Image categorization using Fisher kernels of non-iid image models

Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid

Image categorization using Fisher kernels of non-iid image models

Ramazan Gokberk Cinbis, Jakob Verbeek and Cordelia Schmid
LEAR, INRIA Grenoble, France Laboratoire Jean Kuntzmann
firstname.lastname@inria.fr

Abstract

The bag-of-words (BoW) model treats images as an unordered set of local regions and represents them by visual word histograms. Implicitly, regions are assumed to be identically and independently distributed (iid), which is a poor assumption from a modeling perspective. We introduce non-iid models by treating the parameters of BoW models as latent variables which are integrated out, rendering all local regions dependent. Using the Fisher kernel we encode an image by the gradient of the data log-likelihood w.r.t. hyper-parameters that control priors on the model parameters. Our representation naturally involves discounting transformations similar to taking square-roots, providing an explanation of why such transformations have proven successful. Using variational inference we extend the basic model to include Gaussian mixtures over local descriptors, and latent topic models to capture the co-occurrence structure of visual words, both improving performance. Our models yield state-of-the-art categorization performance using linear classifiers; without using non-linear transformations such as taking square-roots of features, or using (approximate) explicit embeddings of non-linear kernels.

1. Introduction

Bag of visual words (BoW) image representations [3, 19] are predominant in current state-of-the-art image categorization and retrieval systems. The BoW model represents an image as a histogram over visual word counts. The histograms are constructed by mapping local feature vectors in images to cluster indices, where the clustering is typically learned using k-means. Recently, Perronnin and Dance [15] have enhanced this basic representation using the notion of Fisher kernels [6]. In this case local descriptors are soft-assigned to components of a mixture of Gaussian (MoG) density, and the image is represented using the gradient of the log-likelihood of the local descriptors w.r.t. the MoG parameters. As we show below, both BoW as well as MoG Fisher vector representations are based on models that assume that local descriptors are independently and identically distributed (iid). However, the iid assumption is a very poor one from a modeling perspective, see Figure 1.

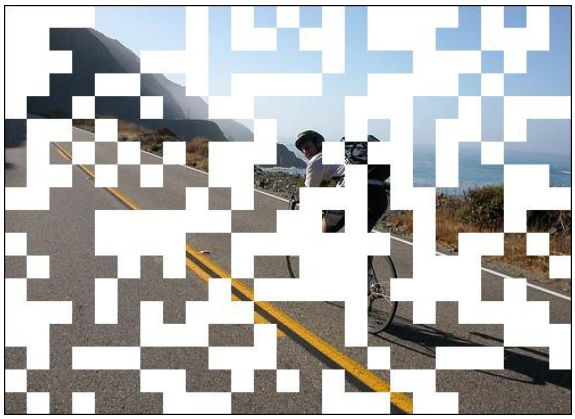

Figure 1. Local image patches are *not* iid: the visible patches are informative on the masked-out ones; one has the impression to have seen the complete image by looking at half of the patches.

In this paper we consider models that capture the dependencies among local image regions by means of non-iid but completely exchangeable models, *i.e*. like iid models our models still treat the image as an unordered set of regions. We treat the parameters of the BoW models as latent variables with prior distributions learned from data. By integrating out the latent variables, all image regions become mutually dependent. We generate image representations from these models by applying the Fisher kernel principle, in this case by taking the gradient of the log-likelihood of the data in an image w.r.t. the hyper-parameters that control the priors on the latent model parameters.

We first present the multivariate Pólya model which represents the set of visual word indices of an image as independent draws from an unobserved multinomial distribution drawn from a Dirichlet prior distribution. By integrating out the latent multinomial distribution, a model is obtained in which all visual word indices are mutually dependent. Interestingly, we find that our non-iid models yield gradients

that are qualitatively similar to popular ad-hoc transformations of BoW image representations, such as square-rooting histogram entries [8, 16, 17, 20]. Therefore, our first contribution is to show that such transformations appear naturally if we remove the poor iid assumption, *i.e*., to provide an explanation why such transformations are beneficial.

Our second model assumes that the region descriptors (*e.g*. SIFT) are iid samples from a latent MoG distribution, and we integrate out the mixing weights, means and variances of the MoG distribution. In this case the computation of the gradients is intractable. Our second contribution is to overcome this technical difficulty by computing a variational free-energy bound on the log-likelihood, and compute gradients w.r.t. the bound instead. This leads to a representation that performs on par with the Fisher vector representation of [17] based on iid MoG models, which includes square-root transformations and was found to be state-of-the-art in a recent independent evaluation study [2].

