Exponential family Fisher vector for image classification

Jorge Sánchez[a,b,*], Javier Redolfi[b,c]

[a] *CONICET, Haya de la Torre S/N, Ciudad Universitaria, X5016ZAA Córdoba, Argentina*
[b] *Universidad Nacional de Córdoba, X5000HUA, Córdoba, Argentine*
[c] *CIII, Universidad Tecnológica Nacional, Facultad Regional Córdoba, X5016ZAA, Córdoba, Argentine*

ABSTRACT

One of the fundamental problems in image classification is to devise models that allow us to relate the images to higher-level semantic concepts in an efficient and reliable way. A widely used approach consists on extracting local descriptors from the images and to summarize them into an image-level representation. Within this framework, the Fisher vector (FV) is one of the most robust signatures to date. In the FV, local descriptors are modeled as samples drawn from a mixture of Gaussian pdfs. An image is represented by a gradient vector characterizing the distributions of samples w.r.t. the model. Equipped with robust features like SIFT, the FV has shown state-of-the-art performance on different recognition problems. However, it is not clear how it should be applied when the feature space is clearly non-Euclidean, leading to heuristics that ignore the underlying structure of the space. In this paper we generalize the Gaussian FV to a broader family of distributions known as the *exponential family*. The model, termed *exponential family Fisher vectors* (`eFV`), provides a unified framework from which rich and powerful representations can be derived. Experimental results show the generality and flexibility of our approach.

1. Introduction

In this work we focus on the problem of image classification, i.e. the task of assigning labels to images based on its content. Motivated by the tremendous growth on the volume and complexity of the image-related data, the problem has attracted great interest. Currently, not only the number of images has grown but also the nature of the visual information is changing towards more complex modalities, e.g. the use of deep information with the advent of RGBD cameras (Wang et al., 2014; Gupta et al., 2014) or the recent interest on hyperspectral imaging (Salamati et al., 2014) for solving different perception problems. Devising methods that allow us to capture the semantically rich information encoded in the images remains a major concern.

In the literature, one of the most successful approaches to tackle the problem has been to represent the images with summary statistics computed from a set of local patch descriptors and to use these "signatures" to learn the classifiers. Perhaps the most

*Corresponding author: Tel.: +54-351-4334051 Int. 309
e-mail: `jsanchez@famaf.unc.edu.ar` (Jorge Sánchez)

emblematic example of these models is the Bag-of-Visual-Words (BoVW) (Csurka et al., 2004; Sivic and Zisserman, 2003). In the BoVW, local descriptors are first encoded into fixed-length vectors using an auxiliary representation known as the visual codebook. Next, these vectors are aggregated into a global representation by a pooling operation (e.g. an average) and given as input to a classifier.

The BoVW model has been generalized to account for higher-order statistics with the introduction of the Fisher Vector (FV) (Perronnin and Dance, 2007), the VLAD (Jégou et al., 2010) and the Super Vector (Zhou et al., 2010), to name a few. Among these, FVs have shown to perform best in classification (Chatfield et al., 2011; Huang et al., 2014).

An underlying assumption in all of the above models is that local descriptors are –at least locally– normally distributed. For the BoVW, VLAD and SVs this is motivated by the use of the Euclidean distance during the encoding step while in the FV this follows from the explicit use of a mixture of Gaussian (GMM) pdfs to model the distribution of local features. Despite the great success of these models when built on top of robust descriptors like SIFT (Lowe, 2004), it is not clear how they should be applied in cases where the local feature space is clearly non-Gaussian, e.g. binary (Calonder et al., 2012; Alahi et al., 2012) or defined over the space of $n \times n$ symmetric positive definite (SPD) matrices (Tuzel et al., 2006; Ma et al., 2014). Note that this observation also holds for feature spaces which are subsets of $\mathbb{R}^n$, e.g. normalized histograms in the standard $(n-1)$-simplex or local features projected onto the unit sphere by a normalization operation. When having to address this problem in practice, it is common to pre- or post-process the data so that the assumptions made by the model are better fulfilled, in which case the core formulation remains unchanged. An example of such a strategy is the PCA projection step in the widely used SIFT + FV pipeline (Sánchez et al., 2013).

Although effective in practice, these heuristics are not very satisfactory from a modeling point of view since they effectively ignore the natural underlying structure of the data. As an illustrative example, let us consider binary descriptors as those proposed in (Calonder et al., 2012; Alahi et al., 2012).This family of features enjoy several properties which make them very appealing for large-scale recognition problems, e.g. they are very fast to compute (orders of magnitude faster than SIFT) and have a smaller memory footprint than the real counterpart. Nevertheless, they have so far being restricted mostly to matching (Heinly et al., 2012) and instance-level recognition problems.

One of the first attempts to use modern binary features in higher-level recognition problems can be found in (Gálvez-López and Tardós, 2011). In their work, the authors propose to learn a Bag-of-Binary-Words (BoBW) model using standard k-means followed by a rounding operation on the elements of the codebook. Using a more principled approach, (Zhang et al., 2013) proposed a learning scheme based on the Hamming distance that proved to be useful in classification. More related to our work, (Uchida and Sakazawa, 2013) derived a FV based on mixtures of Bernoulli pdfs which was shown to perform better than the BoBW in an object retrieval task. In (Caetano et al., 2014), the authors propose a model that extends the BoVW by computing histograms of distances

between the set of descriptors and each element in the codebook, learned using the k-medians algorithm and the Hamming distance.

Beyond the binary descriptor case, there has been a growing interest on using covariance matrices as local descriptors. Since covariance matrices lie on a rather complex manifold, dealing with them properly is a quite challenging. In this line of work, (Tuzel et al., 2006) considered the use of covariance descriptors built from simple features computed at pixel level (including the pixel location, color information and first- and second-order spatial derivatives). For classification, they relied on a boosting scheme using kNN classifiers and a distance metric specialized to covariance matrices. In the context of 3D shape analysis, (Tabia et al., 2014) proposed a model that extends the BoVW by using geodesic distances on the manifold of SPD matrices. The approach showed superior performance in shape matching and retrieval tasks compared to other descriptor-based approaches. In the same spirit, (Faraki et al., 2014) proposed a Bag-of-Riemanninan-Words model based on the Karcher mean (Pennec, 2006) (codebook learning) and the Stein divergence (Sra, 2011) (sample assignment). In the same work, the authors also proposed a "Fisher Tensor" (FT) model which consists on an embedding of the manifold of SPD matrices into a vector space so that a Gaussian FV can be learned on that space. These models were successfully applied to the classification of human cells from 2D images.

