Dirichlet-based Histogram Feature Transform for Image Classification

Takumi Kobayashi
National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, Japan
takumi.kobayashi@aist.go.jp

Abstract

Histogram-based features have significantly contributed to recent development of image classifications, such as by SIFT local descriptors. In this paper, we propose a method to efficiently transform those histogram features for improving the classification performance. The (L_1-normalized) histogram feature is regarded as a probability mass function, which is modeled by Dirichlet distribution. Based on the probabilistic modeling, we induce the Dirichlet Fisher kernel for transforming the histogram feature vector. The method works on the individual histogram feature to enhance the discriminative power at a low computational cost. On the other hand, in the bag-of-feature (BoF) framework, the Dirichlet mixture model can be extended to Gaussian mixture by transforming histogram-based local descriptors, e.g.*, SIFT, and thereby we propose the method of Dirichlet-derived GMM Fisher kernel. In the experiments on diverse image classification tasks including recognition of subordinate objects and material textures, the proposed methods improve the performance of the histogram-based features and BoF-based Fisher kernel, being favorably competitive with the state-of-the-arts.*

1. Introduction

Histogram-based features are fundamental and play a key role for recently developed methods especially for image classification. For example, SIFT features are widely used as a local descriptor [29], HOG features are for object detection [6] and the bag-of-word (BoW) methods, *i.e.*, visual word histograms, encourage the image classification in the last decade [5]. In the bag-of-feature (BoF) framework, the SIFT local descriptors are also fed into the Fisher kernel [36] to significantly improve the performance.

The histogram captures the statistical feature basically by counting the certain types of symbol. Besides, it is also used to measure the significance of those symbols by voting *uncountable* weights. In SIFT/HOG, the histogram of gradient orientations is constructed by voting the weight derived from the gradient magnitude into the orientation bins. And, soft weights are effectively employed to describe (quantize) a continuous input space in the form of the histogram, such as in soft coding for BoW [40]. Such soft weighting degrades the nature of countable histogram, while BoW in document texts is inherently a histogram of countable words. Thus, the histogram-based feature requires appropriate transformation such as normalization to form an effective feature vector for image classifications.

In order to transform the feature into the form favorable for classification, the statistical methods such as PCA, ICA and LDA have been widely applied. Those methods produce the projection of the feature vectors into a (sub)space based on the statistical criterion which is specific to the objective task. On the other hand, various types of normalization are also applicable to the histogram-based features for the feature transform. Those methods generally work on the features apart from the tasks. The L_1 normalization $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}$ provides histogram-based features with probabilistic perspective, that is, the L_1-normalized histogram is regarded as the probability mass function over the histogram bins. For general image features, the L_2 normalization $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$ is one of the most frequently used feature transform and in recent years, L_2-Hellinger normalization $\frac{\sqrt{\boldsymbol{x}}}{\|\sqrt{\boldsymbol{x}}\|_2} = \sqrt{\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}}$ is intensively applied [1, 36].[1] In this work, we focus on the latter feature transform that is generally applicable to histogram-based features, followed by the classification methods specific to the recognition tasks.

The feature transform is also addressed in a form of a kernel function for kernel-based methods [37]. While Gaussian kernel is generally applied to feature vectors, there are some kernels specialized for the histogram-based features; *e.g.*, χ^2 kernel [47] and intersection kernel [23]. However, the kernel function inevitably requires the kernel-based methods that is computationally expensive. Thus, the kernel feature map is proposed in [41, 31] to circumvent that issue by representing the (additive) kernel in an explicit

[1] The SIFT descriptor with L_2-Hellinger normalization is known as RootSIFT [1].

linear feature form. The mapping enables us to leverage the kernel's discriminative power in the linear features, but it drastically augments the feature dimensionality to approximate the non-linear kernel function, requiring more computational cost compared to the original features.

In this paper, we propose a method to efficiently transform the histogram-based features. From the probabilistic viewpoint, the (L_1-normalized) histogram feature is regarded as a probability mass function, which can be modeled by Dirichlet distribution. Based on the probabilistic model, we induce the Dirichlet Fisher kernel to transform the histogram-based features, enhancing the discriminative power without augmenting the feature dimensionality. Thus, the transformed features are favorably fed into the linear classifier. And, the method does not require cumbersome parameter tuning; instead, it is generally determined based on the statistics of the histogram features, *e.g.*, *natural SIFT statistics*. Note that the proposed method is applicable to any types of histogram-based features including those of countable symbols, *e.g.*, BoW, as well as those constructed by voting weights, *e.g.*, SIFT, whereas the Pólya model [4] only accepts the former type of histograms.

On the other hand, in the BoF framework, a plenty of histogram-based local descriptors, *e.g.*, SIFT, are modeled by the Dirichlet mixture model. We extend it to the Gaussian mixture model via the feature transform and thereby propose the method of Dirichlet-derived GMM Fisher kernel. The method produces effective image features at a low extra computational cost for generic image classification, improving the classification performance.

Our main contributions are 1) Dirichlet Fisher kernel with the parameter determined by natural statistics of the target histogram features, such as *natural SIFT statistics*, 2) Dirichlet-derived GMM Fisher kernel for generic image classifications, and 3) the thorough comparative experiments on a variety of image classification tasks.

2. Dirichlet Fisher kernel

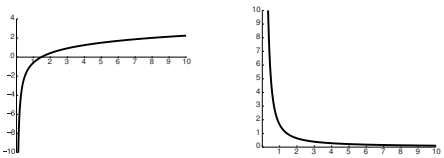
We propose Dirichlet Fisher kernel to transform a histogram feature into a discriminative (linear) form of the same dimensionality, which is suitable for the subsequent linear classifier.

