Annales Univ. Sci. Budapest., Sect. Comp. 40 (2013) 201–214

FISHER KERNELS FOR IMAGE DESCRIPTORS: A THEORETICAL OVERVIEW AND EXPERIMENTAL RESULTS

Bálint Daróczy, András A. Benczúr and Lajos Rónyai
(Budapest, Hugary)

Dedicated to Professors Zoltán Daróczy and Imre Kátai on their 75th birthday

Communicated by Zoltán Horváth
(Received March 29, 2013; accepted June 10, 2013)

Abstract. Visual words have recently proved to be a key tool in image classification. Best performing Pascal VOC and ImageCLEF systems use Gaussian mixtures or k-means clustering to define visual words based on the content-based features of points of interest. In most cases, Gaussian Mixture Modeling (GMM) with a Fisher information based distance over the mixtures yields the most accurate classification results.
In this paper we overview the theoretical foundations of the Fisher kernel method. We indicate that it yields a natural metric over images characterized by low level content descriptors generated from a Gaussian mixture. We justify the theoretical observations by reproducing standard measurements over the Pascal VOC 2007 data. Our accuracy is comparable to the most recent best performing image classification systems.

Key words and phrases: Classification, images, Fisher kernel, Fisher information, Gaussian mixture model, generative model, discriminative model.
2010 Mathematics Subject Classification: 62F99, 68U10, 94A08.
Discussions on information theoretic topics with András Krámli and on image related topics with István Petrás are gratefully acknowledged.
The Project is supported by OTKA Grants NK 105645, K 77476 and K 77778. This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013). The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group (`www.ictlabs.elte.hu`).

1. Introduction

Image classification consists of assigning one or multiple labels to an image based on its semantic content. Although much progress has been made, in particular in the context of the PASCAL VOC [9] and ImageCLEF evaluation campaigns [16], the problem remains challenging. Several approaches model the distribution of low level features: bag of keypatches [7] or bag of visual terms [15], irrespective of their absolute or relative location. Categorization requires the estimation of the visual vocabulary, which is typically done by k-means [22, 7, 23], Gaussian Mixture Modeling (GMM) [18], mean-shift [14] or LDA [10].

Following the work of Jaakkola and Haussler [12], Perronnin and Dance [17] introduced Fisher kernels over a Gaussian mixture generative image model. The starting point of our experiments is the Perronnin-Dance method that proved to be very powerful especially for concept type classes, including best performance at the ImageCLEF and PASCAL VOC classification tasks [1, 19, 5]. In this paper we thoroughly define the generative model used in recent image classification systems and indicate why Fisher kernels capture a natural metric over the models. We give the theoretical background in Section 2. In Section 3 we describe our own experiments over the Pascal VOC 2007 data.

2. Generative image models, Fisher kernel and the Fisher metric

Powerful methods for image similarity and classification are based on a generative content model. Image regions or points of interest are generated from a Gaussian mixture as seen in Fig. 1. In this section we show why the Fisher distance is a natural metric to measure image dissimilarity under the above generative model. The model assumes that the D dimensional low level image descriptors originate from the mixture of N Gaussian distributions. We may think of these Gaussians (denoted by $\mathcal{N}_1, ..., \mathcal{N}_N$) as clusters.

Generative probability models (such as hidden Markov models) and discriminative approaches (such as support vector machines) are very important tools in the area of statistical classification of various types of data. Jaakkola and Haussler [12] proposed a remarkable and highly successful approach to combine the two, somewhat complementary approaches. Kernel methods for discriminative classification employ a real valued kernel function K to measure the similarity of two examples X, Y in terms of the value $K(X, Y)$. In many

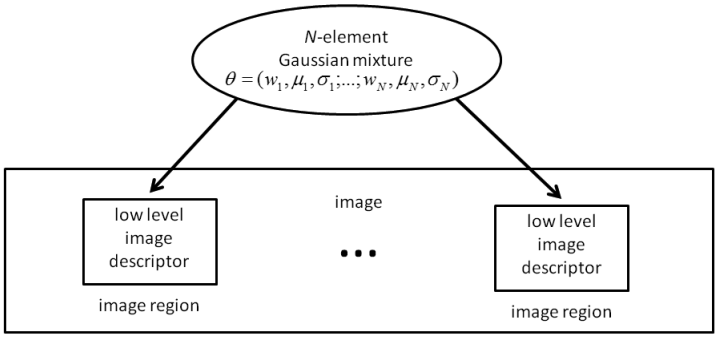

Figure 1. In the naive independence model, image regions are generated independent of each other according to the Gaussian mixture $P(X|\theta)$.