Our third contribution is to use the same variational framework to compute Fisher vector representations based on the latent Dirichlet allocation (LDA) model [1], in order to capture the co-occurrence statistics missing in BoW representations. We compare performance to Fisher vectors of PLSA [5], a topic model that does not treat the model parameters as latent variables. We find that topic models improve over BoW models, and that the LDA improves over PLSA even when square-rooting is applied.

In the following section we motivate our work in relation to the most relevant related work, and in Section 3 we present our non-iid latent variable models of local image descriptors. We present experimental results in Section 4, and summarize our conclusions in Section 5.

2. Motivation and related work

The use of non-linear feature transformations in BoW image representations is widely recognized to be beneficial for image categorization [8, 16, 17, 20, 22]. These transformations alleviate an obvious shortcoming of linear classifiers on BoW image representations: the fact that a fixed change Δ in a BoW histogram , from h to $h + \Delta$, leads to a score increment that is independent of the original histogram h: $f(h+\Delta) - f(h) = w^\top(h+\Delta) - w^\top h = w^\top \Delta$. Therefore, the score increment from images (a) though (d) in Figure 2 will be comparable, which is undesirable: the classifier score for *cow* should sharply increase from (a) to (b), and then remain stable among (b), (c), and (d).

Popular remedies to this problem include the use of chi-square kernels [22], or taking the square-root of histogram entries [16, 17], also referred to as the Hellinger kernel [20]. The effect of these is similar. Both transform the features such that the first few occurrences of visual words will have a more pronounced effect on the classifier score than if the count is increased by the same amount but starting at a larger value. This is desirable, since now the first patches providing evidence for an object category can significantly impact the score, *e.g*. making it easier to detect small object instances. The qualitative similarity is illustrated in Figure 3, where we compare the ℓ_2, chi-square, and Hellinger distances on the range $[0, 1]$.

Figure 2. The score of a linear 'cow' classifier will increase similarly from images (a) through (d) due to the increasing number of cow patches. This is undesirable: the score should sharply increase from (a) to (b), and remain stable among (b), (c), and (d).

The motivation for square-root and similar transformations tends to vary across papers. Sometimes it is based on empirical observations of improved performance [16, 20], by reducing sparsity in Fisher vectors [17], or in terms of variance stabilization transformations [8, 21]. To the best of our knowledge, we are the first to motivate them by showing that such discounting transformations appear naturally in models that do not make the unrealistic iid assumption.

Similar transformations are also used in image retrieval to counter burstiness effects [7], *i.e*., if rare visual words occur in an image, they tend to do so in bursts due to the locally repetitive nature of natural images. Burstiness also occurs in text, and the Dirichlet compound multinomial distribution, also known as multivariate Pólya distribution, has been used to model it [13]. This model places a Dirichlet prior on a latent per-document multinomial, and words in a document are sampled independently from it. In the next section, we use the multivariate Pólya distribution as our basic non-iid image model, and the Fisher kernel framework to compute image representations as the gradient w.r.t. the hyper-parameters of the Dirichlet prior. This differs from [13] which trained class-conditional Pólya models for use in a generative classification approach.

To apply the same idea in combination with the MoG Fisher kernel image representations of [15] is technically more involved. In this case, the latent model parameters (mixing weights, means, and variances) cannot be integrated out analytically, and the computation of the gradients is no longer tractable as in the MoG case of [15]. To over-

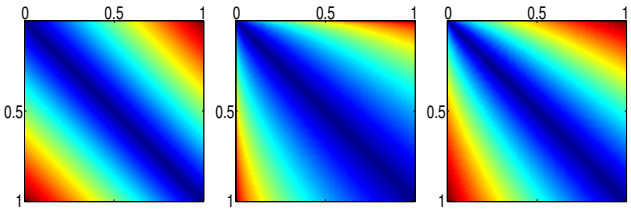

Figure 3. Comparison of (left to right) ℓ_2, Hellinger, and chi-square distances for x and y values ranging from 0 to 1. Both the Hellinger and chi-square distance discount the effect of small changes in large values unlike the ℓ_2 distance.

come this difficulty we rely on the variational free-energy bound [9], which is obtained by subtracting the Kullback-Leibler divergence between an approximate posterior on the latent variables and the true posterior. By imposing a certain independence structure on the approximate posterior, tractable approximate inference techniques can be devised. We then compute the gradient of the variational bound as a surrogate for the intractable exact log-likelihood. This differs from [14], which uses the variational free-energy to define an alternative encoding, replacing the Fisher kernel.

