Global Gaussian Approach for Scene Categorization Using Information Geometry

Hideki Nakayama Tatsuya Harada Yasuo Kuniyoshi
Grad. School of Information Science and Technology, The University of Tokyo
{*nakayama, harada, kuniyosh*}*@isi.imi.i.u-tokyo.ac.jp*

Abstract

Local features provide powerful cues for generic image recognition. An image is represented by a "bag" of local features, which form a probabilistic distribution in the feature space. The problem is how to exploit the distributions efficiently. One of the most successful approaches is the bag-of-keypoints scheme, which can be interpreted as sparse sampling of high-level statistics, in the sense that it describes a complex structure of a local feature distribution using a relatively small number of parameters. In this paper, we propose the opposite approach, dense sampling of low-level statistics. A distribution is represented by a Gaussian in the entire feature space. We define some similarity measures of the distributions based on an information geometry framework and show how this conceptually simple approach can provide a satisfactory performance, comparable to the bag-of-keypoints for scene classification tasks. Furthermore, because our method and bag-of-keypoints illustrate different statistical points, we can further improve classification performance by using both of them in kernels.

1. Introduction

With the significant advances in computer systems, appearance-based image recognition methods using statistical learning have drawn increasing attention. In particular, automatic scene and object categorization has been one of the most important challenges, and has seen substantial progress in the last decade. One notable breakthrough is the advances in local feature description, which provides a powerful cue to represent the semantics of images. In this approach, a bunch of local features, typically SIFT [15] or SURF [4], are extracted from local sub-windows of an image. These local features form a probabilistic distribution in the feature space, which is expected to contain rich information related to the local and global image structures. Therefore, the basic question is: How to efficiently exploit the distribution of local features?

The most well-known and practical example of local approach methods is the bag-of-keypoints scheme (bag-of-visual-words) [8]. The first step of this method is to perform vector quantization of the local features of training images using clustering algorithms to obtain centroids, which represent the visual words. The resulting feature is the histogram of visual word occurrences in the image. This method can be interpreted as a sparse sampling of high-level statistics of local feature distributions, in the sense that it describes the complex structure of a distribution using some small number of representative parameters (# of visual words). High-level statistics are suited for detecting locally distinctive patterns in an image, which are thought to be especially important for capturing the characteristics of solid objects.

In this work, we focus on the opposite approach: dense sampling of low-level statistics. We simply model a local feature distribution of each image as a Gaussian, which we call the global Gaussian (GG) approach. This is equivalent to sampling all zeroth- and first-order statistical moments. While the bag-of-keypoints (BoK) basically captures the local frequency in the feature space, a Gaussian provides fundamental global information. This is thought to be suitable for describing abstract scenes, where solid distinctive cues are not always available. Following the information geometry approach [3], we derive theoretically supported similarity measures for Gaussians, and then apply them to kernel functions for training a classifier. Using an information geometry technique, we can derive both an optimal kernel and an efficient linear approximation.

Although some previous studies are closely related to ours [18, 26], a Gaussian based approach has not always received enough attention compared to BoK in the field of scene recognition. In experiments of scene categorization, using three state-of-the-art datasets, we show that the global Gaussian approach achieves promising results well comparable to that of bag-of-keypoints. Furthermore, because our method and BoK illustrate different statistical points, we can further improve classification performance by using both of them in kernels. Overall, our contributions are:

978-1-4244-6983-3/10/$26.00 ©2010 IEEE

Table 1. Summary of previous work and our work from the viewpoint of local feature statistics.

	High-level	Low-level
Dense	Non-parametric [5] GMM [18, 30]	Covariance [25, 26] Ours (Gaussian)
Sparse	Bag-of-keypoints [8]	

- A thorough experiment with the global Gaussian approach in the context of scene categorization.
- Theoretically deriving optimal kernel metrics and a scalable linear approximation using the information geometry technique.
- Proposing the idea of combining the global Gaussian and BoK techniques and demonstrating the effectiveness of that combination.

2. Related Work

Generally, the local features of an image are generated from a complex hidden distribution. However, the number of statistically independent samples that can be extracted from one image is severely restricted. Therefore, estimating the distribution is an extremely difficult task.

With this in mind, Table 1 summarizes the approaches of both previous work and ours. A straightforward approach is to use raw local features for matching images. Boiman *et al*. [5] proposed the Naive-Bayes nearest neighbor (NBNN) classification algorithm, which finds the nearest patch in the training corpus for all patches in the query image. This method showed excellent performance, probably because a non-parametric approach can handle the complex structure of real data relatively stably using a limited number of examples. However, the computational cost of this method is immense because they need to preserve all raw local features in the training images for use in classification.