cases the kernel can actually be viewed as an inner product:

$$K(X,Y) = \phi_X^T \phi_Y,$$

where the feature vectors $\phi_X, \phi_Y \in \mathbb{R}^k$ are obtained via a fixed, problem specific map $X \mapsto \phi_X$ which describes the examples X in terms of a real vector of length k.

The main innovation of Jaakkola and Haussler [12] is to obtain the kernel function *directly from a generative probability model* and therefore obtain a kernel quite closely related to the underlying model. They consider a parametric class of probability models $P(X|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^l$ for some positive integer l. In the image content generative model (Fig. 1) $P(X|\theta)$ is given by N Gaussians $N(\mu_i, \sigma_i)$ with weights w_i for $i = 1,\ldots,N$.

Provided that the dependence on θ is sufficiently smooth, the collection of models with parameters from Θ can then be viewed as a (statistical) manifold M_Θ. M_Θ can be turned into a Riemannian manifold* [13] by giving a scalar product at the tangent space of each point $P(X|\theta) \in M_\Theta$ via a positive semidefinite matrix $F(\theta)$, which varies smoothly with the base point θ. Such positive semidefinite matrices are provided by the Fisher information matrix

$$F(\theta) := \mathbf{E}(\nabla_\theta \log P(X|\theta) \nabla_\theta \log P(X|\theta)^T),$$

*A Riemannian manifold M is a smooth real manifold, where for each point $p \in M$ there is an inner product defined on the tangent space of p. This inner product varies smoothly with p. One can define the length of a tangent vector via this inner product on the tangent space. This makes possible to define the length of a curve $\gamma(t)$ on M by integrating the length of the tangent vector $\dot{\gamma}(t)$. This in turn allows to define a metric on M. The distance between two points Q and Q' is just the length of the shortest curve on M from Q to Q'.

where the gradient vector $\nabla_\theta \log P(X|\theta)$ is

$$\nabla_\theta \log P(X|\theta) = \left(\frac{\partial}{\partial \theta_1} \log P(X|\theta), \ldots, \frac{\partial}{\partial \theta_l} \log P(X|\theta) \right),$$

and the expectation is taken over $P(X|\theta)$. In particular, if $P(X|\theta)$ is a probability density function, then the ij-th entry of $F(\theta)$ is

$$f_{ij} = \int_X P(X|\theta)(\frac{\partial}{\partial \theta_i} \log P(X|\theta))(\frac{\partial}{\partial \theta_j} \log P(X|\theta))dX.$$

The vector $U_X = \nabla_\theta \log P(X|\theta)$ is called the *Fisher score* of the example X. Now the mapping $X \mapsto \phi_X$ of examples to feature vectors can be $X \mapsto F^{-\frac{1}{2}} U_X$ (we suppressed here the dependence on θ). Thus, to capture the generative process, the gradient space of the model space M_Θ is used to derive a meaningful feature vector. The corresponding kernel function

$$K(X, Y) := U_X^T F^{-1} U_Y$$

is called the *Fisher kernel*.

An intuitive interpretation is that U_X gives the direction where the parameter vector θ should be changed to fit best the data X (see Section 2 in [17]).

2.1. Fisher distance: a univariate Gaussian example

The question arises why we use the Fisher metric on Θ instead of e.g. the Euclidean distance inherited from the ambient space $\mathbb{R}^l$? As a first step in discussing this issue, we follow [6] to consider the family of univariate Gaussian probability density functions

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right),$$

parameterized by the points of the upper half-plane H of points $(\mu, \sigma) \in \mathbb{R}^2$ with $\sigma > 0$. Fix values $0 < \sigma_1 < \sigma_2$ and $\mu_1 < \mu_2$. The Euclidean distance of $A = (\mu_1, \sigma_1)$ and $B = (\mu_2, \sigma_1)$ is $\mu_2 - \mu_1$, the same as the distance of $C = (\mu_1, \sigma_2)$ and $D = (\mu_2, \sigma_2)$. At the same time, an inspection of the graphs of the density functions shows[†] that the dissimilarity of the distributions attached to C and D is smaller than the dissimilarity of the distributions with parameters A and B. This suggests that a distance reflecting the dissimilarity of the