2.1. Definition

The L_1-normalized histogram feature $\boldsymbol{x} \in \mathbb{R}^D$ ($\boldsymbol{x} \geq 0$, $\|\boldsymbol{x}\|_1 = 1$) is supposed to be drawn from the Dirichlet distribution which is defined by

$$\mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\mathtt{B}(\boldsymbol{\theta})}\prod_{i=1}^{D} x_i^{\theta_i - 1}, \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D_+$ is the parameter vector and $\mathtt{B}$ is the beta function, $\mathtt{B}(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{D}\Gamma(\theta_i)}{\Gamma(\theta_0)}$ using the gamma function Γ



(a) digamma ψ (b) trigamma ψ'

Figure 1. Digamma and trigamma functions.

with $\theta_0 = \sum_{i=1}^{D}\theta_i$. Note that the Dirichlet distribution (1) is defined over the simplex domain of the discrete probability distribution $\boldsymbol{x}$ as shown in Fig. 4a. The Fisher kernel [15] of the Dirichlet distribution is given by the derivatives of the log probabilities w.r.t $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}} \log \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) = \underline{\log}(\boldsymbol{x}) - \{\underline{\psi}(\boldsymbol{\theta}) - \psi(\theta_0)\mathbf{1}\} = \underline{\log}(\boldsymbol{x}) - \boldsymbol{\mu}_\theta, \tag{2}$$

where $\mathbf{1} \in \mathbb{R}^D$ is the vector whose components are all 1, ψ is the digamma function (Fig. 1a), and the underlined functions $\underline{\log}$ and $\underline{\psi}$ act on respective components of the input vector. The Fisher information matrix is then written by

$$\begin{aligned}\boldsymbol{H} &= \int \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\log\mathtt{p}(\boldsymbol{x};\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^{\top}\log\mathtt{p}(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} \\ &= \mathtt{diag}\{\underline{\psi'}(\boldsymbol{\theta})\} - \psi'(\theta_0)\mathbf{1}\mathbf{1}^{\top}, \end{aligned} \tag{3}$$

where ψ' indicates the trigamma function which is the first derivative of ψ (Fig. 1b), and $\mathtt{diag}(\cdot)$ produces the diagonal matrix from the input vector. Thus, the Dirichlet Fisher kernel is obtained by the following *linear* feature form;

$$\boldsymbol{H}^{-\frac{1}{2}}\nabla_{\boldsymbol{\theta}}\log\mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{H}^{-\frac{1}{2}}[\underline{\log}(\boldsymbol{x}) - \boldsymbol{\mu}_\theta]. \tag{4}$$

It should be noted that the Dirichlet Fisher kernel results in the same dimensionality D as the original feature $\boldsymbol{x}$.

2.2. Diagonal approximation

In the Dirichlet Fisher kernel (4), it is computationally exhaustive to apply (inverse of) the full Fisher information matrix $\boldsymbol{H}$. Therefore, we approximate $\boldsymbol{H}$ in the following diagonal form;

$$\boldsymbol{H} \approx \mathtt{diag}\big(\underline{\psi'}(\boldsymbol{\theta}) - \psi'(\theta_0)\mathbf{1}\big) = \mathtt{diag}(\boldsymbol{\sigma}_\theta^2). \tag{5}$$

Since the trigamma function $\psi'(\theta_0)$ approaches to zero at larger θ_0 as shown in Fig. 1b, $\psi'(\theta_0)$ is negligible especially in the case of large dimension D, and the Fisher information matrix $\boldsymbol{H}$ is dominated by the diagonal components $\boldsymbol{\sigma}_\theta^2 \triangleq \underline{\psi'}(\boldsymbol{\theta}) - \psi'(\theta_0)\mathbf{1} > 0$. Thus, (4) reduces to

$$\mathtt{diag}(\boldsymbol{\sigma}_\theta)^{-1}[\underline{\log}(\boldsymbol{x}) - \boldsymbol{\mu}_\theta], \tag{6}$$

which consists only of component-wise operations.

2.3. Empirical approximation

In the above formulations (4, 6), the parameter $\boldsymbol{\theta}$ could be estimated from the training samples by the EM method [32]. However, the parameter $\boldsymbol{\theta}$ fortunately appears only in the forms of $\boldsymbol{\mu}_\theta$, $\boldsymbol{\sigma}_\theta$ which are empirically estimated as mean and standard deviation of $\underline{\log}(\boldsymbol{x})$ as follows.

Since $E_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \log \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta})] = \int \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} = \mathbf{0}$, the mean of $\underline{\log}(\boldsymbol{x})$ is given by

$$E_{\boldsymbol{x}}[\underline{\log}(\boldsymbol{x})] = \underline{\psi}(\boldsymbol{\theta}) - \psi(\textstyle\sum_i^D \theta_i)\mathbf{1} = \boldsymbol{\mu}_\theta, \tag{7}$$

and then the diagonal Fisher information matrix results in the variance of $\underline{\log}(\boldsymbol{x})$ as

$$\begin{aligned}\sigma_{\theta i} = H_{ii} &= \int \mathtt{p}(\boldsymbol{x};\boldsymbol{\theta}) \left\{\log(x_i) - E_{\boldsymbol{x}}[\log(x_i)]\right\}^2 d\boldsymbol{x} \\ &= Var_{\boldsymbol{x}}[\log(x_i)]. \end{aligned} \tag{8}$$

Those mean and variance of $\underline{\log}(\boldsymbol{x})$ are empirically estimated from the training samples $\{\boldsymbol{x}_j\}_{j=1}^N$ by

$$\hat{\boldsymbol{\mu}} = E_{\boldsymbol{x}}[\underline{\log}(\boldsymbol{x})] = \frac{1}{N}\sum_{j=1}^N \underline{\log}(\boldsymbol{x}_j), \tag{9}$$

$$\hat{\sigma}_i^2 = Var_{\boldsymbol{x}}[\log(x_i)] = \frac{1}{N}\sum_{j=1}^N \left\{\log(x_{ji}) - \hat{\mu}_i\right\}^2, \tag{10}$$