As examples of parametric estimation, Vasconcelos *et al*. exploited a Gaussian mixture model (GMM) for modeling the distribution [18, 28]. To apply their generative model to discrimination, they proposed a kernel function for defining a similarity between two distributions, and used it on a support vector machine (SVM). In addition, Zhou *et al*. [30] estimated a GMM for each image and used its parameters as the appearance feature. These methods are interpreted as sampling of the high-level statistics of local feature distributions. Ideally, this will give an optimal representation of a distribution. However, as we mentioned above, it is nearly impossible to estimate a large-scale GMM using local features sampled from each individual image. Therefore, [30] applied a hierarchical estimation of GMM, where a distribution of each image is estimated as a deviation from the entire training corpus. This approach, however, cannot always provide an effective representation for each image because the estimation will be severely affected by the characteristics of the training corpus.

A similar problem occurs in the codebook generation in the bag-of-keypoints (bag-of-visual-words) scheme [8], because a standard k-means tends to place its clusters around the densest regions in the training corpus. Many recent works are dedicated to this problem. For example, Jurie *et al*. [12] exploited a radius-based mean-shift clustering to generate a more appropriate codebook. Wu *et al*. [29] showed that a histogram intersection is generally a better metric for clustering local features. Moreover, Tuytelaars *et al*. [24] presented a lattice-based vector quantization instead of a data-driven approach. Further, there are many studies focused on improving BoK related to other aspects. For example, the soft assignment strategy [22, 27] is shown to create more descriptive visual word histograms.

Probably, the study by Tuzel *et al*. [26] is the closest to ours. They extracted a covariance matrix of the local features of an image, and described it as a point on a Riemannian manifold. Further, they performed LogitBoost learning using the structure of the manifold by means of differential geometry. They achieved excellent performance with a human detection task. This method can be interpreted as using the shape of a Gaussian for describing an image. Covariances are typical examples of low-level statistics and are expected to be relatively stable, although they are sampled from each image independently. However, an obvious problem is that they lose the mean information. That is, two Gaussians at different points having similar shapes are indistinguishable. Moreover, because our method is based on a "flat" manifold, we can effectively exploit the structure of tangent spaces.

3. Our Approach

In this study, an image is represented as a probability distribution of its local features. Suppose a bag of d-dimensional local features $\{\boldsymbol{x}_k\}$ are extracted from an image I_j. Then I_j can be explained with the distribution $p_j(\boldsymbol{x};\boldsymbol{\theta}(j))$ having $\boldsymbol{\theta}(j)$ as the parameters. We plot each sample on a flat Riemannian manifold using the information geometry technique. We derive some theoretically supported similarity metrics on the manifold and use them for kernel functions so that they can be applicable to discrimination. As a natural result, it is shown that a theoretically optimal kernel is the one based on the Kullback-Leibler (KL) divergence. Basically, this kernel becomes the same one used in [18], and is expected to provide the upper limit performance of our global Gaussian approach. However, the scalability of a KL divergence based method is low because it requires high-cost nonlinear computation. Therefore, we also derive an approximate linear kernel based on the Riemannian metric.

3.1. Brief Summary of Information Geometry

Information geometry, which is based on differential geometry, began as the geometric study of statistical estimation [3]. It expresses the model space of a certain family of parametric probability functions as a Riemannian manifold. Each sample, which constitutes a probabilistic distribution, is represented as a point on the manifold. Let us consider the manifold S formed with a probabilistic model $p(\boldsymbol{x};\boldsymbol{\theta})$ having n-dimensional parameters $\boldsymbol{\theta} = (\theta^1, ..., \theta^n)$. An information geometry framework gives a statistically natural structure to the manifold. First, we exploit a Fisher information matrix as a Riemannian metric.

$$G^{\theta}_{lm}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta^l}\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta^m}\right]. \qquad (1)$$

Next, we apply a symmetric connection called an α-connection [1], having α as a parameter to determine the structure of the manifold. For some special probabilistic models, we find a flat manifold by taking an appropriate affine coordinate system ξ, where tangent spaces are flatly connected. If such a coordinate ξ exists, the model space is defined as α-flat, and ξ is defined as the α-affine coordinate system. In an α-flat space, a geodesic is represented as a line on an α-coordinate system (α-geodesic). It is known that an α-flat space is always $-\alpha$-flat and that we can take another affine coordinate system that is dual to ξ. As discussed in more detail below, $\alpha = \pm 1$ becomes especially important in information geometry [2]. Actually, it is known that there exist ± 1-coordinate systems for many practical probabilistic models which are used widely for statistical learning. Therefore, information geometry has been successfully applied to the analysis and interpretation of many kinds of learning methods such as EM algorithm [2], boosting [19] and variational Bayes [10]. For further details, refer to [3].