[†]Let f_A, f_B, f_C, f_D be the density functions corresponding to A, B, C, D and let I be a small interval close to μ_2. Then $\int_I |f_C - f_D| dx$ will be smaller than $\int_I |f_A - f_B| dx$.

distributions is not the Euclidean one. It turns out that the Fisher distance reflects dissimilarity much better in this case. In fact, the Fisher distance $d_F(P,Q)$ of two points $P=(\mu_1,\sigma_1)$ and $Q=(\mu_2,\sigma_2)$ is related nicely to the hyperbolic distance $d_H(P,Q)$ measured in the Poincaré half-plane model of hyperbolic geometry (formula (4) in [6]):

$$d_F(P,Q)=\sqrt{2}d_H\left(\left(\frac{\mu_1}{\sqrt{2}},\sigma_1\right),\left(\frac{\mu_2}{\sqrt{2}},\sigma_2\right)\right).$$

The significance of Fisher metric is highlighted by a fundamental result of N. N. Čencov [3] stating that it exhibits an invariance property under some maps which are quite natural in the context of probability[‡]. Moreover it is essentially the unique Riemannian metric with this property. This invariance property is discussed in Campbell [2] and it is extended by Petz and Sudár to a quantum setting [20]. We remark here that in the work [2] Campbell refers to the monograph [8] by János Aczél and Zoltán Daróczy as the primary source on information measures. Thus, one can view the use of Fisher kernel as an attempt to introduce a natural comparison of the examples on the basis of the generative model (see Section 4 in [12]).

2.2. The Fisher metric over general distributions

The Fisher metric over the Riemannian space

$$\Delta=\{(p_1,\ldots,p_n);\ p_i\geq 0,\ \sum p_i=1\}\subseteq\mathbb{R}^n$$

of finite probability distributions $(p_1,p_2,\ldots,p_n)$ has a beautiful connection to the metric of the sphere $S\subseteq\mathbb{R}^n$ of points $(x_1,\ldots,x_n)$ with $\sum_i x_i^2=4$. This goes back to Sir Ronald Fisher and is discussed in [2], [11] and [20]. A point $(p_1,\ldots,p_n)$ of the probability simplex Δ corresponds to a unique point of the positive "quadrant" of S^+ of S via $4p_i=x_i^2$, $i=1,2,\ldots,n$. This is actually an *isometry* if one considers the spherical metric on S^+. In fact, let $x(t)$ be a curve on S^+. Then the squared length of the tangent vector to $x(t)$ is

$$\|\dot{x}(t)\|^2=\sum_{i=1}^n(\dot{x_i}(t))^2=\sum_{i=1}^n((2\sqrt{p_i(t)})')^2=$$

$$=\sum_{i=1}^n\left(\frac{\dot{p_i}(t)}{\sqrt{p_i(t)}}\right)^2=\sum_{i=1}^n p_i(t)((\log p_i(t))')^2,$$

[‡]These maps are congruent embeddings by Markov morphisms.

which is the squared length of $\dot{p}(t)$ in the Fisher metric on Δ. The Fisher distance $d_F(P,Q)$ between probability distributions $P=(p_1,\dots,p_n)$ and $Q=$ $=(q_1,\dots,q_n)$ can then be calculated along a great circle of S. It will be

$$d_F(P,Q)=2\arccos\left(\sum_{i=n}^{n}\sqrt{p_iq_i}\right).$$

2.3. The Fisher metric over Gaussian mixtures: the image classification setup

For classification tasks Perronnin and Dance [17] proposed the Fisher metric over the Gaussian mixture image content generative model as a content based distance between two images. Let $X=x_1,..,x_T$ be a set of samples extracted from a particular image I_X. In the naive independence model, the probability density function of X is equal to

$$P(X|\theta)=\Pi_{t=1}^{T}P(x_t|\theta). \tag{2.1}$$

We obtain that the Fisher score of X is a sum over the Fisher scores of the samples of X

$$U_X=\nabla_\theta\log P(X|\theta)=\nabla_\theta\sum_{t=1}^{T}\log P(x_t|\theta).$$

The GMM assumption means that

$$P(x_t|\theta)=\sum_{i=1}^{N}w_iP_i(x_t|\theta),$$

where $(w_1,\dots,w_N)$ is a finite probability distribution and P_i is the density of $\mathcal{N}_i$, a D dimensional Gaussian distribution with mean vector $\mu_i\in\mathbb{R}^D$ and diagonal covariance matrix with diagonal $\sigma_i\in\mathbb{R}^D$.