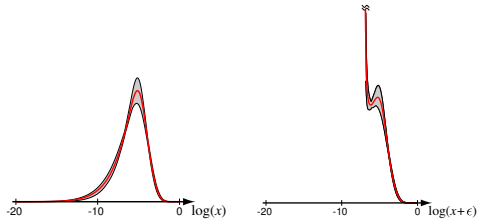
where x_{ji} denotes the i-th component of the j-th sample vector $\boldsymbol{x}_j$. These statistics are efficiently and stably computed in contrast to the model parameter estimation for $\boldsymbol{\theta}$ by the EM method. The Dirichlet Fisher kernel is finally obtained via these empirical approximations (9, 10) by

$$\mathtt{diag}(\hat{\boldsymbol{\sigma}})^{-1}\{\underline{\log}(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}\}. \tag{11}$$

This resultant formulation is quite simple and efficient without augmenting the feature dimensionality unlike the kernel map [41, 31]. This might also be regarded as standardization for $\underline{\log}(\boldsymbol{x})$, not for $\boldsymbol{x}$. It should be noted that this standardization is theoretically derived from the Fisher kernel of the Dirichlet model.

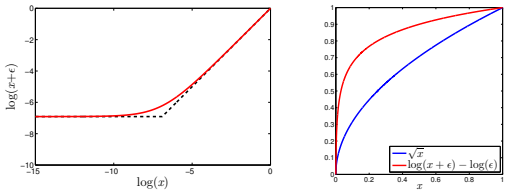
2.4. Modified log-function

The "log" function plays an important role in the Dirichlet Fisher kernel (11). It, however, causes a numerical issue at zero, approaching $-\infty$ as $x \to 0$, and some histogram bins would practically result in empty. To prevent it, $\log(x)$ is usually modified to $\log(x+\epsilon)$ using small fraction parameter $\epsilon \ll 1$. In this section, we address the effect of the fraction ϵ from the viewpoint of probability distribution over the logarithm of the variable, which has not been intensively discussed so far; especially, we mention the *natural SIFT statistics*.



(a) $\log(x)$ (b) $\log(x+\epsilon)$

Figure 2. Empirical distribution of transformed SIFT features marginalized over feature components. The gray region indicates the deviations (between min and max) across the datasets and the red solid curve is the mean distribution.



(a) from $\log(x)$ to $\log(x+\epsilon)$ (b) $\log(x+\epsilon)$ vs $\sqrt{x}$

Figure 3. Transform function. (a) modification of log by introducing ϵ, and (b) comparison to $\sqrt{x}$ with scaling into $[0, 1]$.

The marginal of the Dirichlet distribution is given by the Beta distribution;

$$\mathtt{p}(x;\boldsymbol{\theta}) = \frac{x^{\theta_i - 1}(1-x)^{\theta_0 - \theta_i - 1}}{\mathtt{B}(\theta_i, \theta_0 - \theta_i)} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathtt{B}(\alpha,\beta)} \tag{12}$$

$$= \mathtt{p}(x;\alpha,\beta), \tag{13}$$

where we consider the marginal across all feature components under the assumption that the parameters θ_i are not so drastically changed over i. Then, we transform the variable x into $v = \log(x)$ to modify (13) into

$$\mathtt{p}(v;\alpha,\beta) = \frac{\exp(\alpha v)\{1-\exp(v)\}^{\beta-1}}{\mathtt{B}(\alpha,\beta)}. \tag{14}$$

For example, Fig. 2a shows the empirical distributions of SIFT extracted from various datasets used in Sec.4.3, omitting the features of $x = 0$. We can find that the marginal distributions actually result in almost the same form of the unimodal (modified) Beta distribution regardless of the dataset sources. Thus, this is regarded as the *natural SIFT statistics*, similarly to the natural image statistics [20]. Based on this distribution, we give ϵ in a general form.[2]

Modifying $\log(x)$ into $\log(x+\epsilon)$ corresponds to the variable transform by $\tilde{v} = \log\{\exp(v) + \epsilon\}$ (Fig. 3a), which accordingly reformulates the probability distribution into

$$\mathtt{p}(\tilde{v};\alpha,\beta) = \frac{\{\exp(\tilde{v}) - \epsilon\}^{\alpha-1}\{1+\epsilon-\exp(\tilde{v})\}^{\beta-1}\exp(\tilde{v})}{\mathtt{B}(\alpha,\beta)}. \tag{15}$$

[2]Here, we demonstrate the case of SIFT, but we empirically confirmed that it holds for the other types of histogram features.

The empirical distribution of so transformed SIFT is shown in Fig. 2b. The function $\log(x+\epsilon)$ works in the following two aspects; First, it roughly rounds off the $\log(\theta)$ of $\theta < \epsilon$ into $\log(\epsilon)$ as shown in Fig. 3a. Tiny histogram values are regarded as zeros, exhibiting the negative evidence [16], while the differences of the moderately small x are enhanced like a local contrast enhancement (Fig. 3b). Second, by adequately determining ϵ, $\hat{v} = \log(x+\epsilon)$ is smoothly distributed above $\log(\epsilon)$ (lower bound) while preserving the behavior around the mode of distribution. Too small ϵ pushes the lower bound $\log(\epsilon)$ far away from the mode, which emphasizes the negative evidence too much. On the other hand, larger ϵ unfavorably merges the negative evidence and the mode. The appropriate ϵ renders the smooth distribution between the mode and the lower bound (negative evidence). For that purpose, we determine ϵ based on the cumulative percentile of the empirical distribution; for general setting, we suggest 25% percentile, *e.g.*, producing $\epsilon = \mathsf{P}^{-1}(0.25) \approx 0.001$ for SIFT descriptors where $\mathsf{P}(\epsilon) = \int_0^\epsilon \mathtt{p}(x)dx$.