In this paper, we generalize the FV formalism to a broader family of distributions known as the *exponential family*. Since members of this family are defined on a variety of domains, our model –termed the *exponential family Fisher vector* (`eFV`)– provides a unified framework from which flexible and powerful representations can be derived.

Our main contributions are the following ones. We provide a complete derivation of the FV on sets, considering also the case of varying sample cardinalities (Section 2). We propose a model that generalizes the state-of-the-art FV encoding to mixtures of non-Gaussian pdfs in a unified and natural way (Section 5). We extend the diagonal normalization in the original FV formulation to a block-diagonal form and provide a simple and general method for its estimation. We analyze the case of finite input spaces and show that, in this case, linear classification becomes independent of the model complexity (Section 5.1). We show on two very different and challenging classification problems (Section 6) the power and flexibility of the proposed approach.

The code used to learn the models and to compute `eFV` signatures will be made available on the project website (`http://www.famaf.unc.edu.ar/~jsanchez/efv`).

2. The Fisher kernel framework

Let $S = \{p_\lambda\}$ be a family of distributions on $\mathcal{X}$, parameterized by a vector $\lambda = (\lambda_1, \ldots, \lambda_M)^T$. The set S can be regarded as a M-dimensional Riemannian manifold with a metric given by the Fisher information matrix (Amari, 1985). We can attach to each λ (a point on the manifold) a n-dimensional vector space known as the *tangent space* of S at λ; let us denote it as $T_\lambda S$. A vector on $T_\lambda S$ is a linear combination of basis vectors $\partial_i \stackrel{\text{def}}{=} \partial_{\lambda_i}$, $i = 1, \ldots, M$. Among all vectors on $T_\lambda S$, the *natural gradient* (Amari, 1998) gives the direction of steepest ascent for functions on the manifold. Let us consider the function $\log p_\lambda(X)$ viewed as a function of X. In

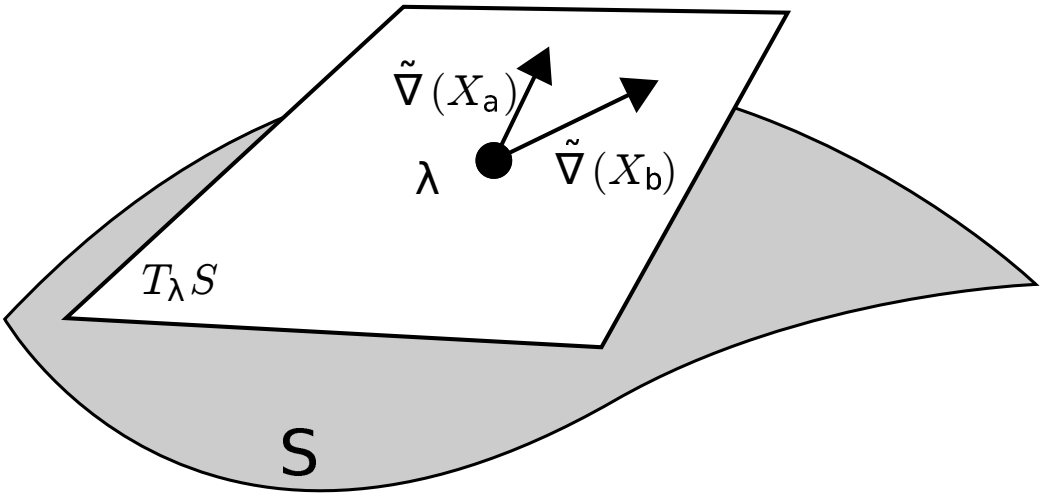

Figure 1. Illustration of the FK on S. $\tilde{\nabla}(X)$ denotes the natural gradient vector applied to $\log p_\lambda(X)$.

the statistical literature, the gradient of the log-likelihood w.r.t. the parameters is known as the *score*, and it plays a fundamental role in estimation theory. The Fisher kernel (FK) (Jaakkola and Haussler, 1998) is the inner product between natural gradient vectors acting on the function $\log p_\lambda(X)$ relative to the local Riemannian metric at λ. Figure 1 illustrates the concept. Concretely, let X_a and X_b be two samples drawn from $\mathcal{X}$. The FK $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is defined as:

$$K(X_a, X_b) \stackrel{\text{def}}{=} [\nabla_\lambda \mathcal{L}(X_a; \lambda)]^T I_\lambda^{-1} [\nabla_\lambda \mathcal{L}(X_b; \lambda)] \,, \tag{1}$$

where $\mathcal{L}(\mathrm{X}; \lambda) \stackrel{\text{def}}{=} \log p_\lambda(X)$ and I_λ is the Fisher information matrix (FIM) for p_λ. The FK can be regarded as a measure of the similarity between samples based on how they would affect the model (in a maximum-likelihood sense) if they were used to update its parameters from λ to $\lambda + d\lambda$ along the manifold.

Fisher vector (FV). For the matrix I_λ, the following decomposition holds: $I_\lambda^{-1} = {L_\lambda}^T L_\lambda$. Eq. (1) can thus be rewritten as $K(X_a, X_b) = [L_\lambda \nabla_\lambda \mathcal{L}(X_a; \lambda)]^T [L_\lambda \nabla_\lambda \mathcal{L}(X_b; \lambda)]$. The vector generated from X by the mapping $g : \mathcal{X} \to \mathbb{R}^M$,

$$g(\mathrm{X}) \stackrel{\text{def}}{=} L_\lambda \nabla_\lambda \mathcal{L}(X; \lambda) \tag{2}$$

is known as the FV encoding of X.