Our use of latent Dirichlet allocation (LDA) [1] differs from earlier work on using topic models such as LDA or PLSA [5] for object recognition [11, 18]. The latter use topic models to *compress* BoW image representations by using the inferred document-specific topic distribution. We, instead, use the Fisher kernel framework to *expand* the image representation by decomposing the original BoW histogram into several bags-of-words, one per topic, so that individual histogram entries not only encode how often a word appears, but also in combination with which other words it appears. Whereas compressed topic model representations were mostly found to at best maintain BoW performance, we find significant gains by using topic models.

3. Non-iid image representations

In this section we present our non-iid models. We start with a model for BoW quantization indices, and then extend it to a model over sets of local feature vectors, such as SIFT. Finally, we extend the model to capture co-occurrence statistics across visual words using LDA in Section 3.3.

3.1. Bag-of-words and the multivariate Pólya model

The standard BoW image representation can be interpreted as applying the Fisher kernel framework [6] to a simple iid multinomial model over visual word indices [10]. Let $w_{1:N} = \{w_1, \ldots, w_N\}$ denote the visual word indices corresponding to N patches sampled in an image, and let π be a learned multinomial over K visual words, parameterized in log-space, *i.e.* $p(w_i = k) = \pi_k$ with $\pi_k = \exp(\gamma_k)/\sum_{k'} \exp(\gamma_{k'})$. The gradient of the data log-likelihood is in this case given by $\frac{\partial \sum_i \ln p(w_i)}{\partial \gamma_k} = n_k - N\pi_k$, where n_k denotes the number of occurrences of visual word k among the set of indices $w_{1:N}$. This is a shifted version of the standard BoW histogram, where the mean of all image representations is centered at the origin. We stress that this multinomial interpretation of the BoW model assumes that the visual word indices across all images are iid.

Our first non-iid model assumes that for each image there is a different, a-priori unknown, multinomial generating the visual word indices in that image. In this model visual word indices within an image are mutually dependent, since knowing some of the w_i provides information on the underlying multinomial π, and thus also provides information on which subsequent indices could be sampled from it. The model is parameterized by a non-symmetric Dirichlet prior over the latent image-specific multinomial, $p(\pi) = \mathcal{D}(\pi|\alpha)$ with $\alpha = (\alpha_1, \ldots, \alpha_K)$, and the w_i are modeled as iid samples from π. The marginal distribution on the w_i is obtained by integrating out π:

$$p(w_{1:N}) = \int_\pi p(\pi) \prod_i p(w_i|\pi). \qquad (1)$$

This model is known as the multivariate Pólya, or Dirichlet compound multinomial [13], and the integral simplifies to

$$p(w_{1:N}) = \frac{\Gamma(\hat{\alpha})}{\Gamma(N+\hat{\alpha})} \prod_k \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}, \qquad (2)$$

where $\Gamma(\cdot)$ is the Gamma function, and $\hat{\alpha} = \sum_k \alpha_k$. See panels (a) and (b) in Figure 4 for a graphical representation of the BoW multinomial model, and the Pólya model.

Following the Fisher kernel framework, we represent an image by the gradient w.r.t. the hyper-parameters α_k of the log-likelihood of the visual word indices $w_{1:N}$:

$$\frac{\partial \ln p(w_{1:N})}{\partial \alpha_k} = \psi(\alpha_k + n_k) - \psi(\hat{\alpha} + N) - \psi(\alpha_k) + \psi(\hat{\alpha}), \qquad (3)$$

where $\psi(x) = \partial \ln \Gamma(x)/\partial x$ is the digamma function.

Only the first two terms in Eq. (3) depend on the counts n_k, and for fixed N the gradient is determined up to additive constants by $\psi(\alpha_k + n_k)$, *i.e.* it is given by a transformation of the visual word counts n_k. Figure 5 shows the transformation $\psi(\alpha + n)$ for various values of α, along with the square-root function for reference. We see that the same monotone-concave discounting effect is obtained as by taking the square-root of histogram entries. This transformation arises naturally in our latent variable model, and suggests that such transformations are successful *because* they correspond to a more realistic non-iid model, *c.f.* Figure 1.