2.5. Discussion

The Dirichlet Fisher kernel is connected to tf-idf as follows. Roughly speaking, the digamma function $\psi(\theta)$ is approximated by $\log(\theta)$ at large θ[3], which further provides the following approximations:

$$\begin{aligned}\underline{\psi}(\boldsymbol{\theta}) - \psi(\theta_0)\mathbf{1} &\approx \underline{\log}(\boldsymbol{\theta}) - \log(\theta_0)\mathbf{1} = \underline{\log}\left(\frac{\boldsymbol{\theta}}{\theta_0}\right)\\ &= \underline{\log}\big(E_{\boldsymbol{x}}[\boldsymbol{x}]\big) = \underline{\log}(\bar{\boldsymbol{x}}),\\ \psi'(\theta_i) - \psi'(\theta_0) &\approx \frac{1}{\theta_i} - \frac{1}{\theta_0} = \frac{1}{\theta_0}\left(\frac{\theta_0}{\theta_i} - 1\right) = \frac{1}{\theta_0}\left(\frac{1}{\bar{x}_i} - 1\right),\end{aligned}$$

where we use $\bar{\boldsymbol{x}} \triangleq E_{\boldsymbol{x}}[\boldsymbol{x}] = \frac{\boldsymbol{\theta}}{\theta_0}$, the mean of $\boldsymbol{x}$. Thus, the i-th component in the Dirichlet Fisher kernel (6) is approximately reformulated to

$$\frac{1}{\theta_0}\left(\frac{1}{\bar{x}_i} - 1\right)\log\left(\frac{x_i}{\bar{x}_i}\right). \tag{16}$$

This is some sort of tf-idf [35]; apart from the constant $\frac{1}{\theta_0}$, the feature x_i is first normalized by its mean and then weighted by $\frac{1}{\bar{x}_i} - 1$. Those weights emphasize the components of low means $\bar{x}_i$ which correspond to the rarely observed symbols. Such weighting that prefers rare symbols more than common ones is motivated in the same way as tf-idf. Therefore, the Dirichlet Fisher kernel works similarly to tf-idf for enhancing the discriminativity in the features.

The Dirichlet Fisher kernel is also related to the Pólya Fisher kernel which has been first proposed in [7] for text categorization and then presented in [4] for visual classification. The Pólya model accounts for symbol (word) burstiness, which is measured discretely, by means of the compound of Dirichlet and multinomial models [30]. Thereby, it is suitable for transforming the histogram composed of *countable* symbols, *e.g.*, BoW, but is inapplicable to those of uncountable voting weights, *e.g.*, SIFT/HOG. In contrast, the proposed method deals with various types of features into which those countable/uncountable histograms are L_1-normalized. In addition, the Pólya method [7, 4] inevitably requires to learn hyper parameters, which is computationally exhaustive in the case of large-scale high-dimensional histogram features.

[3] $\psi(\theta)$ is usually approximated by $\log(\theta - 0.5)$ for $\theta \geq 1$. However, in order to give light on the connection to tf-idf, we roughly approximate it by $\log(\theta)$.

3. Extension to BoF-based Fisher kernel

In Sec.2, we proposed the Dirichlet Fisher kernel that is applied to individual histogram features. In this section, it is extended in the bag-of-feature framework to the BoF-based Fisher kernel which has exhibited promising performance on image classifications [36].

BoF represents an image by plenty of local descriptors extracted at various points (grid points) with various scales in the image. Suppose we employ histogram-based local descriptors, typically SIFT descriptors, the distribution of which is modeled by the Dirichlet distribution as described in the previous section.

3.1. Dirichlet mixture model (DMM)

It is straightforward to apply the Dirichlet mixture model (DMM) to describe the bag of the local descriptors;

$$\mathtt{p}(\boldsymbol{x}; \{\boldsymbol{\theta}_k\}_{k=1}^K) = \sum_{k=1}^K \omega_k \frac{1}{\mathtt{B}(\boldsymbol{\theta}_k)} \prod_{i=1}^D x_i^{\theta_{ki}-1}, \tag{17}$$

where ω_k are the prior weights, $\sum_k^K \omega_k = 1, \omega_k \geq 0 \,\forall k$. This modeling naturally derives the following DMM Fisher kernel by using (6),

$$\mathcal{G}_k = \frac{1}{\sqrt{N\omega_k}} \sum_i^N \mathtt{p}(k|x_i)\mathtt{diag}(\boldsymbol{\sigma}_{\theta k})^{-1}\big\{\underline{\log}(\boldsymbol{x}_i) - \boldsymbol{\mu}_{\theta k}\big\}, \tag{18}$$

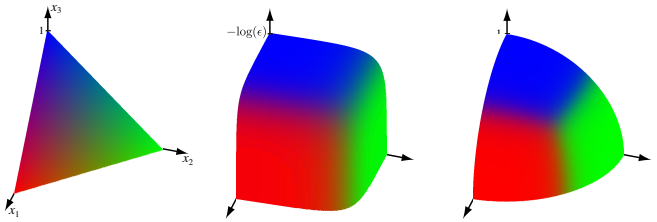
$$\boldsymbol{\mu}_{\theta k} = \underline{\psi}(\boldsymbol{\theta}_k) - \psi(\theta_{k0})\mathbf{1},\ \boldsymbol{\sigma}_{\theta k} = \sqrt{\underline{\psi'}(\boldsymbol{\theta}_k) - \psi'(\theta_{k0})\mathbf{1}}.$$

Note that the k-th part of DMM Fisher kernel $\mathcal{G}_k$ is the D-dimensional vector since the Dirichlet model contains only the parameters $\boldsymbol{\theta}_k \in \mathbb{R}^D$.