The exponential family is among the most basic and important probabilistic models for practical applications. It also plays an important role in the information geometry framework. A distribution of the exponential family is represented as follows:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{n}\theta^i F_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + C(\boldsymbol{x})\right). \qquad (2)$$

Here, $\boldsymbol{\theta}$ is the model parameter, F is a function of the observed variable $\boldsymbol{x}$, $\psi(\boldsymbol{\theta})$ is the potential function, and $C(\boldsymbol{x})$ is a constant function independent of $\boldsymbol{\theta}$. The exponential family is 1-flat, taking $\boldsymbol{\theta}$ as the corresponding affine coordinate system. We can take another affine coordinate system $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ which is dual to $\boldsymbol{\theta}$ and is defined as $\eta_i = E_{\boldsymbol{\theta}}[F_i(\boldsymbol{x})]$. The $\boldsymbol{\eta}$-coordinate system is interpreted as the space of sufficient statistics and is -1-flat. The Riemannian metric of the $\boldsymbol{\eta}$-coordinate system becomes the inverse of that of the $\boldsymbol{\theta}$-coordinate system (G^{θ}, Eq. 1). This can be explicitly described using the following conversion.

$$G^{\eta}_{lm} = \frac{\partial \theta^l}{\partial \eta_m}. \qquad (3)$$

3.2. Gaussian Embedding

A Gaussian also belongs to the exponential family and is described by $n = d + d(d+1)/2$ parameters. Let μ and Σ denote the sample mean and covariance respectively. Letting

$$\begin{aligned} C(\boldsymbol{x}) = 0,\;\; F_i(\boldsymbol{x}) = x_i,\;\; F_{ij}(\boldsymbol{x}) = x_i x_j \; (i \le j),\\ \theta^i = \sum_{j=1}^{d}(\Sigma^{-1})_{ij}\mu_j,\quad \theta^{ii} = -\frac{1}{2}(\Sigma^{-1})_{ii},\\ \theta^{ij} = -(\Sigma^{-1})_{ij}\; (i < j), \end{aligned} \qquad (4)$$

then, a Gaussian is represented as follows:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left[\sum_{1\le i\le d}\theta^i F_i(x) + \sum_{1\le i\le j\le d}\theta^{ij}F_{ij}(x) - \psi(\boldsymbol{\theta})\right]. \qquad (5)$$

Here,

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\log(2\pi)^d|\Sigma|. \qquad (6)$$

Also, the $\boldsymbol{\eta}$-coordinates become as follows:

$$\eta_i = \mu_i, \quad \eta_{ij} = \Sigma_{ij} + \mu_i\mu_j \; (i \le j). \qquad (7)$$

The $\boldsymbol{\theta}$-coordinate is based on model parameters and the $\boldsymbol{\eta}$-coordinate is based on sufficient statistics. In an ideal situation where we can obtain perfect information from samples, we may take any of them for the image feature space. However, usually we have only a limited amount of observations (local features) for each sample (an image). Therefore, we take the estimated sufficient statistics from the observations and plot each sample on the $\boldsymbol{\eta}$-coordinates. Let $\boldsymbol{e}_i, \boldsymbol{e}_{ij}$ denote the basis vectors corresponding to η_i and η_{ij} respectively. Then the $\boldsymbol{\eta}$-coordinate system is described as:

$$\begin{aligned} \boldsymbol{\eta} &= \sum_{1\le i\le d}\eta_i\boldsymbol{e}_i + \sum_{1\le i\le j\le d}\eta_{ij}\boldsymbol{e}_{ij}\\ &= (\eta_1, ..., \eta_d, \eta_{11}, ..., \eta_{1d}, \eta_{22}, ...\eta_{2d}, ..., \eta_{dd})^T\\ &= \big(\hat{\mu}_1, ..., \hat{\mu}_d, \hat{\Sigma}_{11} + \hat{\mu}_1^2, ..., \hat{\Sigma}_{1d} + \hat{\mu}_1\hat{\mu}_d,\\ &\qquad \hat{\Sigma}_{22} + \hat{\mu}_2^2, ..., \hat{\Sigma}_{dd} + \hat{\mu}_d^2\big)^T. \end{aligned} \qquad (8)$$

[1] $\alpha = 0$ corresponds to the Levi-Civita connection.

[2] In information geometry, terms such as 1-connection, 1-flat are specifically called e-connection and e-flat (e:exponential), and -1-connection and -1-flat are called m-connection and m-flat (m:mixture). However, we do not change the terminology in this paper for simplicity.