By introducing the occupancy probability

$$\gamma_t(i)=\mathbb{P}(i|x_t,\theta)=\frac{w_iP_i(x_t|\theta)}{\sum_{j=1}^{N}w_jP_j(x_t|\theta)},$$

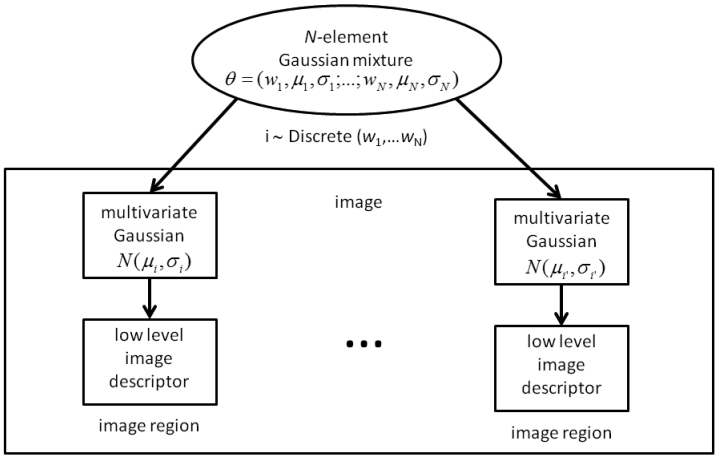

Figure 2. In this variant of the naive independence model, image regions are generated by first selecting one component of the mixture from a discrete distribution and then the low level descriptors are given by the selected multivariate Gaussian $N(\mu_i, \sigma_i)$.

the following formulae are obtained in [17] for the final gradients:

$$\frac{\partial \log P(X|\theta)}{\partial w_i} = \sum_{t=1}^{T} [\frac{\gamma_t(i)}{w_i} - \frac{\gamma_t(1)}{w_1}], \tag{2.2}$$

$$\frac{\partial \log P(X|\theta)}{\partial \mu_i^d} = \sum_{t=1}^{T} \gamma_t(i) [\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2}], \tag{2.3}$$

$$\frac{\partial \log P(X|\theta)}{\partial \sigma_i^d} = \sum_{t=1}^{T} \gamma_t(i) [\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d}], \tag{2.4}$$

where in the first equation (2.2) we consider $i > 1$ only, since $\sum_i w_i = 1$. The superscript d refers to the d-th coordinate of a vector from $\mathbb{R}^D$.

Despite the compact form of the above derivatives, the computation of the Fisher information remains a challenging problem. To overcome this difficulty, Perronnin and Dance further simplified the naive independence model of Fig. 1 as follows. In the model illustrated in Fig. 2, they assume that the sample x_t for image region $t \in \{1, \ldots, T\}$ is generated by first selecting one Gaussian $\mathcal{N}_j$ from the mixture according to the distribution $(w_1, \ldots, w_N)$ and then considering x_t as a sample from $\mathcal{N}_j$. In other words, they assume that the distribution of the occupancy probability is sharply peaked [17], resulting in only one Gaussian per sample with non-zero (≈ 1) occupancy probability. They also assume that

T, the number of regions generated for an image, is constant. We note that the assumptions on sharp peaks and a constant T are not entirely valid in our experiments, nevertheless we used the above formulas.

