Fisher Vectors for Fine-Grained Visual Categorization

Jorge Sánchez, Florent Perronnin, Zeynep Akata

Fisher Vectors for Fine-Grained Visual Categorization

Jorge Sánchez, Florent Perronnin and Zeynep Akata
Textual and Visual Pattern Analysis (TVPA) group
Xerox Research Centre Europe (XRCE)

Abstract

The bag-of-visual-words (BOV) is certainly the most popular image representation to date and it has been shown to yield good results in various problems including Fine-Grained Visual Categorization (FGVC) [3, 4]. Our contribution is to show that the Fisher Vector (FV) – which describes an image by its deviation from an "average" model – is an excellent alternative to the BOV for the FGVC problem. In this extended abstract we first provide a brief introduction to the FV. We then present theoretical as well as practical motivations for using the FV for FGVC. We finally provide experimental results on four ImageNet subsets: fungus, ungulate, vehicle and ImageNet10K. Compared to [4] which uses spatial pyramid (SP) BOV representations, we report significantly higher classification accuracies. For instance, **on ImageNet10K we report 16.7% vs 6.4% top-1 accuracy (a 160% relative improvement).**

1. The Fisher Vector in a Nutshell

We only provide a very brief introduction to the FV. More details can be found in [8, 9]. Let $X = \{x_t, t = 1 \dots T\}$ be a set of T local descriptors extracted from an image (*e.g.* SIFT descriptors [7]). Let u_λ be a Gaussian Mixture Model (GMM) with parameters λ which models the generation process of local descriptors in any image. u_λ is generally referred to as a visual vocabulary and is estimated offline from a large number of descriptors extracted from a representative set of images. The FV $\mathcal{G}_\lambda^X$ characterizes the sample X by its deviation from u_λ:

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X. \quad (1)$$

G_λ^X is the gradient of the log-likelihood with respect to λ:

$$G_\lambda^X = \frac{1}{T} \nabla_\lambda \log u_\lambda(X). \quad (2)$$

L_λ is the Cholesky decomposition of the inverse of the Fisher information matrix F_λ of u_λ, *i.e.* $F_\lambda^{-1} = L'_\lambda L_\lambda$.

Here, $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \dots N\}$ where w_i, μ_i and σ_i are respectively the mixture weight, mean vector and standard deviation vector of Gaussian i. Let $\gamma_t(i)$ be the soft assignment of descriptor x_t to Gaussian i. We have the following formulas for the gradients with respect to μ_i and σ_i:

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

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

The FV $\mathcal{G}_\lambda^X$ is the concatenation of the $\mathcal{G}_{\mu,i}^X$ and $\mathcal{G}_{\sigma,i}^X$ vectors.

As shown in [9], square-rooting and L2-normalizing the FV can greatly enhance the classification accuracy. Also, following the SP matching approach of Lazebnik *et al*. [6], one can split an image into several regions, compute one FV per region and concatenate the per-region FVs.

2. The Fisher Vector For FGVC

We now motivate the usefulness of the FV for the FGVC problem. Especially, (i) compared to the BOV it includes higher order information, (ii) it describes an image by what makes it different from other images and (iii) it is scalable.

The FV includes higher-order information. The quantization process in the BOV is lossy and reduces the discriminative power of the representation. Since for FGVC we have to make use of subtle cues to distinguish between similar categories, we argue that we need to keep as much information as possible from the raw descriptors. It is obvious from equations (3) and (4) that the FV contains much more descriptor information than the BOV for the same number of Gaussians. Indeed, the FV extends the BOV by going beyond counting (0-order statistics): it also encodes statistics (up to the second order) about the distribution of descriptors assigned to each visual word. Actually, if each patch was assigned to a different Gaussian, then we could perfectly reconstruct the patches from the FV [1].