As Eq. 8 shows, the $\boldsymbol{\eta}$-coordinates consist of all means and correlations of the elements of observed local features. The Riemannian metric of the $\boldsymbol{\eta}$-coordinate system is as follows:

$$\begin{aligned}
G^{\eta}_{ij} &= (\Sigma^{-1})_{ij}(1+\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})+ \\
&\qquad \sum\nolimits_{k=1}^{d} \mu_k(\Sigma^{-1})_{ki} \sum\nolimits_{k=1}^{d} \mu_k(\Sigma^{-1})_{kj}, \\
G^{\eta}_{i(pq)} &= -(\Sigma^{-1})_{pi} \sum\nolimits_{k=1}^{d} \mu_k(\Sigma^{-1})_{kq} - \\
&\qquad (\Sigma^{-1})_{qi} \sum\nolimits_{k=1}^{d} \mu_k(\Sigma^{-1})_{kp} \;\; (p<q), \\
G^{\eta}_{i(pp)} &= -(\Sigma^{-1})_{pi} \sum\nolimits_{k=1}^{d} \mu_k(\Sigma^{-1})_{kp} \\
G^{\eta}_{(pq)(rs)} &= (\Sigma^{-1})_{ps}(\Sigma^{-1})_{qr} + (\Sigma^{-1})_{qs}(\Sigma^{-1})_{pr} \\
&\qquad (p<q, r<s), \\
G^{\eta}_{(pq)(rr)} &= (\Sigma^{-1})_{pr}(\Sigma^{-1})_{rq} \;\; (p<q), \\
G^{\eta}_{(pp)(rr)} &= \frac{1}{2}(\Sigma^{-1})^2_{pr}.
\end{aligned} \tag{9}$$

Above, the suffixes correspond to Eq. 8. For example, $G^{\eta}_{i(pq)} = \langle \boldsymbol{e}_i, \boldsymbol{e}_{pq} \rangle$, and $G^{\eta}_{(pq)(rr)} = \langle \boldsymbol{e}_{pq}, \boldsymbol{e}_{rr} \rangle$.

3.3. Kernel Functions

KL divergence based kernel

In information geometry, the α-divergence between two points $P : f(\boldsymbol{x})$, $Q : g(\boldsymbol{x})$ in a dually-flat space is defined as follows:

$$D^{(\alpha)}(P||Q) = \psi\big(\boldsymbol{\theta}(P)\big) + \varphi\big(\boldsymbol{\eta}(Q)\big) - \sum_{i=1}^{n} \theta^i(P)\eta_i(Q). \tag{10}$$

Here, $\varphi(\boldsymbol{\eta})$ is the potential function of the $\boldsymbol{\eta}$-coordinate system. The α-divergence is an important metric for information geometry. Intuitively, it represents the dissimilarity between two points; strictly speaking, it is different from a mathematical distance, because a symmetric property does not hold unless P and Q are sufficiently close. Also, the dual $-\alpha$-divergence becomes $D^{(-\alpha)}(P||Q) = D^{(\alpha)}(Q||P)$. In the case of the exponential family, 1-divergence ($\alpha = 1$) becomes equal to the KL divergence between $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$:

$$k(f||g) = \int f(\boldsymbol{x}) \left[\log f(\boldsymbol{x}) - \log g(\boldsymbol{x})\right] d\boldsymbol{x}. \tag{11}$$

Also, the dual -1-divergence ($\alpha = -1$) is equal to $k(g||f)$.

Since we take the -1-flat $\boldsymbol{\eta}$-coordinate system, we consider -1-divergence. However, since this is an asymmetric metric, we cannot use it directly as a kernel function. Therefore, we define a distance between two samples by symmetrizing the divergence following the approach of [18].

$$\begin{aligned}
&dist\big(\boldsymbol{\eta}(P), \boldsymbol{\eta}(Q)\big) \\
&= D^{(-1)}(P||Q) + D^{(-1)}(Q||P) \\
&= k(g||f) + k(f||g) \\
&= tr(\Sigma_P\Sigma_Q^{-1}) + tr(\Sigma_Q\Sigma_P^{-1}) - 2d + \\
&\qquad tr\left((\Sigma_P^{-1} + \Sigma_Q^{-1})(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)^T\right).
\end{aligned} \tag{12}$$

To define a kernel that satisfies the Mercer conditions, we simply exponentiate the distance following [18]:

$$K_{kl}(P,Q) = \exp\big(-a\; dist\big(\boldsymbol{\eta}(P), \boldsymbol{\eta}(Q)\big)\big). \tag{13}$$

Above, a is a smoothing parameter. KL divergence requires computing the inverse of a covariance matrix, which can be unstable when only a small number of features are available. Therefore, we add a regularization matrix to the covariance matrices for improving numerical stability. That is, we let $\Sigma \rightarrow \Sigma + bI$. This process is equivalent to adding artificial white noise to local features.