3.2. Dirichlet-derived GMM

We rewrite the Dirichlet distribution (1) by using the essential form of $\boldsymbol{v} = \underline{\log}(\boldsymbol{x})$ as

$$\mathtt{p}(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\theta}^\top \boldsymbol{v}\}}{\mathtt{B}(\boldsymbol{\theta})} \approx \frac{\exp\{\boldsymbol{\theta}^\top \tilde{\boldsymbol{v}}\}}{\mathtt{B}(\boldsymbol{\theta})} = \mathtt{p}(\tilde{\boldsymbol{v}}; \boldsymbol{\theta}), \tag{19}$$



(a) $\boldsymbol{x}$ (simplex) (b) $\underline{\log}(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}$ (c) $\frac{\underline{\log}(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}}{\|\underline{\log}(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}\|_2}$

Figure 4. Feature transform. (a) input histogram feature, (b) transformed feature by modified log and (c) transformed feature so as to be suitable for GMM. Pseudo colors show the correspondence among three feature spaces. This figure is best viewed in color.

where $\tilde{\boldsymbol{v}} = \underline{\log}(\boldsymbol{x}+\epsilon)$. Since $\boldsymbol{x}$ lies on the simplex in $\mathbb{R}^D$ (Fig. 4a), its transform $\tilde{\boldsymbol{v}}$ is distributed on the convex surface ($\sum_i e^{\tilde{v}_i} = 1$) as shown in Fig. 4b. We roughly approximate the convex surface by using the sphere surface in the positive orthant space (Fig. 4c). These two surfaces are homeomorphic and furthermore similarly convex. This approximation leads to

$$\mathrm{p}(\tilde{\boldsymbol{v}};\boldsymbol{\theta}) \propto \exp\left[\boldsymbol{\theta}^\top\{\tilde{\boldsymbol{v}}-\log(\epsilon)\mathbf{1}\}\right] \tag{20}$$

$$\approx \exp\left\{\boldsymbol{\eta}^\top \frac{\tilde{\boldsymbol{v}}-\log(\epsilon)\mathbf{1}}{\|\tilde{\boldsymbol{v}}-\log(\epsilon)\mathbf{1}\|_2}\right\}, \tag{21}$$

where $\boldsymbol{\eta}$ is the parameter vector that scales $\boldsymbol{\theta}$ to compensate the difference in the radii of the two convex surfaces. (21) implies the von Mises Fisher distribution [22] w.r.t $\boldsymbol{z} = \frac{\tilde{\boldsymbol{v}}-\log(\epsilon)\mathbf{1}}{\|\tilde{\boldsymbol{v}}-\log(\epsilon)\mathbf{1}\|_2}$ which is often replaced with the Gaussian distribution [12][4]. As a result, we insist that GMM is applicable to model the bag of the transformed local descriptors which are L_2-normalized $\frac{\log(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}}{\|\log(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}\|_2}$, and consequently the Dirichlet-derived GMM Fisher kernel is proposed by applying the GMM Fisher kernel [36] to those transformed local descriptors.

It is advantageous to model the distribution of local descriptors by using Gaussian (GMM) in the following points. First, the von Mises Fisher distribution (21), getting back to Dirichlet distribution, is circularly symmetric around $\boldsymbol{\eta}$ and unable to characterize the anisotropic dispersity orthogonal to $\boldsymbol{\eta}$, while the Gaussian model can exploit it by *variance*; the GMM Fisher kernel [36] produces two types of features related to the *variance* and the *mean*, which doubles the feature dimensionality of the DMM Fisher kernel (18). The DMM Fisher kernel (18) is similar to the proposed Fisher kernel derived from the *mean*,

$$\frac{1}{\sqrt{N\omega_k}}\sum_i^N \mathrm{p}(k|x_i)\mathtt{diag}(\boldsymbol{\sigma}_k)^{-1}\Big[\boldsymbol{z}_i-\boldsymbol{\mu}_k\Big], \tag{22}$$

where $\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k$ are the GMM parameters estimated by applying EM method to the transformed descriptors $\boldsymbol{z} = \frac{\log(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}}{\|\log(\boldsymbol{x}+\epsilon)-\log(\epsilon)\mathbf{1}\|_2}$. Without applying L_2 normalization in $\boldsymbol{z}$, (22) reduces to the DMM Fisher kernel (18). Thus, the DMM Fisher kernel is regarded as the subset of the proposed Dirichlet-derived GMM Fisher kernel.

From the viewpoint of transforming the SIFT descriptors, our method is related to RootSIFT [1] which has also been applied to Fisher kernel [18]. RootSIFT is the SIFT descriptor that is transformed by L_2-Hellinger normalization, $\sqrt{\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}}$. In the RootSIFT, the deviations around the smaller feature values are enhanced by means of the square root, while the proposed transform further enlarges them (Fig. 3b) with pushing the tiny values into the negative evidence as described in Sec.2.4.

4. Experimental results

We evaluate the proposed methods[5] on a variety of image classification tasks by applying the linear SVM classifier. Note that the fraction ϵ is determined based on the 25% percentile of the marginal feature distribution (Sec.2.4).

4.1. Joint histogram of gradient orientation

The proposed Dirichlet Fisher kernel (Sec.2) is applied to transform the gradient local auto-correlation (GLAC) feature [19] for discriminating the pedestrian images on Daimler-Chrysler dataset [33]. The feature is computed by the joint (co-occurrence) histogram of local gradient orientations with uncountable voting weights, to which the Pólya model [4] is inapplicable.