3. Fisher vectors on sets

Let $\mathrm{X} = \{x_n\}_{n=1}^N$ be a set of i.i.d. samples drawn from $\mathcal{X}$ and let p_λ be a valid distribution on $\mathcal{X}$. We consider two cases, according to whether N can be regarded as a constant or it depends on each particular X. In the first case, the FV encoding of any given X can be written as:

$$g(\mathrm{X}) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} L_\lambda \nabla_\lambda \log p_\lambda(x_n) \,. \tag{3}$$

The factor $1/\sqrt{N}$ results from the decomposition of the FIM for the product distribution $\prod_{n=1}^{N} p_\lambda(x_n)$. In this case the dot-product between the FVs of X_a and X_b *is* the FK between the samples. However, when N is variable, the dot product between the embeddings generated by (3) no longer corresponds to the explicit decomposition of the FK as before. Nevertheless, we can extend the above formulation by introducing the cardinality of the sample explicitly into the model as follows. Let us define $N = \mathrm{card}\,(\mathrm{X})$ as a random variable following a Poisson distribution of parameter θ and consider the following joint model for N and X:

$$p_{(\theta,\lambda)}(\mathrm{X}, N) = p_\theta(N) p_\lambda(\mathrm{X}|N) = p_\theta(N) \prod_{n=1}^{N} p_\lambda(x_n)\ . \tag{4}$$

$p_\theta(N) \overset{\text{def}}{=} \theta^N \exp(-\theta)/N!$ and $\theta = \mathbb{E}_\theta[N] \in \mathbb{R}_+$ is the parameter of the distribution. The FIM for this model can be written as:

$$I_{(\theta,\lambda)} = \begin{pmatrix} \frac{1}{\theta} & 0^T \\ 0 & \theta I_\lambda \end{pmatrix}\ , \tag{5}$$

where I_λ is the FIM for p_λ. The new mapping $\hat{g} : \mathcal{X} \times \mathbb{Z}^+ \to \mathbb{R}^{M+1}$ becomes:

$$\hat{g}(\mathrm{X}, N) = \frac{1}{\sqrt{\theta}} \begin{pmatrix} N - \theta \\ \sum_{n=1}^{N} L_\lambda \nabla_\lambda \log p_\lambda(x_n) \end{pmatrix}\ . \tag{6}$$

From Eq.(6), the meaning of the Poisson term becomes clear: it encodes the deviation from the mean of the number of elements in the set. It reduces to the standard FV formulation for problems dealing with fixed cardinalities. In what follows, we focus on the case $N = 1$ since its extension to arbitrary sets is straightforward with the above definitions.

4. The exponential family mixture model

Working with the FV requires choosing an appropriate parametric form for p_λ and which model to choose depends heavily on the particularities of the data. In practice, it is often the case that little is known or can be assumed about the structure of the data beyond the range of values for which it is defined. Even if our model is well formulated, different problems might require different levels of complexity in order to be able to capture the subtleties and particularities of the actual data. Based on this observations, we extend the FV by considering mixture distributions of the form:

$$p_\lambda(x) = \sum_{k=1}^{K} w_k p_k(x), \quad w_k > 0\ \forall k, \quad \sum_{k=1}^{K} w_k = 1\ , \tag{7}$$

and where $p_k : \mathcal{X} \to \mathbb{R}_+$ is chosen to be a member of the q-parameter exponential family, i.e. distributions of the form:

$$p_k(x) \equiv p(x; \eta_k) = h(x) \exp\left(\langle \eta_k, T_k(x) \rangle - \psi(\eta_k)\right)\ . \tag{8}$$

$\eta_k \in \mathbb{R}^q$ is the vector of natural parameters, $T_k(x) \in \mathbb{R}^q$ is the vector of sufficient statistics for the distribution, $\psi : \mathbb{R}^q \to \mathbb{R}$ is known as log-partition and $h : \mathcal{X} \to \mathbb{R}$ is a normalizer. Table 1 show some examples of distributions which are members of the exponential family.

Table 1. Examples of 1-dimensional (top) and multivariate (bottom) exponential family. † $\mathcal{S}(D)$ denotes the space of SPD $D \times D$-matrices.

Distribution	$\mathcal{X}$	$T(x)$	$\psi(\eta)$	$h(X)$	$H(t)$
Gaussian	$\mathbb{R}$	$\begin{pmatrix} x & x^2 \end{pmatrix}^T$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$	1	$\begin{pmatrix} \frac{t_1}{t_2 - t_1^2} & -\frac{1}{2}\frac{1}{t_2 - t_1^2} \end{pmatrix}^T$
Bernoulli	$\{0, 1\}$	x	$\log(1 + e^\eta)$	1	$\log\left(\frac{p}{1-p}\right)$
Exponential	$\mathbb{R}_+$	x	$-\log(-\eta)$	1	$-t$
Poisson	$\mathbb{N}$	x	e^η	$\frac{1}{x!}$	$\log(t)$
Multivariate ext.	$\mathcal{X}^D$	$\begin{pmatrix} T(x_1) & \dots & T(x_D) \end{pmatrix}^T$	$\sum_{i=1}^D \psi(\eta_i)$	$\prod_{i=1}^D h(x_i)$	$\begin{pmatrix} H(t_1) & \dots & H(t_D) \end{pmatrix}^T$
Dirichlet	$[0, 1]^D$	$\begin{pmatrix} \log(x_1) & \dots & \log(x_D) \end{pmatrix}^T$	$\sum_{i=1}^D \log\Gamma(\eta_i + 1) - \log\Gamma\left(\sum_{i=1}^D (\eta_i + 1)\right)$	1	(Minka, 2000)
Wishart†, n dof	$\mathcal{S}(D)$	x	$\log\Gamma_p\left(\frac{n}{2}\right) - \frac{n}{2}\log\lvert\eta\rvert$	$\lvert\eta\rvert^{(n-D-1)/2}$	$-\frac{n}{2}t^{-1}$

We follow (Krapac et al., 2011) and re-write the mixture weights as $w_i = \exp(\alpha_i)/\sum_{k=1}^K \exp(\alpha_k)$ in order to avoid having to enforce explicitly the normalization constraint in Eq. (7). For each choice of $p(\cdot; \eta_k)$, the vector $\lambda = \left(\alpha_1, \dots, \alpha_K, \eta_1^T, \dots, \eta_K^T\right)^T \in \mathbb{R}^{K(q+1)}$ fully characterizes the mixture distribution. The parameters can be readily estimated from a finite pool of samples using the EM algorithm (Redner and Walker, 1984). This would involve iterations of the form:

$$\eta_k^{(t+1)} \leftarrow H\left(\frac{\sum_{n=1}^N \gamma_k^{(t)}(x_n) T(x_n)}{\sum_{n=1}^N \gamma_k^{(t)}(x_n)}\right) \tag{9}$$

$$w_k^{(t+1)} \leftarrow \frac{1}{N}\sum_{n=1}^N \gamma_k^{(t)}(x_n), \tag{10}$$

where H is a maximum-likelihood estimator for the η_ks and $\gamma_i(X) \stackrel{\text{def}}{=} w_i p_i(X)/\sum_{k=1}^K w_k p_k(X)$ is the posterior of the sample given the ith component of the mixture.