Observe that in the limit of $\alpha \to \infty$ the transfer function becomes linear, since for large α the Dirichlet prior tends to a delta peak on the simplex and thus removes the uncertainty on the underlying multinomial, with an observed multinomial BoW model as its limit. In the limit of $\alpha \to 0$,

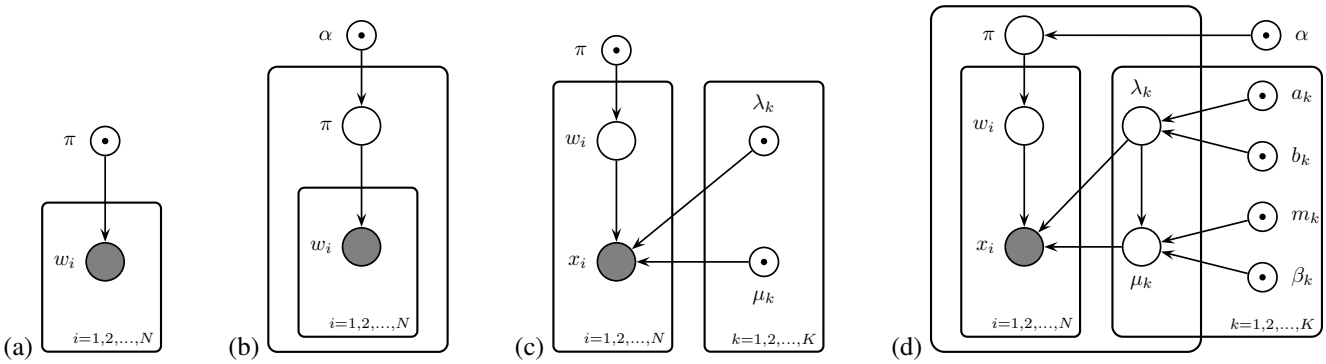

Figure 4. Graphical representation of the models in Section 3.1 and Section 3.2: (a) multinomial BoW model, (b) Pólya model, (c) MoG model, (d) latent MoG model. The outer plates in (b) and (d) refer to images. The index i runs over the N patches in an image, and index k over visual words. Nodes of observed variables are shaded, and those of (hyper-) parameters are marked with a central dot in the node.

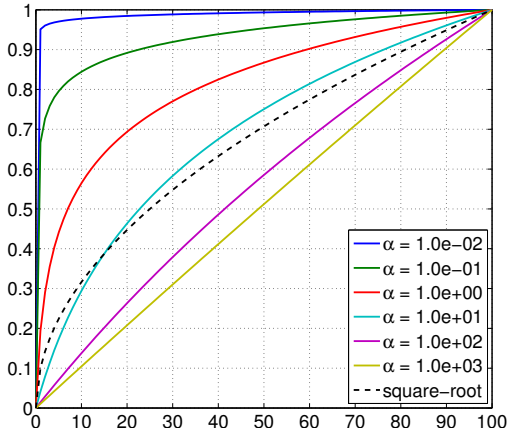

Figure 5. Digamma functions $\psi(\alpha + n)$ for various α, and $\sqrt{n}$ as a function of n; functions have been rescaled to the range $[0, 1]$.

corresponding to priors that concentrate their mass at sparse multinomials, the transfer function becomes a step function. This is intuitive, since in the limit of ultimately sparse distributions only one word will be observed, and its count no longer matters, we only need to know which word is observed to determine which α_k should be increased.

3.2. Modeling descriptors using latent MoG models

In this section we turn to the state-of-the-art image representation of [15] that applies the Fisher kernel framework to mixture of Gaussian (MoG) models over local descriptors.

A MoG density $p(x) = \sum_k \pi_k \mathcal{N}(x; \mu_k, \sigma_k)$ is defined by mixing weights $\pi = \{\pi_k\}$, means $\mu = \{\mu_k\}$ and variances $\sigma = \{\sigma_k\}$.[1] The K Gaussian components of the mixture correspond to the K visual words in a BoW model. In [15], local descriptors across images are assumed to be iid samples from a single MoG model underlying all images. They represent an image by the gradient of the log-likelihood of the descriptors $x_{1:N}$ sampled from it. For local descriptors of dimension D, *e.g.* $D = 128$ for SIFT, this yields an image representation of size $K(1+2D)$, since for each of the K visual words there is one derivative w.r.t. its mixing weight, and $2D$ derivatives for the means and variances in the D dimensions. This representation thus stores more information about the descriptors assigned to a visual word than just their count, as a result higher performance is obtained using a limited number of visual words.