The final representation of image I_X is

$$G_X = F^{-\frac{1}{2}} U_X. \tag{2.5}$$

For this computation in practice a diagonal approximation of F is used. The diagonal terms of this approximation are

$$f_{w_i} \approx T(\frac{1}{w_i} + \frac{1}{w_1}); \tag{2.6}$$

$$f_{\mu_i^d} \approx \frac{Tw_i}{(\sigma_i^d)^2}; \tag{2.7}$$

$$f_{\sigma_i^d} \approx \frac{2Tw_i}{(\sigma_i^d)^2}. \tag{2.8}$$

For images I_X, and I_Y the Fisher kernel $K(I_X, I_Y)$ is the following bilinear kernel over the Fisher vectors G_X and G_Y:

$$K(I_X, I_Y) = U_X^T F^{-1} U_Y = U_X^T F^{-1/2} F^{-1/2} U_Y = G_X^T G_Y. \tag{2.9}$$

The dimension of the Fisher vector is $2ND + N - 1$, where D is the dimension of the samples. Since this value depends on N, the number of Gaussians in the mixture, one has to find a good balance between the accuracy of the mixture model and the computational cost.

2.4. Learning via the Fisher kernel

Jaakkola and Haussler in [12] propose the use of Fisher kernels for classification tasks. They introduce the notion of differential extension of models and show under reasonable assumptions that in this framework logistic regression with the Fisher kernel provides at least as powerful classification method as the underlying generative model.

Fisher kernels can be applied for image classification by computing the parameters of the generative model in Fig. 1 and then by using these parameters in the equations of the preceding subsection.

The mixture parameters in the generative model (Fig. 1) can be determined by Gaussian mixture decomposition via the standard expectation maximization (EM) algorithm [24]. In a popular interpretation, the mixture gives vocabularies of visual words in a "bag-of-words" representation of the image. In particular in the simplified model of Fig. 2, each Gaussian is a word of an N-element vocabulary and each region represents one visual word.

Table 1. Average MAP on Pascal VOC 2007

	LLC [5]	SV [5]	IFK [19]	IFK [19]	IFK [5]	Exp 1	Exp 2
Fine sampling	yes	yes	no	no	yes	yes	very
Descriptor	SIFT	SIFT	SIFT	SIFT	SIFT	HOG	HOG
Codebook size	25k	1024	256	256	256	507	507
Spatial Pooling	yes	yes	no	yes	yes	no	no
Dimension	200k	1048k	41k	327k	327k	97k	97k
MAP	.573	.582	.553	.583	.617	.579	.588

3. Experiments

We carried out our experiments by using the Pascal VOC 2007 data set [9], the most popular benchmark for image categorization. The Pascal VOC 2007 task uses 5011 training images and a test set with 4952 images, each image annotated manually into predefined object classes such as cat, bus, person or airplane. Our choice of dataset gave us an opportunity to compare our experiments to the winner methods (without detection) of later challenges including the SuperVector coding (SV) and Locality-constrained Linear Coding (LLC) [5]. To justify our experiments, we compare them to the Improved Fisher Kernel (IFK) results in [19] and [5].

3.1. Feature generation and modeling

We extracted multiple feature vectors per images to describe the visual content. We employed two different fine sampling procedures, the very dense sampling (Exp. 2 in Table 1) resulting in approximately 300,000 while the other (Exp. 1) about 72,000 (step size is equal to 3, similarly to [5]) keypoints (regions) per image. To describe the keypoints, we calculated HOG (Histogram of Oriented Gradients) with different sub-block sizes (4x4, 8x8, 12x12, 16x16 for Exp. 2 and 4x4, 6x6, 8x8, 10x10 for Exp. 1 as suggested in [5]). We reduced the original dimension (144) of the samples (low-level descriptors) to 96 by PCA. The Gaussian Mixture Model (GMM) was trained on a sample set of 3 million descriptors with 512 Gaussians. Our overall procedure is shown in Fig. 3.

We used the resulting kernels after applying the normalizations suggested in [19] with $\alpha = 0.125$ for training linear SVM models by the LibSVM package [4] for each of the 20 Pascal VOC 2007 concepts independently.