4.1. Multivariate extension

For multivariate (vectorial) data, we can extend the 1-dimensional distributions to an arbitrary number of dimensions as follows. Let $p_\lambda : \mathcal{X} \to \mathbb{R}_+$ be a (univariate) distribution which is a member of the exponential family; we define its D-dimensional extension $\tilde{p}_\lambda : \mathcal{X}^D \to \mathbb{R}_+$ as:

$$\tilde{p}_\lambda(x) = \prod_{i=1}^D p_{\lambda_i}(x_i). \tag{11}$$

It is easy to verify that $\tilde{p}_\lambda$ can be expressed as in (8) and it is thus a member of the exponential family. If we consider mixtures of pdfs defined as in Eq. (11), we can show that:

$$\operatorname{cov}[X] = \sum_{k=1}^K w_k\left(\Sigma_k + \mu_k\mu_k^T\right) - \mathbb{E}[X]\mathbb{E}[X]^T \tag{12}$$

where $\mathbb{E}[X] = \sum_{k=1}^K w_k\mu_k$, $\mu_k = \mathbb{E}_{X\sim\tilde{p}_k}[X]$ and $\Sigma_k = \operatorname{cov}_{X\sim\tilde{p}_k}[X]$. Since the covariance in (12) is not diagonal, mixtures of distributions as Eq. (11) can still capture some of the correlations between the dimensions (Bishop, 2006).

5. The exponential family Fisher vector

In order to derive a FV mapping for the model in (7) we need to compute the gradient w.r.t. λ and to derive an expression for the normalizer in (2). For the gradients, we have:

$$\partial_{\alpha_k}\mathcal{L}(X;\lambda) = \gamma_k(X) - w_k \tag{13}$$

$$\nabla_{\eta_k}\mathcal{L}(X;\lambda) = \gamma_k(X)\Big[T(X) - \nabla_{\eta_k}\psi(\eta_k)\Big]\,. \tag{14}$$

The computation of L_λ is rather costly since it requires the decomposition and inversion of a $K(q+1)\times K(q+1)$ matrix. This quickly becomes impractical as the number of mixture components or the input dimensionality increases. In order to make the problem tractable, we follow (Perronnin and Dance, 2007) and assume that the assignment of samples to the components of the mixture is almost "hard", in which case the matrix L_λ can be expressed as (see Appendix A for details of the derivation):

$$L_\lambda \approx \begin{pmatrix} L_\alpha & 0 & \cdots & 0 \\ 0 & L_1/\sqrt{w_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_K/\sqrt{w_K} \end{pmatrix}. \tag{15}$$

$L_\alpha \overset{\text{def}}{=} \text{diag}\Big(\frac{1}{\sqrt{w_1}},\ldots,\frac{1}{\sqrt{w_K}}\Big)$ and L_i is the matrix that results from $I_i^{-1} = L_i^T L_i$, with I_i the FIM for the ith mixture component considered alone. These can be readily estimated by sampling from each component separately and computing the sample covariances of the scores, which is more efficient than sampling from the mixture distribution and approximating the block normalizers directly. Computing the normalizer (15) requires the inversion and decomposition of K SPD matrices of dimension $q\times q$ instead of the full matrix.

Using (13)–(15), we define a family of models of the form $g(x) \overset{\text{def}}{=} \Big(g_\alpha(x)^T, g_1(x)^T, \ldots, g_K(x)^T\Big)^T$, where:

$$g_\alpha(x) = \left(\frac{\gamma_1(x)-w_1}{\sqrt{w_1}},\ldots,\frac{\gamma_K(x)-w_K}{\sqrt{w_K}}\right)^T, \tag{16}$$

$$g_i(x) = \frac{\gamma_i(x)}{\sqrt{w_i}} L_i[T(x) - \nabla_i\psi(\eta_i)], \quad i = 1,\ldots,K\,, \tag{17}$$

and term this family the *exponential family Fisher vector* (`eFV`).

An interpretation for the `eFV`. For distributions that belong to the exponential family it holds that $\nabla_\eta\psi(\eta) = \mathbb{E}_{x\sim p}[T(X)]$. The gradients in Eq (14) can thus be written as:

$$\nabla_{\eta_k}\mathcal{L}(X;\lambda) = \gamma_k(x_n)\Big[T(x_n) - \mathbb{E}_{x\sim p_k}[T(X)]\Big]\,. \tag{18}$$

From which it follows that the `eFV` encodes the deviation from its expected value of the sufficient statistics of the sample w.r.t. the model p_λ.

5.1. Linear classification and finite input spaces

To close this section, we discuss the problem of linear classification with the `eFV` when the input space $\mathcal{X}$ is finite, i.e. there exists an $n \in \mathbb{N}$ such that $\iota : \mathcal{X} \to \{0, 1, \cdots, n-1\}$ is a bijection. A linear classifier f_w acting on the `eFV` encoding of a set $\mathrm{X} = \{x_i\}_{i=1}^N$ would predict a score:

$$f_w(g(\mathrm{X})) = w_0 + w^T g(\mathrm{X}) = \frac{1}{\sqrt{\theta}} \sum_{n=1}^{N} f_w(g(x_n)) \,. \tag{19}$$

Since $\mathcal{X}$ is finite, we can precompute the classification scores for each $x \in \mathcal{X}$ and build a table (indexed by $\iota(x)$) so that, at test time, the composition $f_w \circ g$ reduces to a simple table look-up. In this case, the cost of classifying a new sample is independent of the number of mixture components K. This goes one step further than (Cinbis et al., 2013; Oneata et al., 2014) since we do not even need to compute the `eFV` for the elements in the set.

The same approach can be applied to index the norms of the `eFV`s associated to each $x \in \mathcal{X}$ in order to compute an approximate L_p-normalization as in (Oneata et al., 2014).

6. Experiments

We evaluate different aspects of the `eFV` encoding as well as its overall performance on two challenging classification problems. We run experiments using different combinations of local features and mixture distributions in order to show the flexibility and generality of the approach. We first describe the datasets used in the evaluations and describe our experimental setup. We then report results.