In analogy to the previous section, we remove the iid assumption by defining a MoG model per image and treating its parameters as latent variables. We place conjugate priors on the image-specific parameters: a Dirichlet prior on the mixing weights, and a combined Normal-Gamma prior on the means μ_k and precisions $\lambda_k = \sigma_k^{-1}$:

$$p(\lambda_k) = \mathcal{G}(\lambda_k | a_k, b_k), \quad (4)$$

$$p(\mu_k | \lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \lambda_k)^{-1}). \quad (5)$$

The distribution on the descriptors $x_{1:N}$ in an image is obtained by integrating out the latent MoG parameters:

$$p(x_{1:N}) = \int_{\pi,\mu,\lambda} p(\pi) p(\mu, \lambda) \prod_{i=1}^{N} p(x_i | \pi, \mu, \lambda), \quad (6)$$

$$p(x_i | \pi, \mu, \lambda) = \sum_k p(w_i = k | \pi) p(x_i | w_i = k, \lambda, \mu), \quad (7)$$

where $p(w_i = k|\pi) = \pi_k$, and $p(x_i|w_i = k, \lambda, \mu) = \mathcal{N}(x_i | \mu_k, \lambda_k^{-1})$ is the Gaussian corresponding to the k-th visual word. See Figure 4 (c) and (d) for graphical representations of the MoG model and the latent MoG model.

Unfortunately, computing the log-likelihood in this model is intractable, and so is the computation of the gradient of the log-likelihood which we need for both hyper-parameter learning and to extract the Fisher vector representation. To overcome this problem we propose to approximate the log-likelihood by means of a variational lower

[1] We present here the uni-variate case for clarity, extension to the multivariate case with diagonal covariance matrices is straightforward.

bound [9], and compute gradients w.r.t. the bound $F \leq \ln p(x_{1:N})$ instead of the intractable log-likelihood, where

$$\begin{aligned} F &= \ln p(x_{1:N}) - D\big(q(\pi,\mu,\lambda,w_{1:N})||p(\pi,\mu,\lambda,w_{1:N}|x_{1:N})\big) \\ &= H(q) + \mathbb{E}_q[\ln p(x_{1:N},w_{1:N},\pi,\mu,\lambda)]\,, \end{aligned} \tag{8}$$

where $D(q||p)$ denotes the Kullback-Leibler divergence between distributions q and p. This is a valid bound for any choice of q, and the bound is tight when q matches the posterior on the hyper-parameters. If the bound is tight, it is easy to show that its gradient equals that of the data log-likelihood. By constraining q to factorize over the assignments w_i of local descriptors to visual words, and the latent MoG parameters π, λ, and μ,

$$q(\pi,\mu,\lambda,w_{1:N}) = q(\pi)\prod_k q(\mu_k|\lambda_k)q(\lambda_k)\prod_i q(w_i), \tag{9}$$

we obtain a bound for which we can tractably compute its value and gradient w.r.t. the hyper-parameters.

Given the hyper-parameters we can update the variational distributions $q(w_i)$ and $q(\pi), q(\mu_k|\lambda_k), q(\lambda_k)$ to improve the quality of the bound (although in general it will not be tight due to the decomposition imposed on q), the update equations are detailed in Appendix A.