Daimler-Chrysler pedestrian [33]. The task is to classify image patches of 18×36 pixels into a pedestrian (positive) or clutter (negative). For details of the dataset and the evaluation protocol, refer to [33]. The GLAC feature of the setting in [19] is extracted respectively from 2×4 spatial bins over the 18×36 image, resulting in 2,592 dimensions.

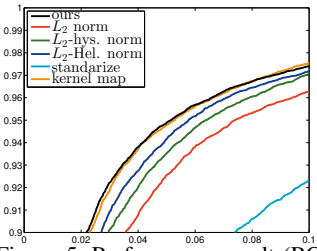
As is the case with Sec.2.4, the empirical feature distributions on the pedestrian and non-pedestrian sets result in almost the same distribution with a small deviations, based on which the fraction ϵ is determined; $\epsilon = \mathrm{P}^{-1}(0.25)$.

The performance results are shown in Fig. 5 with comparison to the other types of feature transforms; L_1 normalization $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1}$, L_2 $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$, L_2-hysteresis $\frac{\min[\boldsymbol{x}/\|\boldsymbol{x}\|_2,\tau]}{\|\min[\boldsymbol{x}/\|\boldsymbol{x}\|_2,\tau]\|_2}$ with $\tau = \frac{1}{\sqrt{D}}$[6], L_2-Hellinger $\frac{\sqrt{\boldsymbol{x}}}{\|\sqrt{\boldsymbol{x}}\|_2}$, standarization $\mathtt{diag}(\boldsymbol{\sigma}_X)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_X)$ with the mean $\boldsymbol{\mu}_X$ and standard deviation $\boldsymbol{\sigma}_X$ estimated from $\boldsymbol{x}$, and the explicit feature map of χ^2 kernel proposed in [41]. Fig. 5 shows the performance results measured by ROC and equal error rate (EER).

[4] Actually, although the SIFT descriptors [29] are L_2-hysteresis normalized, forming spherical distributions, the GMM is directly applied on those distribution in the GMM Fisher kernel [36].

[5] Codes are available at https://staff.aist.go.jp/takumi.kobayashi/codes.html#DirFT

[6] In the case of SIFT, $\tau = 0.2$ as suggested in [29].



Method	EER (%)
ours	**94.95**
L_1 norm	78.26
L_2 norm	93.91
L_2-hys. norm	94.31
L_2-Hel. norm	94.63
standarize	91.31
kernel map [41]	94.89

Figure 5. Performance result (ROC) on Daimler Chrysler pedestrian dataset [33].

The proposed method outperforms the other normalization methods and is competitive with [41]. It should be noted that, while the method [41] augments the feature dimensionality D into $7D$ significantly increasing the computation time in classification, our method efficiently operates on respective feature components at a low computational cost without enlarging the dimensionality.

4.2. Bag-of-words histogram

We next apply the Dirichlet Fisher kernel to the bag-of-word (BoW) feature which is composed of *countable* word histogram. We test the method on PASCAL-VOC2007 dataset [8] in the following experimental setup.

A lot of SIFT local descriptors [29] are densely extracted at spatial grid points in 4 pixel step with three scales of $\{16, 24, 32\}$ pixels. To construct 16,384 visual words, we apply k-means clustering to a million of (transformed) SIFT descriptors randomly drawn from the training images. An image is partitioned into sub-regions in three levels of spatial pyramid as $1\times1, 2\times2$ and 3×1; the BoW histogram features are computed on the respective sub-regions and then concatenated into the image feature vector which is finally fed into the linear SVM classifier.

PASCAL-VOC2007 [8]. The dataset contains object images of 20 categories with large variation regarding appearances and poses as well as complex backgrounds. We follow the standard VOC evaluation protocol.

The performance results are shown in Table 1 with the comparison similarly to Sec.4.1, demonstrating that the proposed method is superior to the other normalization methods. It slightly outperforms the method of Pólya model [4], even though our method is so general as to accept not only this countable BoW histogram but also the other types of histogram features as in Sec.4.1. For comparison, we also applied the BoW method using LLC [44], and the proposed method significantly outperforms it.

4.3. Fisher kernel in bag-of-feature framework

We finally apply the proposed Dirichlet-derived GMM Fisher kernel (Sec.3.2) to various image classification tasks; object recognition [8, 11], fine-grained object classification [43], scene categorization [34, 45], event classification [24], aerial image classification [46] and material recognition [38]. The experimental setup for the BoF-based Fisher kernel is the same as in Sec.4.2 except that $K = 256$ GMM is obtained by EM method instead of visual words.

Table 1. Performance on VOC2007 using BoW.

Methods	mAP (%)	Methods	mAP (%)
ours	**60.91**	Pólya [4]	60.58
L_1 norm	51.93	standarize	58.16
L_2 norm	53.50	kernel map [41]	60.67
L_2-hys. norm [36]	59.58	BoW-LLC [44]	57.37
L_2-Hel. norm [1]	59.10		

Table 2. Performance on VOC2007 using BoF-based FK.

Methods	mAP (%)
ours	**63.83**
DMM-FK	57.51
$\log(x+\epsilon)$	63.48
L_1 norm	59.63
L_2 norm	60.03
L_2-hys. norm [36]	61.40
L_2-Hel. norm [1]	63.04
kernel map [41]	61.95

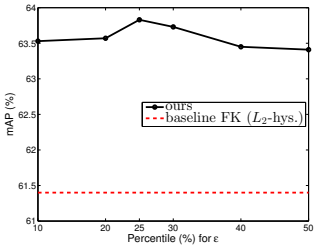


Figure 6. Performance of various ϵ on VOC2007.