Datasets. The Pascal VOC 2007 (Everingham et al., 2007) contains around 10K images depicting 20 different object categories. In spite of its relatively small size, it remains as one of the most challenging datasets in the literature (Torralba and Efros, 2011). For evaluation, we followed the recommended procedure as described in (Everingham et al., 2007). Classification performance is measured using the mean Average Precision (mAP) metric.

The KTH-TIPS2-a dataset (Caputo et al., 2005) contains 4395 images of 11 different texture materials acquired under different scales, poses and illumination conditions. The images are split into 4 subsets (samples *a, b, c* and *d*). We follow the standard protocol of taking each time one of the samples for testing and the remaining three for training. Performance is measured as the average accuracy over the four runs. We also report the accuracy for each individual run. For tuning the parameters we use 5-fold cross-validation.

Experimental setup. We describe the general pipeline used in the experiments. Since different combinations of local features and mixture distributions were used, details regarding each particular choice are given in the corresponding subsection.

- *Local features*. Given an image, we compute a resolution pyramid with 5 levels and a downsampling factor of $2^{-1/2}$. From each layer, we extract local descriptors from patches of 24×24 pixels sampled regularly using a step of 6 pixels.
- *Mixture model*. To fit the parameters we use the Expectation Maximization (EM) algorithm and 1M random samples from the training set. To initialize the EM iterations, we run k-means on a subset of the data and use the proportion and sufficient statistics of the samples assigned to each cluster as initial guess for the w_is and η_is, respectively.
- *eFV signature*. We compute the eFV normalizer by sampling from each component as described in Section 5. The Poisson parameter θ is set to the average number of samples extracted from the images in the training set. As in (Perronnin et al., 2010), we apply the signed square-root transformation and L_2-normalize the resulting vector.
- *Classifiers*. We rely on linear SVMs and a *one-vs-all* strategy. We use LIBLINEAR (Fan et al., 2008) for training.

In the following, we use eFV-X to denote the eFV derived from a *X*MM, e.g. eFV-G will denote a eFV signature computed from a mixture of Gaussian (GMM) pdfs.

6.1. Effect of the sample cardinality

We first evaluate the effect of including the Poisson term into the formulation (Eq. (6)). Since the gradient w.r.t. θ adds a single extra dimension to the –in general high-dimensional– eFV, we could not expect large improvements compared to the basic formulation. Nevertheless, this extra dimension could add valuable information regarding the characteristic size of the objects. The idea is that, since image patches are sampled regularly, the cardinality of the sample relates to the size of the objects in the image. To test the hypothesis we run experiments on the Pascal VOC 2007, using the provided bounding boxes to focus the computations on "foreground" patches only. We say that an image patch belongs to the foreground if it lies inside *any* of the bounding boxes provided for that image.

For this experiment we use SIFT vectors (4×4 cells of 6×6 pixels each) projected to 64-dimensions with PCA (SIFT-PCA, $\mathcal{X} = \mathbb{R}^{64}$) and eFV-G signatures[1]. Besides the block-diagonal normalizer of Eq. (15), we also consider a fully diagonal model obtained by restricting each block to be also diagonal. Figure 2 shows results for increasing values of K.

As it can be observed, there is a small but noticeable effect on adding the extra Poisson term to the representation. Remarkably, this extra term represents only a tiny fraction of the whole signature dimensionality (0.003% for $K = 256$). When we ran the same experiments using the full set of descriptors we did not observe any significant improvement, as expected. In this case, the cardinality of the sample no longer carries any discriminative information. Regarding the different normalizations, we observe that using the block-diagonal formulation leads to slightly better performances (+0.3 absolute points). In the following, we use a block-diagonal as the default normalization model for eFV-G.

[1] Note that the multivariate extension of the 1d Gaussians (Section 4.1) lead to a mixture with diagonal covariances, as in standard FV.