The gradient of F w.r.t. the hyper-parameters depends only on the variational distributions on the MoG parameters of an image $q(\pi) = \mathcal{D}(\pi|\alpha^*)$, $q(\lambda_k) = \mathcal{G}(\lambda_k|a_k^*, b_k^*)$, and $q(\mu_k|\lambda_k) = \mathcal{N}(\mu_k|m_k^*, (\beta_k^*\lambda_k)^{-1})$, and not on the $q(w_i)$. For the precision hyper-parameters we find:

$$\frac{\partial F}{\partial a_k} = [\psi(a_k^*) - \ln b_k^*] - [\psi(a_k) - \ln b_k]\,, \tag{10}$$

$$\frac{\partial F}{\partial b_k} = \frac{a_k}{b_k} - \frac{a_k^*}{b_k^*}, \tag{11}$$

for the hyper-parameters of the means:

$$\frac{\partial F}{\partial \beta_k} = \frac{1}{2}\left(\beta_k^{-1} - \frac{a_k^*}{b_k^*}(m_k - m_k^*)^2 - 1/\beta_k^*\right), \tag{12}$$

$$\frac{\partial F}{\partial m_k} = \beta_k\frac{a_k^*}{b_k^*}(m_k^* - m_k), \tag{13}$$

and for the hyper-parameters of the mixing weights:

$$\frac{\partial F}{\partial \alpha_k} = [\psi(\alpha_k^*) - \psi(\hat{\alpha}^*)] - [\psi(\alpha_k) - \psi(\hat{\alpha})]\,. \tag{14}$$

By substituting the update equation (21) from Appendix A for the variational parameters α_k^* in the gradient Eq. (14), we exactly recover the gradient of the multivariate Pólya model, albeit using soft-counts $s_k^0 = \sum_i q(w_i = k)$ of visual word occurrences here. Thus, the bound leaves intact the qualitative behavior of the multivariate Pólya model. Similar discounting effects can be observed in the gradients of the hyper-parameters of the means and variances.

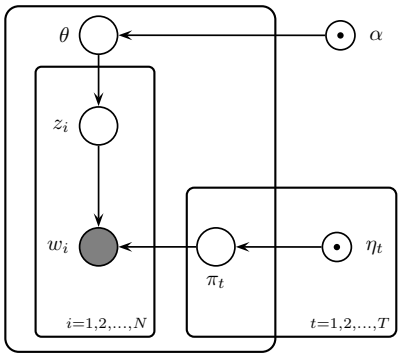

Figure 6. Graphical representation of LDA. The outer plate refers to images. The index i runs over patches, and index t over topics.

Note that in our latent MoG model we have two hyper-parameters (m_k, β_k) associated with each mean μ_k, and similar for the precisions. Therefore, our gradient representation of an image has length $K(1+4D)$, which is almost twice the size of the Fisher vector of the iid MoG model which are of size $K(1+2D)$. So our latent MoG model not only naturally generates the beneficial discounting effects, it also generates a higher dimensional gradient signal that might lead to better separability of object categories.

3.3. Capturing co-occurrence with topic models

In our third model, we extend the Pólya model to capture co-occurrence statistics of visual words using latent Dirichlet allocation (LDA) [1]. We model the visual words in an image as a mixture of T topics, encoded by a multinomial θ mixing the topics, where each topic itself is represented by a multinomial distribution π_t over the K visual words. We associate a variable z_i, drawn from θ, with each patch that indicates which topic was used to draw its visual word index w_i. We place Dirichlet priors on the topic mixing, $p(\theta) = \mathcal{D}(\theta|\alpha)$, and the topic distributions $p(\pi_t) = \mathcal{D}(\pi_t|\eta_t)$, and integrate these out to obtain the marginal distribution over visual word indices as:

$$p(w_{1:N}) = \int_\theta p(\theta)p(\pi)\prod_i p(w_i|\theta,\pi), \tag{15}$$

$$p(w_i = k|\theta,\pi) = \sum_t p(z_i = t|\theta)p(w_i = k|\pi_t). \tag{16}$$

See Figure 6 for a graphical representation of the model.

Both the log-likelihood and its gradient are intractable to compute for the LDA model. As before, however, we can resort to variational methods to compute a free-energy bound $F = \ln p(w_{1:N}) - D\big(q(\theta)\prod_t q(\pi_t)||p(\theta,\pi|w_{1:N})\big)$ on the data log-likelihood. The update equations of the variational distributions $q(\theta) = \mathcal{D}(\theta|\alpha^*)$ and $q(\pi_t) = \mathcal{D}(\pi_t|\eta_t^*)$ to maximize F are given by:

$$\alpha_t^* = \alpha_t + \sum_i q_{it}, \qquad \eta_{tk}^* = \eta_{tk} + \sum_{i:w_i=k} q_{it}, \tag{17}$$