We note that other classification approaches which make use of the raw descriptor information have been suggested, such as the NN-based approach of Boiman *et al*. [1]. However we believe that the FV framework is more scalable

[1] We are not implying that having one patch per Gaussian is necessary to obtain good performance with the FV.

since [1] requires storing all patches of all training images which is not practical for very large datasets.

The FV is discriminative. We assume that the descriptors $\{x_t, t = 1 \dots T\}$ of a given image follow a distribution p and that we can decompose p into a mixture of two parts: a background image-independent part which follows u_λ and an image-specific part which follows an image-specific distribution q. Let $0 \leq \omega \leq 1$ be the proportion of image-specific information contained in the image:

$$p(x) = \omega q(x) + (1 - \omega) u_\lambda(x). \quad (5)$$

It was shown in [9] that the contribution of $(1 - \omega)u_\lambda(x)$ to the FV is approximately cancelled-out in $\mathcal{G}_\lambda^X$. This is a strong property since it shows that the FV implicitly discards the background (*i.e.* non-discriminative) information. We note that this has a direct consequence for our FGVC problem: the notion of background information can be *automatically* adapted for each fine-grained problem, simply by training the GMM u_λ on the relevant data. For instance, if u_λ is trained on fungus images, then background fungus information will be cancelled-out from the FV while if u_λ is trained on vehicle images, then background vehicle information will be cancelled-out.

The FV is scalable. FGVC typically implies that one has to deal with a large number of classes and, consequently, a large number of images (refer to Table 1 for a count of the number of classes and images for the different datasets we experimented on). We argue that the FV can scale to very large datasets containing millions of images.

First, the FV is efficient to compute. Indeed, for the same visual vocabulary size, the FV is much higher dimensional than the BOV and contains more discriminative information. Consequently, the FV gives state-of-the-art results even with tiny visual vocabularies.

Second, it was shown in [9] that the FV yields excellent results with linear classifiers – such as linear SVMs – which can be trained efficiently on a large scale using, for instance, Stochastic Gradient Descent (SGD) [2]. Also the cost of linear classification is independent of the number of support vectors.

Third, the FV can be very significantly compressed (by a factor of 64 in our experiments) with almost zero information loss using product quantization (PQ) [5]. For instance, we can store the 9M compressed FVs of ImageNet10K in approx. 80GB. Integrating the FV decompression in the SGD classifier learning yields a scalable training algorithm [10]. Indeed when a compressed FV is passed to the SGD algorithm, it is decompressed on-the-fly and, once it has been processed, the decompressed version of the sample is discarded. Hence, only one decompressed FV is kept in RAM at a time.

Dataset	Fungus	Ungulate	Vehicle	INet10K
#Classes	134	183	262	10K
#Images	88K	173K	226K	9M
[4]	11.6%	14.5%	24.1%	6.4%
FV	**19.4%**	**29.5%**	**42.3%**	**16.7%**

Table 1. Comparison of the FV with the SPBOV of [4].

3. Experimental Results

Datasets. We now provide experimental results on 4 datasets proposed in [4]: fungus, ungulate, vehicle and ImageNet10K. While the first 3 datasets consider a single fine-grained task, the last one considers multiple fine-grained problems simultaneously. For each dataset, we followed the same experimental protocol as [4]: half of the data is used for training and the other half is used for testing. We compute for each class the top-1 accuracy and report the average.

Experimental Set-up. Images are resized to 100K pixels (if larger). SIFT descriptors [7] are extracted densely at multiple scales. The feature dimensionality is reduced from 128 to 64 with PCA. We train a visual vocabulary with 256 Gaussians (using simple MLE) and use a SP with 4 regions (1 FV for the full image and 3 for the top, middle and bottom regions). This leads to 131,072-dim FVs. For the classification, we use linear SVMs trained with SGD [2].

Results. The results are reported in Table 1. It is clear that the FV leads to significantly higher results than the BOV on these 4 fine-grained datasets.

References

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

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

[3] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[4] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[5] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. on PAMI*, 33(1), 2011.

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

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[8] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[9] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

[10] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, 2011.