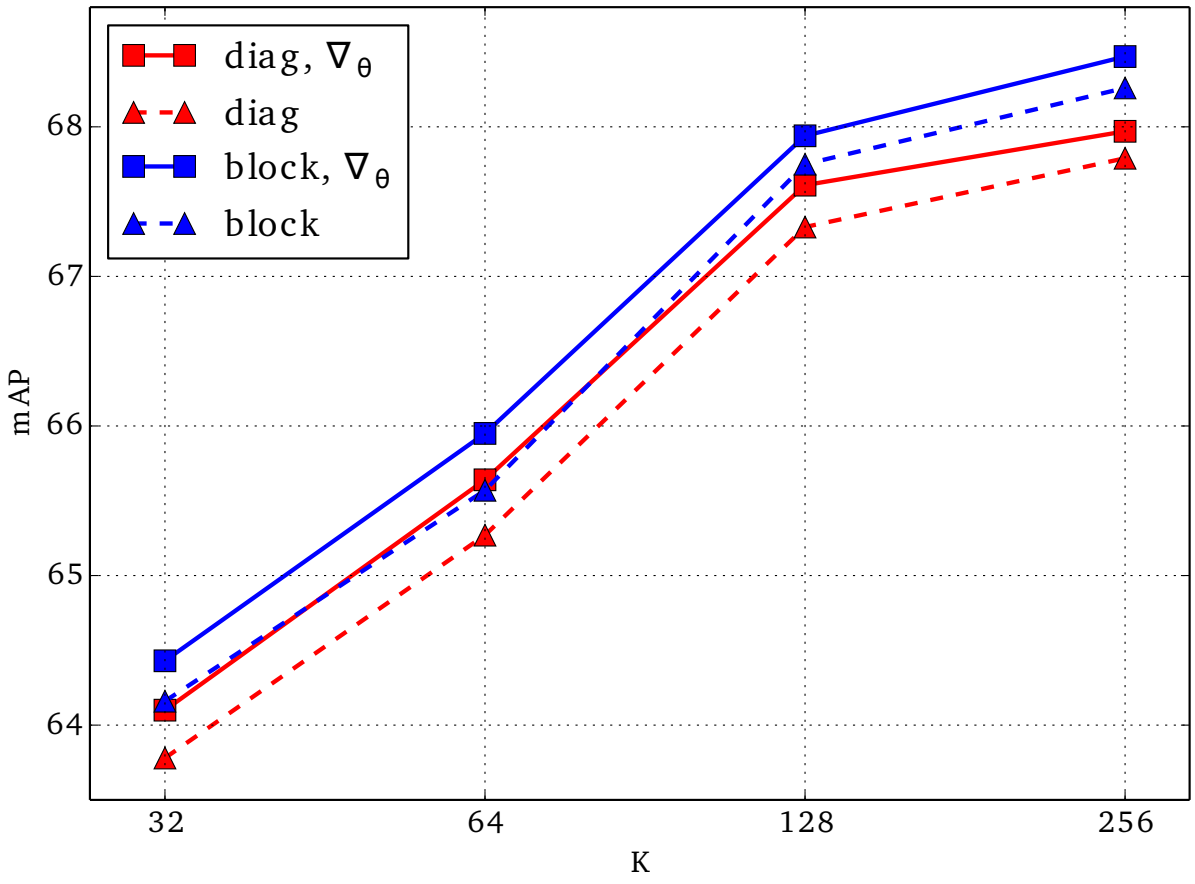

Figure 2. Effect of the different `eFV` formulations as measured on the "masked" Pascal VOC 2007 dataset (see text for details). Diagonal (red) and block-diagonal (blue) normalizer; with (solid) and without (dashed) the gradient w.r.t. θ.

6.2. Classification with binary features

Here we focus on the problem of classification using binary features. We report results on both the PASCAL VOC 2007 and KTH-TIPS2-a datasets. We use a mixture of multivariate Bernoulli pdfs[2]. To fit the model, we use the *random prototype* method of Juan et al. (2004) (with a mixing factor of 0.5) for initialization.

Pascal VOC 2007. Figure 3 shows results on PASCAL VOC 2007 for following configurations: a baseline system based on SIFT-PCA and `eFV`-G signatures; two systems based on binarized SIFT-PCA features[3]: the first based on modeling the distribution of local features with BMMs (SIFT-PCA-bin, `eFV`-B), the second treating the binary data as real-valued and using GMMs as models (SIFT-PCA-bin, `eFV`-G). The later can be seen as an "heuristic" approach to classification with binary features. Finally, we also report results using BRIEF features (256 bits) (Calonder et al., 2012) and `eFV`-Bs (BRIEF-256, `eFV`-B).

The baseline achieved a performance of 59.5% mAP for $K = 512$. This is comparable to the best results published for this

[2]The multivariate extension of the Bernoulli distribution is also known in the literature as *multinoulli* distribution (Murphy, 2012).

[3]Obtained by applying the function $b(z) \stackrel{\text{def}}{=} \max\{0, \text{sign}(z)\}$ dimension-wise

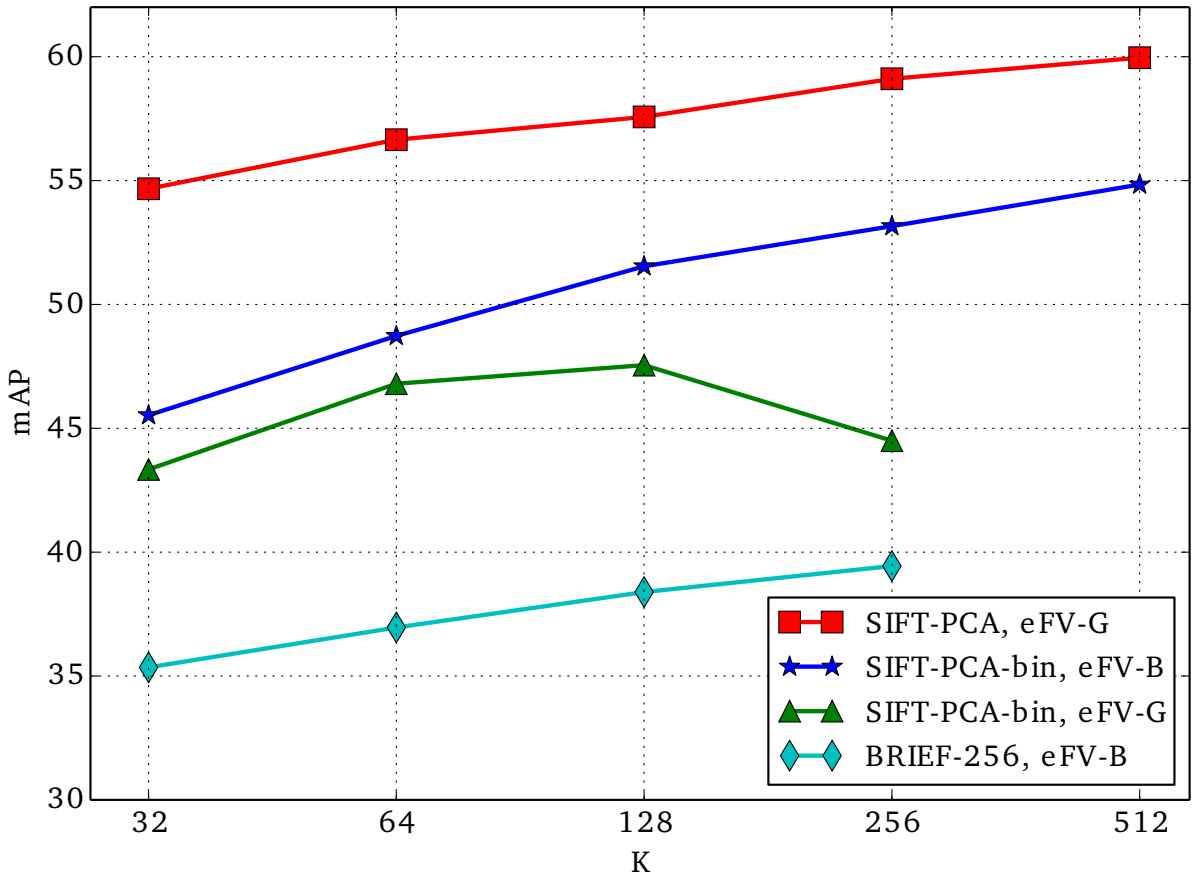

Figure 3. `eFV` **on binary descriptors evaluated on the PASCAL VOC 2007 dataset. SIFT-PCA FV baseline (red);** `eFV`**-G (green) and** `eFV`**-B (blue) on binarized SIFT-PCA vectors. BRIEF features and** `eFV`**-B encoding (cyan).**

dataset using SIFT-PCA features and Gaussian FVs (Sánchez et al., 2013). If we consider the two systems based on SIFT-PCA-bin features, we observe that for the system using GMMs performance reaches a peak of 47.5 at $K = 128$ whereas for the system based on BMMs there is a steady increase, reaching 54.8% mAP at $K = 512$. Note that, for the same value of K, `eFV`-B have roughly half of the dimensions of `eFV`-G (since the GMM parameters include the means and the variances explicitly). For the same dimensionality, the performance achieved by `eFV`-B ($K = 512$) is *only* 5 absolute points below the performance of the baseline. From a storage perspective, the memory requirement for the binarized features are 32 times smaller (for single precision floats) than for their real counterpart.