Ad-hoc linear kernel

First, as the simplest baseline of linear approximation, we simply apply a linear kernel to the $\boldsymbol{\eta}$-coordinate system. This is a strong approximation that ignores the manifold metric, and will be severely affected by the nature of the local descriptors and scaling effects. We call this the ad-hoc linear kernel (ad-linear).

$$K_{ad}(P,Q) = \boldsymbol{\eta}(P)^T\boldsymbol{\eta}(Q). \tag{14}$$

Center tangent linear kernel

For a more strict formulization, we need to exploit the Riemannian metric in Eq. 9, which takes different values at each point of the $\boldsymbol{\eta}$-coordinate. Suppose we have N training images. Following some previous work [1], we take the metric on the mean of them, $\boldsymbol{\eta}_c = \frac{1}{N}\sum_i^N \boldsymbol{\eta}(i)$ for approximation.

$$K_{ct}(P,Q) = \boldsymbol{\eta}(P)^T G^{\eta}(\boldsymbol{\eta}_c)\boldsymbol{\eta}(Q). \tag{15}$$

Here, $G^{\eta}(\boldsymbol{\eta}_c)$ is the metric on $\boldsymbol{\eta}_c$. This process is interpreted as approximating the model space using the tangent space of $\boldsymbol{\eta}_c$. $\boldsymbol{\eta}_c$ corresponds to the Gaussian distribution estimated from the local features of the entire training corpus, which is a reasonably representative point. We call this the center tangent linear kernel (ct-linear). The ct-linear kernel can be efficiently computed by applying a normal linear kernel to the transformed coordinate system $\boldsymbol{\zeta} = \big(G^{\eta}(\boldsymbol{\eta}_c)\big)^{1/2}\boldsymbol{\eta}$. Therefore, we can substantially improve the performance from the ad-hoc linear kernel without losing scalability.

4. Classification Method

In this work, we employ two classification methods. The first is the SVM, which is a common tool for classification in recent work on generic image recognition. The other is the probabilistic discriminant analysis (PDA) [11], a probabilistic interpretation of the classical linear DA. A benefit of PDA is that we can build a multiclass classifier by solving an eigenvalue problem only once, while SVM needs a fusion of binary classifiers. In both SVM and PDA, we apply kernel functions to cope with non-linear structures. For SVM implementation, we use LIBSVM [7].

Below, we introduce the kernel DA, which is the core of the PDA, and the classification rule provided by the PDA framework. Suppose a kernel function $K(\boldsymbol{\eta}(i), \boldsymbol{\eta}(j)) = \langle \phi(\boldsymbol{\eta}(i)), \phi(\boldsymbol{\eta}(j)) \rangle$ is given, where $\phi : \boldsymbol{\eta} \rightarrow \phi(\boldsymbol{\eta})$ denotes the projection that maps an input vector on a high-dimensional feature space. We let N denote the number of training samples, $\boldsymbol{\eta}^K = \left(K(\boldsymbol{\eta}, \boldsymbol{\eta}(1)), ..., K(\boldsymbol{\eta}, \boldsymbol{\eta}(N))\right)^T$ denote the kernel base vector, Σ_w^K denote the within-class covariance matrix of kernel base vectors, and Σ_b^K denote the between-class covariance matrix. The kernel DA is formulated as the following generalized eigenvalue problem.

$$\Sigma_b^K W = \acute{\Sigma}_w^K W \Lambda \quad (W^T \acute{\Sigma}_w^K W = I). \tag{16}$$

Here, $\acute{\Sigma}_w^K = \Sigma_w^K + \gamma I$. γ is a small positive number used to determine the amplitude of the regularization matrix, which is used to control generalization. Λ is a diagonal matrix having eigenvalues as the elements.

Kernel DA is interpreted as performing linear DA on an implicit high-dimensional space using the kernel trick. Therefore, we can exploit the classification rule using the structure of a latent subspace provided by a probabilistic linear DA framework [11]. Let t denote the number of samples in each class, and $\boldsymbol{\mu}_\eta^K$ denote the mean of kernel base vectors over the entire training dataset. The following projection maps an image feature $\boldsymbol{\eta}$ to a point in the latent space:

$$\boldsymbol{u} = \left(\frac{t-1}{t}\right)^{1/2} W^T(\boldsymbol{\eta}^K - \boldsymbol{\mu}_\eta^K). \tag{17}$$

The covariance of the latent values is given by the following expression:

$$\Psi = max\left(0, \frac{t-1}{t}\Lambda - \frac{1}{t}\right). \tag{18}$$

Using this structure, we classify a newly input sample $\boldsymbol{\eta}_s^K$ by the maximum likelihood estimation. We assume that $\boldsymbol{u}_s$, the projected point of $\boldsymbol{\eta}_s^K$, is generated from a certain class C with probability:

$$p(\boldsymbol{u}_s | \boldsymbol{u}_{1...t}^C) = \mathcal{N}\left(\boldsymbol{u}_s | \frac{t\Psi}{t\Psi + I}\bar{\boldsymbol{u}}^C, I + \frac{\Psi}{t\Psi + I}\right). \tag{19}$$

Here, $\boldsymbol{u}_{1...t}^C$ are latent values of n independent training samples that belong to class C, and $\bar{\boldsymbol{u}}^C$ is their mean. We classify $\boldsymbol{\eta}_s^K$ in the class with the largest value of Eq. 19. This is an extremely simple process similar to the nearest-centroid approach.