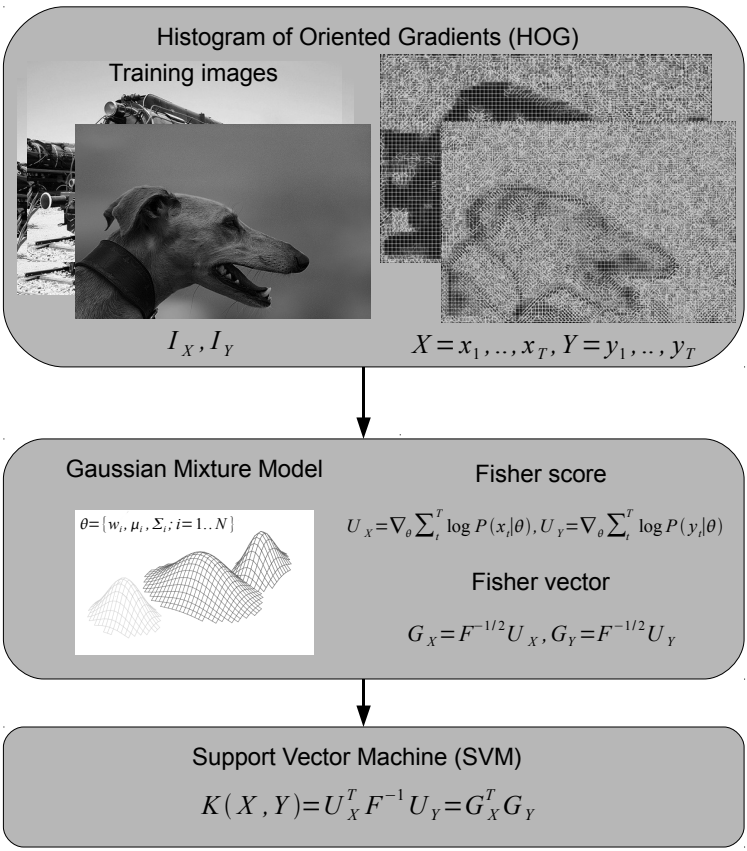

Figure 3. Our classification procedure

Table 2. MAP on Pascal VOC 2007 data set

	air plane	bicycle	bird	boat	bottle
Exp.2 fine no SP	.801	.665	.509	.738	.279
IFK no fine SP [19]	.757	.648	.528	.706	.300
IFK fine SP [5]	.789	.674	.519	.709	.307
	bus	car	cat	chair	cow
Exp.2 fine no SP	.646	.811	.608	.520	.390
IFK non fine SP [19]	.641	.775	.555	556	.418
IFK fine SP [5]	.721	.799	.613	.559	.496
	dining table	dog	horse	motor bike	person
Exp.2 fine no SP	.511	.453	.780	.643	.843
IFK non fine SP [19]	.563	.417	.763	.644	.827
IFK fine SP [5]	.584	.447	.788	.708	.849
	potted plant	sheep	sofa	train	tv/ monitor
Exp.2 fine no SP	.293	.446	.499	.779	.529
IFK non fine SP [19]	.283	.397	.566	.797	.515
IFK fine SP [5]	.317	.510	.564	.802	.574

3.2. Evaluation

Although spatial pooling is a widely used and effective extension to naive bag-of-words models [21, 19, 5], we did not apply it. Our consideration is based on the fact that the standard spatial pooling methods (split the images into 1x1, 3x1, 2x2 regions) contribute a huge increase in the dimension of the representation per image (8 times in [19, 5]). Despite the 3.3 times lower dimension of Exp. 2 the results are comparable to IFK fine SP [5] in five categories (within 5 percent range) and are better in four categories (airplane, boat, car and dog).

4. Conclusion

In this paper we described a Fisher kernel based approach of image classification. We gave theoretical background and provided experimental results. The resulting image classification system is comparable to the best performing

PASCAL VOC systems using SIFT descriptors (see Table 1), in some categories outperforming the best published Fisher vector based systems to date [5, 19] without Spatial Pooling and with 3.3 times lower dimension. Further improvement could be a better approximation of the Fisher information and a generative model capturing the intra image structure. The latter issue is quite serious. If we rearrange the samples (patches of a particular image) in an arbitrary way, then the Fisher vector of the resulting image will be the same as before, while the new image may be radically different.

As the scale of the research collections increases, researchers can no longer afford to spend days of CPU time on refined analysis and have to use simpler methods as fallback. As Gaussian mixture decomposition is one of the most time consuming tasks, we make our very fast graphical coprocessor (GPGPU) source code along with preprocessed visual classification data available for research purposes[§].

References

[1] **Ah-Pine, J., C. Cifarelli, S. Clinchant, G. Csurka and J. Renders,** XRCEs Participation to ImageCLEF 2008 *Working Notes of the 2008 CLEF Workshop*, Aarhus, 2008.