For BRIEF, we achieve a maximum performance of 39.4% mAP ($K = 256$). To the best of our knowledge, this outperforms the best results published on this dataset for this type of features. For instance, Caetano et al. (2014) reports a performance of 36.2% mAP with a model (de Avila et al., 2013) computed based on a vocabulary of 1024 codewords, spatial pyramids and non-linear SVMs.

For `eFV`-B we did not observe any significant difference between the diagonal and the block-diagonal normalizers.

Table 2. Classification results on KTH-TIPS2-a (see text for details).

Method	a	b	c	d	Accuracy
LBPH	71.6	70.4	79.0	63.6	71.1 (5.5)
LBP, eFV-B	73.2	72.7	76.6	65.0	71.9 (4.2)
SIFT-PCA, eFV-G	81.5	78.6	76.3	73.7	77.5 (2.9)
DCov, eFV-W (diag)	85.5	76.2	74.2	69.7	76.4 (5.6)
DCov, eFV-W (block)	85.9	71.2	73.2	73.9	77.5 (5.1)
LHS (Sharma et al., 2012)	–	–	–	–	73.0 (4.7)
DeCAF (Cimpoi et al., 2014)	–	–	–	–	78.4 (2.0)

KTH-TIPS2-a. Local Binary Pattern (LBP) (Ojala et al., 2002) is a popular descriptor in texture classification. In this experiment we use LBPs ($R = 1$, $P = 8$) computed densely and at multiple resolutions (we use the same pyramid configuration as before). We report results for a baseline system using normalized LBP histograms (LPBH) and for a system based on eFV-Bs ($K = 128$). For LBPH we apply the same normalizations as for eFV-B since we observed it greatly improves performance. Results are shown in the first two rows of Table 2. Considered separately, the system based on eFV-Bs obtains an improvement of around 2 points on three of the four samples and a slight improvement on the average over all runs.

6.3. SPD matrix descriptors

For this experiment we consider a variation of the covariance descriptors (DCov) proposed by (Tuzel et al., 2006) and eFV signatures based on mixtures of Wishart pdfs (WMM). From each patch, we compute the sample covariance of following pixel features:

$$F(x,y) = \begin{pmatrix} x & y & \sigma & I & |I_x| & |I_y| & |I_{xx}| & |I_{xy}| & |I_{yy}| \end{pmatrix}. \tag{20}$$

Here, (x, y) and σ are the pixel coordinates and scale of the patch, respectively; $I \stackrel{\text{def}}{=} I(x, y)$ is the image intensity at (x, y) and $I_\xi \stackrel{\text{def}}{=} \frac{\partial I}{\partial \xi}$ ($I_{\xi\zeta} \stackrel{\text{def}}{=} \frac{\partial^2 I}{\partial \xi \partial \zeta}$) denote first (second) derivatives. This results in a 9×9 SPD matrix descriptor per patch.

Table 2 show the results obtained using DCov descriptors and eFV-W ($K = 64$) for both diagonal and block-diagonal normalizers. Parameter n was set to 576, i.e. equal to the number of pixels within a patch. We report results for a baseline system using DSIFT-PCA features and eFV-Gs ($K = 256$). As a comparison, we also show some results reported recently in the literature: the LHS model of (Sharma et al., 2012) and the deep-network based features of (Cimpoi et al., 2014). The first is based on the so-called "difference vectors" and FVs while the second takes as features the output of a deep network for which the last fully connected layer has been removed (DeCAF). In (Cimpoi et al., 2014), results are also reported for a system based on dense SIFT features and FVs (82.2 (4.6)) and for the combination of DeCAF and FVs (84.7 (1.5)). In our case, we use a single feature channel and the same sampling strategy to allow for a fairer comparison.

From the table, we see that for eFV-W using a block-diagonal normalizer might lead to better performance. In this case, the system based on DCov and eFV-Ws achieves a performance which is on par with the state-of-the-art for this dataset.

6.4. Local histograms

We now turn to the case of local features in the standard $(d-1)$-simplex and mixtures of Dirichlet pdfs (DMM). Here, we consider color histograms descriptors (ColH) computed as follows. Image patches are divided into 4×4 square cells. From each cell, we compute a (normalized) RGB histogram using 4 bins per color channel. Finally, cell level features are concatenated and re-normalized to have unit L_1-norm. The resulting descriptors are 192-dimensional.

We consider `eFV`-G and `eFV`-D signatures models and fix the number of components to $K = 512$. On PASCAL VOC 2007, the system based on `eFV`-G signatures achieved 39.7% mAP while that based on `eFV`-D obtained 41.0% mAP. Similar gains were observed for values of K ranging from 32 to 512 components even when, for the same K, the number of dimensions of `eFV`-G is almost twice that of `eFV`-D.

6.5. Computational cost

Next, we consider the cost of encoding a single sample using a mixture of K components for the models listed on Table 1. From Eq. (16) and (17), the only terms that need to be computed at runtime are the posteriors and sufficient statistics for the sample. Since $T(x)$ do not depends on the parameters of the distribution, it can be computed once and reused across components. The overall cost is therefore dominated by the computation of the posteriors and is $O(KD)$, with D the number of input dimensions. FIM normalization takes and extra $O(KD)$ and $O(KD^2)$ for the diagonal and block-diagonal models, respectively. This extra cost is, however, independent of the number of samples.

7. Conclusions

We proposed an image encoding formalism that extends the FV to mixtures of non-Gaussian pdfs. Our model provides a unified framework for the representation of the images using local features defined on general input domains. The model was evaluated empirically on two challenging datasets, for models based on mixtures of Gaussian, Bernoulli, Wishart and Dirichlet pdfs. Results showed the great flexibility and modeling power of our approach.