5. Implementation

5.1. Local Feature Sampling

We use the SIFT [15] descriptor (128-dim) and SURF [4] descriptor (64-dim) as the local feature descriptors. Mikolajczyk *et al*. [17] showed that the SIFT descriptor has averagely the best performance among local feature descriptors. Also, SURF is known as a powerful descriptor comparable to SIFT, although its computational cost is substantially reduced. As the feature sampling method, we take the dense sampling strategy, following many successful works related to image categorization [9, 20, 6, 29, 30]. We space the keypoints five pixels apart, and extract local features from each patch of 16×16 pixels having the keypoint at the center. Note that we extract local features from gray images in all experiments, even if color images are available.

5.2. Use of Spatial Information

We incorporate spatial information of images into our kernels following the standard spatial pyramid kernel [13]. We hierarchically partition images into grids using the zeroth layer (original image) to the L-th layer. Each l-th layer $(0 \leq l \leq L)$ is partitioned into a $2^l \times 2^l$ grid. Then, we generate the local $\boldsymbol{\eta}$-coordinate system independently for each region and compute kernels such as K_{kl} or K_{ct}. Finally, they are merged as follows:

$$K^{GG}(P, Q) = \frac{1}{\sum_{i=0}^{L} \beta^i} \sum_{l=0}^{L} \frac{\beta^l}{2^{2l}} \sum_{k=1}^{2^{2l}} K^{(l,k)}(P, Q). \tag{20}$$

Here, $\beta \in \mathcal{R}$ is a relative weight parameter of layers. The suffix (l, k) indicates that the element belongs to the k-th region of the l-th layer.

As for the implementation of K_{ct} kernel, since computing the metric for each region is expensive, we simply use the one from $L = 0$ for all regions.

5.3. Bag-of-Keypoints Baseline

To provide a quantitative baseline, we implement the BoK method using the same local features sampled for the proposed method. We use the standard k-means method for generating a codebook. We set the number of visual words to 200 and 1000. For training classifiers, we use a histogram intersection kernel and apply the spatial pyramid matching [13]. We let K^{BoK} denote this kernel function in the following.

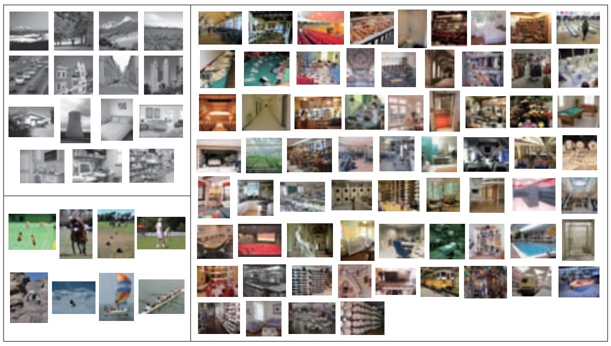

Figure 1. Images from benchmark datasets. Top left: LSP15 [13]. Bottom left: 8-sports [14]. Right: Indoor67 [23].

In some experiments, we merge our proposed kernels (Eq. 20) and those for BoK for further performance improvements. Here, we simply exploit a linear combination.

$$K^{GG+BoK} = \frac{1}{1+\kappa} K^{GG} + \frac{\kappa}{1+\kappa} K^{BoK}. \tag{21}$$

Here, κ is a weight parameter.

6. Experiments

6.1. Datasets

We experiment with three challenging datasets: 15 class scene dataset provided by Lazebnik *et al*. [13] (LSP15), eight class sports events dataset provided by Li *et al*. [14] (8-sports), and a 67 class indoor scene dataset by Quattoni *et al*. [23] (Indoor67). Figure 1 shows some images from each dataset.

LSP15 has been the standard benchmark for scene classification tasks. It consists of 10 outdoor and five indoor classes. The 8-sports dataset has both scene recognition and object recognition aspects. Images in this dataset are characterized by background scenes and foreground athletes. Indoor67 is a new scene dataset published in 2009, and is the largest scene dataset currently available. It is characterized by a large number of classes and their high intra-class variations. Also, it is pointed out in [23] that indoor scene categorization is more difficult than natural scene categorization.

We follow standard experimental protocols used in previous work. In LSP15, we randomly choose 100 training samples for each class and use the remaining samples for testing. Also, we randomly choose 70 training and 60 testing samples in 8-sports, and 80 training and 20 testing samples in Indoor67. Performance is evaluated in terms of the mean of the classification rate of each class [3]. This score is averaged over many trials, wherein the training and testing samples are replaced randomly. In all experiments in this study, we take the average over 10 trials.

Table 2. Basic results of the global Gaussian approach with the LSP15 and 8-sports datasets using different kernels (%). No spatial information is used here.