where $q_{it} = q(z_i = t)$, which is itself updated according to $q_{it} \propto \exp[\psi(\alpha_t^*) - \psi(\hat{\alpha}^*) + \psi(\eta_{tk}^*) - \psi(\hat{\eta}_t^*)]$. The gradients w.r.t. the hyper-parameters are obtained from these as

$$\frac{\partial F}{\partial \alpha_t} = \psi(\alpha_t^*) - \psi(\hat{\alpha}^*) - [\psi(\alpha_t) - \psi(\hat{\alpha})], \quad (18)$$

$$\frac{\partial F}{\partial \eta_{tk}} = \psi(\eta_{tk}^*) - \psi(\hat{\eta}_t^*) - [\psi(\eta_{tk}) - \psi(\hat{\eta}_t)]. \quad (19)$$

The gradient w.r.t. α encodes a discounted version of the topic proportions as they are inferred in the image. The gradients w.r.t. the hyper-parameters η_t can be interpreted as decomposing the bag-of-word histogram over the T topics, and encoding the soft counts of words assigned to each topic. The entries $\frac{\partial F}{\partial \eta_{tk}}$ in this representation not only code how often a word was observed but also in combination with which other words, since the co-occurrence of words throughout the image will determine the inferred topic mixing and thus the word-to-topic posteriors.

In our experiments we compare LDA with the PLSA model [5]. This model treats the topics π_t, and the topic mixing θ as non-latent parameters which are estimated by maximum likelihood. To represent images using PLSA we apply the Fisher kernel framework and compute gradients of the log-likelihood w.r.t. θ and the π_t.

4. Experimental evaluation

We first describe our experimental setup, and then evaluate our latent BoW and MoG models in Section 4.2. We evaluate the topic model representations in Section 4.3.

4.1. Experimental setup

Results are reported on the PASCAL VOC'07 data set [4] with the interpolated mAP score specified by the VOC evaluation protocol. In order to obtain a state-of-the-art baseline, we use the experimental setup described in the recent evaluation [2]: we sample local SIFT descriptors from the same dense grid (3 pixel stride, across 4 scales), project the local descriptors to 80 dimensions with PCA, and train the MoG visual vocabularies from 1.5×10^6 descriptors. In BoW and Pólya models, we use the soft-assignment of patches to visual words to generate the word counts. We compare global image representations, and representations that capture spatial layout by concatenating the signatures computed over various spatial cells as in the spatial pyramid matching (SPM) method [12]. Again, we follow [2] and combine a 1×1, a 2×2, and a 3×1 grid. Throughout, we use linear SVM classifiers, and we cross-validate the regularization parameter.

In order to speed-up the training process of our non-iid latent variable models, we fix the patch-to-word soft-assignments as obtained from the MoG dictionary, and run the variational EM algorithm only to learn the hyper-parameters and to update the latent MoG parameter posteriors (as detailed in Section 3.2 and Appendix A). The LDA models are trained in a similar way: we first train a PLSA model, and then fit Dirichlet priors on the topic-word and document-topic distributions as inferred by PLSA.

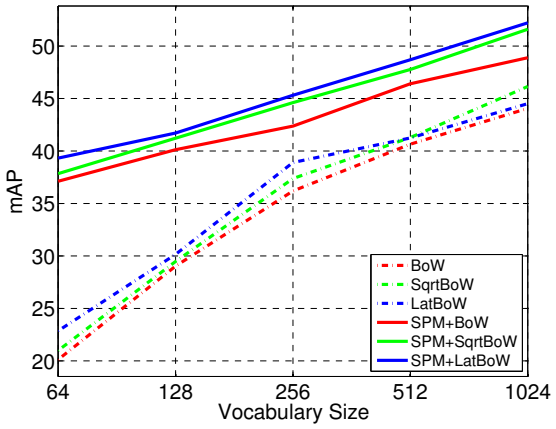

Figure 7. Comparison of BoW representations: plain BoW (red), square-root BoW (green) and Pólya latent BoW model (blue). With SPM (solid) and without (dashed).

SPM	Method	64	128	256	512	1024
No	BoW	20.1	29.0	36.2	40.7	44.1
No	SqrtBoW	21.0	29.5	37.4	**41.3**	**46.1**
No	LatBoW	**22.9**	**30.1**	**38.9**	41.2	44.5
Yes	BoW	37.1	40.1	42.4	46.4	48.9
Yes	SqrtBoW	37.8	41.2	44.6	47.8	51.6
Yes	LatBoW	**39.3**	**41.7**	**45.3**	**48.7**	**52.2**

Table 1. Comparison of BoW representations: plain BoW, square-root BoW and Pólya. The data is the same as in Figure 7.