We analyze the performance of the proposed method from various aspects by using the VOC2007 dataset [8]. The proposed method is compared to DMM Fisher kernel (18) as well as the other types of transformation of SIFT features. In this case, L_2-hysteresis normalization results in the baseline Fisher kernel [36] and L_2-Hellinger normalization produces RootSIFT [1]. The comparison results in Table 2 show that the proposed method outperforms the others. The method of DMM-FK is inferior since it corresponds to only the one part (*mean*) of the proposed FK as described in Sec.3.2. Its extension to GMM-FK using the transform of $\log(x+\epsilon)$ improves the performance by incorporating the FK features derived from the *variance*. Nonetheless, the proposed method is still better due to transforming SIFT features in accordance with von Mieses Fisher which actually reduces to Gaussian. It should be noted again that the method of [41] augments the SIFT dimensionality to $896 = 128 \times 7$, increasing the computational cost.

We then investigate the effects of ϵ which is determined based on the percentile of the distribution in Fig. 2a. Fig. 6 shows the performance of various ϵ with comparison to the baseline FK of L_2-hysteresis normalization [36]. The best performance is obtained at the 25% percentile, though all values of ϵ produces superior performance to the baseline. By using ϵ of the 25% percentile, the distribution of the features is transformed as shown in Fig. 2b; The smaller feature values are favorably rounded-off into $\log(\epsilon)$ (the negative evidence) while the higher features are diversely distributed to exploit the discriminative information. Throughout the experiments, we apply the proposed method with ϵ fixed to the value of 25% percentile, *i.e.*, $\epsilon = 0.001$ for SIFT, though fine tuning such as by cross validation per dataset

would further improve the performance.

The proposed method produces 63.83% which is superior to the VOC2007 winner (59.4%) and is favorably compared to the state-of-the-art result (63.5%) in [13] that resorts to time-consuming object detection.

In the following experiments on various datasets, the proposed Dirichlet-derived GMM FK is compared with the other types of normalization as well as the state-of-the-art results reported in the referenced papers.

MIT-Scene [34]. This dataset contains 15,620 images in 67 indoor scene categories with the large within-class and small between-class variability. We report the classification accuracy according to the experimental setting in [34].

UIUC-Sports [24]. For image-based event classification, Li and Fei-Fei [24] collected 1,792 images of eight sport categories, each of which contains 137$\sim$250 images with large variations of poses and sizes across diverse categories in the cluttered backgrounds. Following the experimental setup used in [24], we report the averaged classification accuracies.

Land-Use [46]. Yand *et al*. [46] collected the dataset of aerial orthoimagery downloaded from the United States Geological Survey (USGS) National Map in 21 land-use categories. Each category contains 100 images of 256×256 pixels in a resolution of one foot per pixel, including a variety of spatial patterns. We follow the experimental and evaluation protocol in [46].

Flickr Material [38]. There are 1,000 images of ten material categories and human-labeled binary masks associated with images that describe the location of the object. We extract local SIFT descriptors on the foreground indicated by the binary mask for material recognition. According to the evaluation protocol in [26], the averaged classification accuracies are reported.

CUB200-2011 [43]. This is a challenging dataset of 200 bird species, including 11,788 images in total, for fine-grained (subordinate) object recognition. We used full uncropped images and evaluated the methods by using the provided training/test split.

Caltech-256 [11]. This dataset is composed of 30,607 images in 256 object categories. There are large intra-class variances regarding such as object locations, sizes and poses in the images, which makes this dataset a challenging benchmark dataset for object recognition. According to the standard experimental setting, we randomly pick up 15, 30, 45, and 60 training images per category and (at most) 50 images for test. We report the averaged classification accuracy over three trials.

SUN-397 [45]. This dataset contains roughly 100K images of 397 scene categories covering as many of visual world scenes as possible. It is a challenging dataset since even "good" human workers in AMT achieve the classification performance of 68.5% on an average. Following the protocol of [45], we used 50 training and 50 test samples per category and measured the classification accuracies averaged over the given 10 training/test partitions.

The performance results on these datasets are shown in Table 3. The proposed method is favorably competitive with the other feature transforms, outperforming the other methods; compared to the baseline of L_2-hysteresis normalization, the performance is improved by $2\sim3\%$. The proposed method surely improves the performance of the Fisher kernel by effectively transforming the local descriptors at a low extra computational cost.

5. Conclusion

We have proposed methods to transform histogram-based features for improving classification performance. The (L_1-normalized) histogram features regarded as probability mass functions are modeled by using Dirichlet distribution from the probabilistic perspective. Based on the probabilistic model, we induce the Dirichlet Fisher kernel which transforms the individual histogram features, and in the BoF framework, the Dirichlet mixture model is extended to GMM via the feature transform, which leads to the Dirichlet-derived GMM Fisher kernel. The proposed methods do not require cumbersome parameter tuning; it is generally determined based on the statistics of the features, *e.g*., *natural SIFT statistics*. In the experiments on a variety of image classifications tasks, the proposed methods favorably improve the performance of the histogram-based features and the BoF-based Fisher kernel.

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[2] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *NIPS*, 2011.

[3] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *NIPS workshop on deep learning*, 2012.

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Image categorization using fisher kernels of non-iid image models ramazan. In *CVPR*, 2012.

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

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] C. Elkan. Deriving tf-idf as a fisher kernel. In *SPIRE*, 2005.

[8] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results.

[9] S. Gao, I.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely - laplacian sparse coding for image classification. In *CVPR*, 2010.

[10] P. Gehler and S. Nowozin. On feature combination for multiclass object classificatio. In *ICCV*, 2009.