		LSP15		8-sports	
		SIFT	SURF	SIFT	SURF
PDA	ad-linear	77.3	75.9	77.9	72.4
	ct-linear	78.8	78.5	79.7	78.1
	KL div.	80.4	81.5	81.7	79.6
SVM	ad-linear	69.9	72.1	70.6	70.2
	ct-linear	75.7	77.7	75.5	73.3
	KL div.	76.3	78.3	78.3	74.9

Table 3. Performance comparison with spatial information on LSP15 (%). The SURF descriptor is used.

		L=0	L=1	L=2
GG	PDA (ad-linear)	75.9	78.8	79.8
	PDA (ct-linear)	78.5	81.6	82.3
	PDA (KL div.)	81.5	84.8	86.1
	SVM (ad-linear)	72.1	73.2	74.3
	SVM (ct-linear)	77.7	80.1	80.7
	SVM (KL div.)	78.3	82.2	83.1
BoK200	PDA	71.9	78.5	81.1
	SVM	70.6	76.3	78.6
BoK1000	PDA	77.1	80.7	82.5
	SVM	74.9	78.0	79.4

Table 4. Performance comparison with spatial information on the 8-sports dataset (%). The SIFT descriptor is used.

		L=0	L=1	L=2
GG	PDA (ad-linear)	77.9	79.3	80.2
	PDA (ct-linear)	79.7	81.5	82.9
	PDA (KL div.)	81.7	83.2	84.4
	SVM (ad-linear)	70.6	71.6	71.7
	SVM (ct-linear)	75.5	77.2	78.8
	SVM (KL div.)	78.3	80.2	81.4
BoK200	PDA	72.0	76.9	79.6
	SVM	71.7	76.3	77.7
BoK1000	PDA	77.8	80.6	81.5
	SVM	76.2	78.1	79.1

6.2. In-depth study with LSP15 and 8-sport datasets

First, we investigate the effectiveness of our approach using LSP15 and 8-sports datasets. Table 2 shows the basic performance without the use of spatial information. We test both SIFT and SURF descriptors. The notation "ad-linear" denotes the ad-hoc linear kernel, "ct-linear" denotes

[3] Average of diagonal elements of the confusion matrix.

Table 5. Performances of global Gaussian, BoK, and combined approach (%). $L = 2$ spatial pyramid is implemented. Kernel PDA is used for classification. SURF descriptor is used for LSP15 and SIFT descriptor is used for 8-sports.

	LSP15	8-sports
GG (KL)	86.1±0.5	84.4±1.4
GG (ct-linear)	82.3±0.4	82.9±1.0
BoK200	81.1±0.7	79.6±1.1
BoK1000	82.5±0.7	81.5±1.7
GG (ct-linear) + BoK200	85.0±0.5	83.2±0.9
GG (ct-linear) + BoK1000	85.3±0.5	83.4±0.7

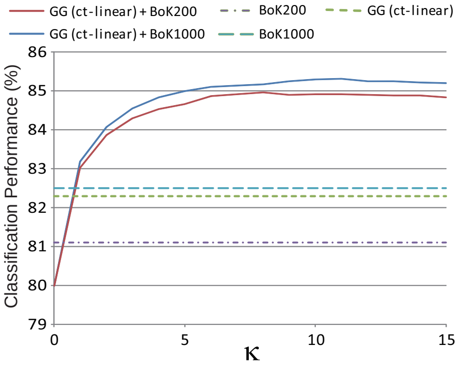

Figure 2. Merging the global Gaussian and BoK with the LSP15 dataset. κ is the parameter for weighting kernels (Eq. 21).

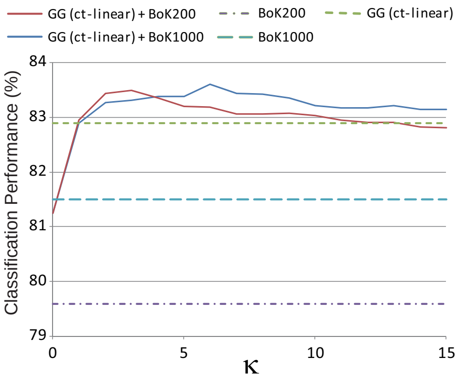

Figure 3. Merging the global Gaussian and BoK with the 8-sports dataset. κ is the parameter for weighting kernels (Eq. 21).

the center tangent linear kernel, and “KL div.” denotes the KL divergence based kernel. As shown, the KL divergence based kernel achieves the best performance, followed by ct-linear and ad-linear. The ct-linear kernel substantially improves performance compared to the ad-linear kernel, and PDA obtains better performance than that of SVM (LIBSVM). In addition, the result shows that SURF is superior in LSP15 and SIFT is superior in 8-sports.

Next, we investigate the effect of spatial information in our method. Here, we implement BoK to provide a baseline. In both our method and BoK, we use spatial pyramids up to $L = 2$. We use SURF for LSP15 and SIFT for 8-sports here. Table 3 shows the results with LSP15, and Table 4 shows the results with 8-sports. Our method obtains a satisfactory result that compares well with BoK using 1000 visual words. Also, the result show that spatial information can reasonably improve the performance of our method.

