.9	51.5	48.1	55.8
sofa	44.0	54.4	54.4	57.9	55.1	55.0	56.0
train	77.8	78.8	77.6	85.9	84.7	84.2	83.0
tvmonitor	61.2	63.5	67.2	68.0	65.2	66.6	68.4
average	54.9	57.5	60.4	66.5	64.3	62.6	65.3

Table 3: Average accuracy for the late-fusion based approach with the best performing systems on table 2.

	VOC2008			VOC2009			
Class	[26]	A+O	LF	[27]	[13]	A+O	LF
average	54.9	60.4	60.8	66.5	64.3	65.3	65.7

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, In: ECCV Int. Workshop on Statistical Learning in Computer Vision, pp. 1–22, 2004.

[3] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, In: Proc. Int. Conf. on Computer Vision, pp. 1470–1477, 2003.

[4] J. Farquhar, S. Szedmak, H. Meng, J. Shawe-Taylor, Improving "bag-of-keypoints" image categorisation, Tech. rep., University of Southampton (2005).

[5] J. van Gemert, J.-M. Geusebroek, C. Veenman, A. Smeulders, Kernel codebooks for scene categorization, In: Proc. European Conf. on Computer Vision, pp. 696–709, 2008.

[6] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1794-1801, 2009.

[7] J. Yang, K. Yu, T. Huang, Efficient highly over-complete sparse coding using a mixture model, In: Proc. European Conf. on Computer Vision, pp. 113–126, 2010.

[8] Y.-L. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 2559–2566, 2010.

[9] Y.-L. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, In: Proc. Int. Conf. on Machine Learning, pp. 111–118, 2010.

[10] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1–8, 2007.

[11] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, In: Proc. European Conf. on Computer Vision, pp. 143–156, 2010.

[12] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 3304–3311, 2010.

[13] X. Zhou, K. Yu, T. Zhang, T. S. Huang, Image classification using super-vector coding of local image descriptors, In: Proc. European Conf. on Computer Vision, pp. 141–154, 2010.

[14] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, In: Proc. Int. Conf. on Computer Vision, pp. 1458–1465, 2005.

[15] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 2169–2178, 2006.

[16] M. Marszalek, C. Schmid, H. Harzallah, J. van de Weijer, Learning representations for visual object class recognition, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/workshop/marszalek.pdf, 2007.

[17] V. Viitaniemi, J. Laaksonen, Spatial extensions to bag of visual words, In: Proc. ACM Int. Conf. on Image and Video Retrieval, pp. 1–8, 2009.

[18] J. R. R. Uijlings, A. W. M. Smeulders, R. J. H. Scha, What is the spatial extent of an object?, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 770–777, 2009.

[19] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 73–80, 2010.

[20] T. E. deCampos, G. Csurka, F. Perronnin, Images as sets of locally weighted features, In: Computer Vision and Image Understanding, 116 (1) 68–85, 2012.

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

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results, http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[24] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, In: ACM Multimedia, pp. 1469–1472, 2010.

[25] L. Bottou, SGD, http://leon.bottou.org/projects/sgd.

[26] M. A. Tahir, K. van de Sande, J. Uijlings, F. Yan, X. Li, K. Mikolajczyk, J. Kittler, T. Gevers, A. Smeulders, UvA and Surrey at PASCAL VOC

2008, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/tahir.pdf (2008).

[27] Y. Gong, T. Huang, F. Lv, J. Wang, C. Wu, W. Xu, J. Yang, K. Yu, T. Zhang, X. Zhou, Image classification using Gaussian mixture and local coordinate coding, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/workshop/yu.pdf (2009).

[28] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, S M. Hu, Global contrast based salient region detection, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 409–416, 2011.

[29] J. Krapac, J. J. Verbeek, F. Jurie, Modeling spatial layout with fisher vectors for image categorization, In: Proc. Int. Conf. on Computer Vision, pp. 1487–1494, 2011.

[30] P. Koniusz, K. Mikolajczyk, Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match, In: Proc. Int. Conf. on Image Processing, pp. 661-664, 2011.

[31] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, In: Proc. British Machine Vision Conference, pp. 1–12, 2011

[32] O. M. Parkhi, A. Vedaldi, A. Zisserman, C. V. Jawahar, Cats and Dogs, In: Proc. Conf. on Computer Vision and Pattern Recognition, pp. 1–8, 2011.