[11] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[12] O. C. Hamsici and A. M. Martinez. Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8:1583–1623, 2007.

Table 3. Performance comparison on various datasets.

(a) MIT-Scene [34]

Method	Acc. (%)
Ours	**63.4**
L_2-hys. norm [36]	60.3
L_2-Hel. norm [1]	60.4
kernel map [41]	61.6
Bo *et al*. [2]	41.8
Zheng *et al*. [48]	47.2
Bo *et al*. [3]	50.5
Juneja *et al*. [18]	63.1

(b) UIUC-Sports [24]

Method	Acc. (%)
Ours	**92.6**$\pm$0.7
L_2-hys. norm [36]	91.3$\pm$1.3
L_2-Hel. norm [1]	92.2$\pm$1.0
kernel map [41]	92.3$\pm$1.4
Liu *et al*. [28]	84.6$\pm$1.5
Gao *et al*. [9]	85.3$\pm$0.5
Bo *et al*. [2]	85.7$\pm$1.3
Zheng *et al*. [48]	87.2

(c) Land-Use [46]

Method	Acc. (%)
Ours	**92.8**$\pm$0.9
L_2-hys. norm [36]	92.2$\pm$0.9
L_2-Hel. norm [1]	92.1$\pm$1.1
kernel map [41]	92.2$\pm$0.6
Yand *et al*. [46]	77.4
Jiang *et al*. [17]	77.8

(d) Flickr Material [38]

Method	Acc. (%)
Ours	**57.3**$\pm$0.9
L_2-hys. norm [36]	55.2$\pm$0.9
L_2-Hel. norm [1]	56.3$\pm$0.8
kernel map [41]	56.8$\pm$1.2
Liu *et al*. [26]	44.6
Li and Fritz [25]	48.1
Liu *et al*. [27]	48.2
Hu *et al*. [14]	54

(e) CUB200-2011 [43]

Method	Acc. (%)
Ours	**27.3**
L_2-hys. norm [36]	25.8
L_2-Hel. norm [1]	26.6
kernel map [41]	26.6
Wah *et al*. [43]	10.3
Wah *et al*. [42]	17.3

(f) Caltech-256 [11]

#sample	15	30	45	60
Ours	**41.8**$\pm$0.2	**49.8**$\pm$0.1	**54.4**$\pm$0.3	**57.4**$\pm$0.4
L_2-hys. norm [36]	39.8$\pm$0.3	48.0$\pm$0.3	52.4$\pm$0.3	55.4$\pm$0.2
L_2-Hel. norm [1]	41.2$\pm$0.3	49.5$\pm$0.2	53.9$\pm$0.4	56.8$\pm$0.3
kernel map [41]	40.3$\pm$0.1	48.5$\pm$0.2	52.9$\pm$0.3	55.9$\pm$0.4
Gehler *et al*. [10]	34.2	45.8	-	-
Kulkarni and Li [21]	39.4	45.8	49.3	51.4
Bo *et al*. [3]	40.5 $\pm$0.4	48.0 $\pm$0.2	51.9 $\pm$0.2	55.2 $\pm$0.3

(g) SUN-397 [45]

Method	Acc. (%)
Ours	**46.1**$\pm$0.1
L_2-hys. norm [36]	43.4$\pm$0.3
L_2-Hel. norm [1]	45.2$\pm$0.2
kernel map [41]	44.3$\pm$0.2
Shen *et al*. [39]	23.1
Xiao *et al*. [45]	38.0

[13] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[14] D. Hu, L. Bo, and X. Ren. Toward robust material recognition for everyday objects. In *BMVC*, 2011.

[15] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

[16] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval: The benefit of pca and whitening. In *ECCV*, 2012.

[17] Y. Jiang, J. Yuan, and G. Yu. Randomized spatial partition for scene recognition. In *ECCV*, 2012.

[18] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[19] T. Kobayashi and N. Otsu. Image feature extraction using gradient local auto-correlations. In *ECCV*, 2008.

[20] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-laplacian priors. In *NIPS*, 2009.

[21] N. Kulkarni and B. Li. Discriminative affine sparse codes for image classification. In *CVPR*, 2011.

[22] K.V.Mardia and P. Jupp. *Directional Statistics (2nd edition)*. John Wiley and Sons Ltd., 2000.

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

[24] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[25] W. Li and M. Fritz. Recognizing materials from virtual examples. In *ECCV*, 2012.

[26] C. Liu, L. Sharan, E. Adelson, and R. Rosenholtz. Exploring features in a bayesian framework for material recognition. In *CVPR*, 2010.

[27] L. Liu, P. Fieguth, G. Kuang, and H. Zha. Sorted random projections for robust texture classification. In *ICCV*, 2011.

[28] L. Liu, L. Wang, and X. Liu. In dense of soft-assignment coding. In *ICCV*, 2011.

[29] D. Lowe. Distinctive image features from scale invariant features. *International Journal of Computer Vision*, 60:91–110, 2004.

[30] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *ICML*, 2005.

[31] S. Maji and A. C. Berg. Max-margin additive classifiers for detection. In *ICCV*, 2009.

[32] T. Minka. Estimating a dirichlet distribution. Technical report, 2000.

[33] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

[34] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[35] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[36] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[37] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2001.

[38] L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784, 2009.

[39] L. Shen, S. Wang, G. Sun, S. Jiang, and Q. Huang. Multi-level discriminative dictionary learning towards hierarchical visual categorization. In *CVPR*, 2013.

[40] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[41] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

[42] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[44] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[45] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[46] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *ICCV*, 2011.

[47] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[48] Y. Zheng, Y.-G. Jiang, and X. Xue. Learning hybrid part filters for scene recognition. In *ECCV*, 2012.