Table 6. Performance comparison with previous work (%). For our method, $L = 2$ spatial pyramid is implemented, and kernel PDA is used for classification. We use the SURF descriptor for LSP15 and Indoor67, and the SIFT descriptor for 8-sports.

Method	LSP15	8-sports	Indoor67
GG (KL-div.)	86.1±0.5	84.4±1.4	45.5±1.1
GG (ct-linear) + BoK1000	85.3±0.5	83.4±0.7	44.9±1.3
Previous	85.2 [30] 84.1 [29] 83.7 [6]	84.2 [29] 73.4 [14]	25.0 [23]

Finally, we try to merge our global Gaussian approach and BoK. Although the KL divergence based kernel achieves a high performance, it is not suitable for practical systems because of its low scalability. Therefore, here we combine the ct-linear kernel of our method and histogram intersection kernel of BoK method as Eq. 21. Table 5 shows that we can further improve the performance by concatenating different statistics of local features provided by Gaussian and BoK. Figure 2 and 3 show the effect of weighting parameter κ. This approach is expected to be feasible in a perfectly linear framework by further incorporating the linear approximation technique of the histogram intersection kernel [16].

6.3. Comparison with previous work

We compare the performance of our approach with that of previous work with LSP15, 8-sports and Indoor67. Table 6 summarizes the best performance of our method and that of previous work. For LSP15, hierarchical Gaussianization [30], which is a GMM-based method, achieved the current best score of 85.2%. Our best score using the KL divergence based kernel is 86.1%. The performance of a more scalable ct-linear + BoK technique is reasonably close at 85.3%. For 8-sports, HIK-codebook [29] achieved 84.2% and we get the slightly better score of 84.4%. Note that [29] improved the performance by sampling local features from an original image and Sobel image at five different scales, while we only extract features from a single scale original image. (Without Sobel images, [29] achieved 81.9%.) In Indoor67, the original work [23] achieved an accuracy of 25.0% by concatenating global description with the Gist descriptor [21] and ROI detection using BoK. We use the SURF descriptor for Indoor67, motivated by its promising performance with the LSP15 scene dataset. Our best score is 45.5%, which is superior to [23] by a large margin. However, the result also shows that we still have a long way to go with this challenging large scale dataset.

7. Conclusion

This study focused attention on the low-level statistics of local feature distributions and applied them to scene categorization. Compared to high-level statistics, low-level statistics can be extracted stably and densely from a single image and can be a powerful discriminative cue. In the proposed approach, we express an image sample with a Gaussian distribution of its local features, and derive some kernel metrics theoretically supported by the information geometry framework. Also, we proposed a scalable linear approximation approach exploiting a tangent space on the Riemannian manifold. A thorough study using three challenging datasets showed the effectiveness of our approach. In addition, we showed that we can further improve the performance by merging our approach with BoK, which can efficiently extract some high-level statistics.

References

[1] S. Akaho. The e-PCA and m-PCA: Dimension reduction of parameters by information geometry. In *Proc. IJCNN*, 2004.

[2] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8, 1995.

[3] S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

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

[6] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[7] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[8] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[9] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE CVPR*, pages 524–531, 2005.

[10] S. Ikeda, T. Tanaka, and S. Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16:1779–1810, 2004.

[11] S. Ioffe. Probabilistic linear discriminant analysis. In *Proc. ECCV*, pages 531–542, 2006.

[12] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. IEEE ICCV*, 2005.

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

[14] L.-J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *Proc. IEEE ICCV*, 2007.

[15] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE ICCV*, pages 1150–1157, 1999.

[16] S. Maji and A. C. Berg. Max-margin additive classifiers for detection. In *Proc. IEEE ICCV*, pages 40–47, 2009.

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[18] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proc. NIPS*, 2003.

[19] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Computation*, 16:1437–1481, 2004.

[20] E. Nowak, F. Jurie, and B. Trigges. Sampling strategies for bag-of-features image classification. In *Proc. ECCV*, pages 490–503, 2006.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[22] F. Perronnin, C. R. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *Proc. ECCV*, pages 464–475, 2006.

[23] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proc. IEEE CVPR*, 2009.

[24] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *Proc. IEEE ICCV*, 2007.

[25] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. ECCV*, pages 589–600, 2006.

[26] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. In *Proc. IEEE CVPR*, 2007.

[27] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proc. ECCV*, pages 696–709, 2008.

[28] N. Vasconcelos, P. P. Ho, and P. J. Moreno. The Kullback-Leibler kernel as a framework for discriminant and localized representations for visual recognition. In *Proc. ECCV*, 2004.

[29] J. Wu and J. M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Proc. IEEE ICCV*, pages 630–637, 2009.

[30] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang. Hierarchical Gaussianization for image classification. In *Proc. IEEE ICCV*, pages 1971–1977, 2009.