Appendix A. A block-diagonal approximation to L_λ

In this appendix we show that the block-diagonal form in Eq. (15) arises naturally under following the *hard-assignment approximation* (HAA): $\gamma_i(x)\gamma_j(x) \approx \gamma_i(x)$ if $i = j$ and 0 otherwise. We define $\ell_\xi \stackrel{\text{def}}{=} \nabla_\xi \mathcal{L}(X;\lambda)$ to avoid clutter.

For mixture distributions as in Eq. (7), the FIM consists of the following blocks: *a)* $\mathbb{E}[\ell_\alpha \ell_\alpha]$, *b)* $\mathbb{E}\left[\ell_\alpha \ell_{\eta_i}\right]$ and *c)* $\mathbb{E}\left[\ell_{\eta_i} \ell_{\eta_j}\right]$, $i, j = 1, \ldots, K$. Cases *a* and b were considered in (Sánchez et al., 2013, Appendix 1). For case *c*, all entries with $i \neq j$ are zero

under the HAA. For $i = j$, we have:

$$\mathbb{E}\Big[\ell_{\eta_i}\ell_{\eta_i}{}^T\Big] = \int_{\mathcal{X}} \gamma_i(x)\gamma_i(x)\Big[\tilde{T}_i(x)\Big]\Big[\tilde{T}_i(x)\Big]^T p_\lambda(x)dx \tag{A.1}$$

$$\approx w_i \int_{\mathcal{X}} p_i(x)\Big[\tilde{T}_i(x)\Big]\Big[\tilde{T}_i(x)\Big]^T dx = w_i I_i \tag{A.2}$$

where $\tilde{T}_i(x) \stackrel{\text{def}}{=} T(x) - \nabla_{\eta_i}\psi(\eta_i)$ and I_i is the FIM corresponding to the ith component considered alone. Inversion and decomposition of (A.2) lead to the $L_i/\sqrt{w_i}$ blocks in (15).

References

Alahi, A., Ortiz, R., Vandergheynst, P., 2012. FREAK: Fast retina keypoint, in: CVPR.

Amari, S., 1985. Differential-Geometrical Methods in Statistic. Springer, New York.

Amari, S.I., 1998. Natural gradient works efficiently in learning. Neural computation 10, 251–276.

de Avila, S.E.F., Thome, N., Cord, M., Valle, E., de Albuquerque Araújo, A., 2013. Pooling in image representation: The visual codeword point of view. CVIU 117, 453–465.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Caetano, C., Avila, S., Guimarães, S., Araújo, A.d.A., 2014. Representing local binary descriptors with bossanova for visual recognition, in: SAC.

Calonder, M., Lepetit, V., Özuysal, M., Trzcinski, T., Strecha, C., Fua, P., 2012. BRIEF: computing a local binary descriptor very fast. IEEE TPAMI .

Caputo, B., Hayman, E., Mallikarjuna, P., 2005. Class-specific material categorisation, in: ICCV.

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

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A., 2014. Describing textures in the wild, in: CVPR.

Cinbis, R.G., Verbeek, J.J., Schmid, C., 2013. Segmentation driven object detection with fisher vectors, in: ICCV.

Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints, in: ECCV SLCV Workshop.

Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A., 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. The Journal of Machine Learing Research 9, 1871–1874.

Faraki, M., Harandi, M.T., Wiliem, A., Lovell, B.C., 2014. Fisher tensors for classifying human epithelial cells. PR 47, 2348–2359.

Gálvez-López, D., Tardós, J.D., 2011. Real-time loop detection with bags of binary words, in: IROS.

Gupta, S., Girshick, R., Arbelaez, P., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation, in: ECCV.

Heinly, J., Dunn, E., Frahm, J.M., 2012. Comparative evaluation of binary features, in: ECCV.

Huang, Y., Wu, Z., Wang, L., Tan, T., 2014. Feature coding in image classification: A comprehensive study. IEEE TPAMI 36, 493–506.

Jaakkola, T., Haussler, D., 1998. Exploiting generative models in discriminative classifiers, in: NIPS.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010. Aggregating local descriptors into a compact image representation, in: CVPR.

Juan, A., García-Hernàndez, J., Vidal, E., 2004. EM initialisation for bernoulli mixture learning., in: SSSPR IAPR Workshop.

Krapac, J., Verbeek, J.J., Jurie, F., 2011. Modeling spatial layout with fisher vectors for image categorization, in: ICCV.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110.

Ma, B., Su, Y., Jurie, F., 2014. Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image and Vision Computing 32, 379–390.

Minka, T., 2000. Estimating a Dirichlet distribution.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI 24, 971–987.

Oneata, D., Verbeek, J., Schmid, C., 2014. Efficient action localization with approximately normalized fisher vectors, in: CVPR.

Pennec, X., 2006. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. JMIV 25, 127–154.

Perronnin, F., Dance, C.R., 2007. Fisher kernels on visual vocabularies for image categorization, in: CVPR.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification, in: ECCV.

Redner, R.A., Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26, 195–239.

Salamati, N., Larlus, D., Csurka, G., Süsstrunk, S., 2014. Incorporating near-infrared information into semantic image segmentation. CoRR .

Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.J., 2013. Image classification with the fisher vector: Theory and practice. IJCV 105, 222–245.

Sharma, G., ul Hussain, S., Jurie, F., 2012. Local higher-order statistics (lhs) for texture categorization and facial analysis, in: ECCV.

Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos, in: ICCV.

Sra, S., 2011. Positive definite matrices and the s-divergence. arXiv .

Tabia, H., Laga, H., Picard, D., Gosselin, P.H., 2014. Covariance descriptors for 3d shape matching and retrieval. CVPR .

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias, in: CVPR.

Tuzel, O., Porikli, F., Meer, P., 2006. Region covariance: A fast descriptor for detection and classification, in: ECCV.

Uchida, Y., Sakazawa, S., 2013. Image retrieval with fisher vectors of binary features, in: IAPR.

Wang, A., Lu, J., Wang, G., Cai, J., Cham, T., 2014. Multi-modal unsupervised feature learning for RGB-D scene labeling, in: ECCV.

Zhang, Y., Zhu, C., Bres, S., Chen, L., 2013. Encoding local binary descriptors by bag-of-features with hamming distance for visual object categorization, in: ECIR.

Zhou, X., Yu, K., Zhang, T., Huang, T.S., 2010. Image classification using super-vector coding of local image descriptors, in: ECCV.