[2] **Campbell, L.L.,** The relation between information theory and the differential geometry approach to statistics. *Information sciences*, **35(3)** (1985), 199–210

[3] **Čencov, N.N.,** *Statistical Decision Rules and Optimal Inference*. Amer Mathematical Society, **53**, 1982.

[4] **Chang, C.-C. and C.-J Lin,** LIBSVM: a library for support vector machines, 2001, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] **Chatfield, K., V. Lempitsky, A. Vedaldi and A. Zisserman,** The devil is in the details: an evaluation of recent feature encoding methods. *British Machine Vision Conference*, Dundee, 2011.

[6] **Costa, S.I.R., S.A. Santos and J.E. Strapasson,** Fisher information distance: a geometrical reading. *preprint arXiv:1210.2354*, 2012.

[7] **Csurka, G., C. Dance, L. Fan, J. Willamowski and C. Bray,** Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, Prague, 2004.

[§] `https://dms.sztaki.hu/en/project/gaussian-mixture-modeling-gmm-and-fisher-vector-toolkit`

[8] **Daróczy, Z. and J. Aczél,** *On Measures of Information and their Characterizations.* Mathematics in Science and Engineering Volume 115, Academic Press, New York, 1975.

[9] **Everingham, M., L. Van Gool, C.K.I. Williams, J. Winn and A. Zisserman,** The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, **88(2)** (2010), 303–338.

[10] **Fei-Fei, L. and P. Perona,** A Bayesian hierarchical model for learning natural scene categories, *Computer Vision and Pattern Recognition*, volume 2, San Diego, 2005.

[11] **Gromov, M.,** In a search for a structure, part 1: On entropy. *Preprint*, 2012.

[12] **Jaakkola, T.S. and D. Haussler,** Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 1999, 487–493.

[13] **Jost, J.,** *Riemannian geometry and geometric analysis.* Springer, 2011.

[14] **Jurie, F. and B. Triggs,** Creating efficient codebooks for visual recognition. *Tenth IEEE International Conference on Computer Vision , ICCV*, volume 1, Beijing, 2005.

[15] **Monay, F., P. Quelhas, D. Gatica-Perez and J. Odobez,** Constructing visual models with a latent space approach. *Lecture notes in computer science*, **3940** (2006), 115–126.

[16] **Nowak, S.,** New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. *Cross Language Evaluation Forum, ImageCLEF Workshop*, Padua, 2010.

[17] **Perronnin, F. and C. Dance,** Fisher kernels on visual vocabularies for image categorization. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2007, 1–8.

[18] **Perronnin, F., C. Dance, G. Csurka and M. Bressan,** Adapted vocabularies for generic visual categorization. *European Conference of Computer Vision, ECCV*, 2006, 464–475.

[19] **Perronnin, F., J. Sánchez T. Mensink,** Improving the fisher kernel for large-scale image classification. *European Conference of Computer Vision, ECCV*, 2010, 143–156.

[20] **Petz, D. and C. Sudar,** Extending the Fisher metric to density matrices. *Geometry of Present Days Science*, 1999, 21–34.

[21] **Lazebnik, C.S.S. and J. Ponce,** Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, New York, 2006.

[22] **Sivic, J. and A. Zisserman,** Video Google: A text retrieval approach to object matching in videos. *Ninth IEEE international conference on computer vision*, Beijing, 2003, 1470–1477.

[23] **van de Sande, K.E.A., T. Gevers and C.G.M. Snoek,** Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32(9)** (2010), 1582–1596.

[24] **Xu, L. and M.I. Jordan,** On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, **8(1)** (1996), 129–151.

B. Daróczy
Institute of Computer Science and Control
Hungarian Academy of Sciences (MTA SZTAKI)
and Eötvös University
Budapest
Hungary
daroczy.balint@sztaki.mta.hu

A. A. Benczúr
Institute of Computer Science and Control
Hungarian Academy of Sciences (MTA SZTAKI)
and Eötvös University
Budapest
Hungary
benczur@sztaki.mta.hu

L. Rónyai
Institute of Computer Science and Control
Hungarian Academy of Sciences (MTA SZTAKI)
and Budapest University of Technology and Economics
Budapest
Hungary
ronyai@sztaki.mta.hu