Pattern Recognition Letters 59 (2015) 26–32

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

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, Córdoba, X5016ZAA, Argentina
[b] Universidad Nacional de Córdoba, Córdoba, X5000HUA, Argentina
[c] CIII, Universidad Tecnológica Nacional, Facultad Regional Córdoba, Córdoba, X5016ZAA, Argentina

ARTICLE INFO

Article history:
Received 10 October 2014
Available online 31 March 2015

Keywords:
Image classification
Fisher kernel
Fisher vectors
Exponential family

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 [17,43] or the recent interest on hyperspectral imaging [34] 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 emblematic example of these models is the Bag-of-Visual-Words (BoVW) [12,37]. 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) [31], the VLAD [21] and the Super Vector [45], to name a few. Among these, FVs have shown to perform best in classification [9,19].

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 [24], it is not clear how they should be applied in cases where the local feature space is clearly non-Gaussian, e.g. binary [1,7] or defined over the space of $n \times n$ symmetric positive definite (SPD) matrices [25,41]. 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 [35].

Although effective in practice, these heuristics are not very satisfactory from a modeling point of view since they effectively ignore

☆ This paper has been recommended for acceptance by R. Davies.

* Corresponding author at: Facultad de Matemática, Astronomía y Física. Universidad Nacional de Córdoba. Av. Medina Allende s/n. Ciudad Universitaria. X5000HUA. Córdoba. Argentina. Tel.: +54 351 4334051 (int. 309); fax: +54 351 4334054.
E-mail address: jsanchez@famaf.unc.edu.ar (J. Sánchez).

http://dx.doi.org/10.1016/j.patrec.2015.03.010
0167-8655/

the natural underlying structure of the data. As an illustrative example, let us consider binary descriptors as those proposed in [1,7]. 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 [18] and instance-level recognition problems.

One of the first attempts to use modern binary features in higher-level recognition problems can be found in [16]. 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. [44] proposed a learning scheme based on the Hamming distance that proved to be useful in classification. More related to our work, Uchida and Sakazawa [42] derived a FV based on mixtures of Bernoulli pdfs which was shown to perform better than the BoBW in an object retrieval task. In [6], 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. [41] 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. [39] 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. [15] proposed a Bag-of-Riemanninan-Words model based on the Karcher mean [30] (codebook learning) and the Stein divergence [38] (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 [2]. 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* [3] 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 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) [20] is the inner product between natural gradient vectors acting on the function $\log p_\lambda(X)$ relative to the local Riemannian metric at λ. Fig. 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:

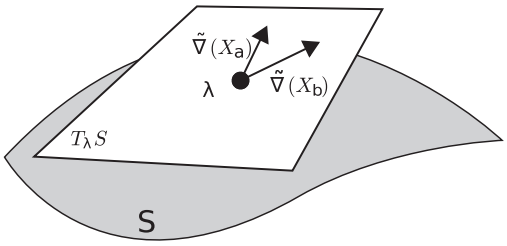

Fig. 1. Illustration of the FK on S. $\tilde{\nabla}(X)$ denotes the natural gradient vector applied to $\log p_\lambda(X)$.

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

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