Before training the classifiers we apply two normalizations to the image representations. First, we whiten the representations so that each dimension is zero-mean and has unit-variance across images, this corresponds to an approximate normalization with the inverse Fisher information matrix [10]. Second, following [17], we also ℓ_2 normalize the image representations.

We compare representations without square-rooting, those with square-rooting applied, and the corresponding latent variable models. As in [17], square-rooting is applied *after* whitening, and *before* ℓ_2 normalization.

4.2. Evaluating latent BoW and MoG models

In Figure 7 and Table 1 we compare the results obtained using standard BoW histograms, square-rooted histograms, and the Pólya model. Overall, we see that the spatial information of SPM is useful, and that larger vocabularies increase performance. We observe that both square-rooting and the Pólya model both consistently improve the BoW representation, across all dictionary sizes, and with or without SPM. Furthermore, the Pólya model generally

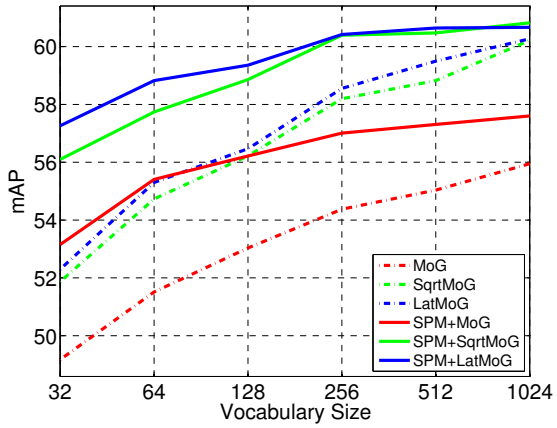

Figure 8. Comparison of MoG representations: plain MoG (red), square-root MoG (green) and latent MoG (blue). With SPM (solid) and without (dashed).

SPM	Method	32	64	128	256	512	1024
No	MoG	49.2	51.5	53.0	54.4	55.0	55.9
No	SqrtMoG	51.9	54.7	56.2	58.2	58.8	60.2
No	LatMoG	**52.3**	**55.3**	**56.5**	**58.6**	**59.5**	**60.3**
Yes	MoG	53.2	55.4	56.2	57.0	57.3	57.6
Yes	SqrtMoG	56.1	57.7	58.9	**60.4**	60.5	**60.8**
Yes	LatMoG	**57.3**	**58.8**	**59.4**	**60.4**	**60.6**	60.7

Table 2. Comparison of MoG representations: plain MoG, square-root MoG and latent MoG. The data is the same as in Figure 8.

leads to larger improvements than square-rooting. These results confirm the observation of Section 3.1 that the non-iid Pólya model generates similar transformations on BoW histograms as square-rooting does, providing an understanding of *why* square-rooting is beneficial.

In Figure 8 and Table 2, we compare image representations based on Fisher vectors computed over MoG models, their square-rooted version, and the latent MoG model of Section 3.2. We can observe that the MoG representations lead to better performance than the BoW ones while using smaller vocabularies. Furthermore, the discounting effect of our latent model and square rooting has a much more pronounced effect here than it has for BoW models, improving mAP scores by around 4 points. Also here our latent models lead to improvements that are comparable and often better than those obtained by square-rooting. So again, *the benefits of square-rooting can be explained by using non-iid latent variable models that generate similar representations.*

4.3. Evaluating topic model representations

To evaluate the performance of topic model representations, we compare Fisher vectors computed on the PLSA model, its square-rooted version, and when using the corresponding latent variable model (LDA) of Section 3.3 instead. We compare to the corresponding BoW representations, and include SPM in all experiments. In Figure 9, we consider topic models using $T=2$ topics for various dictionary sizes, and in Figure 10 we use dictionaries of $K=1024$ visual words, and consider performance as a function of the number of topics. We observe that (i) topic models consistently improve performance over BoW models, and (ii) the plain PLSA representations are consistently outperformed by the square-rooted version and the LDA model. The LDA model requires less topics than (square-rooted) PLSA to obtain similar performance levels. This confirms our findings with the BoW and MoG model of the previous section.

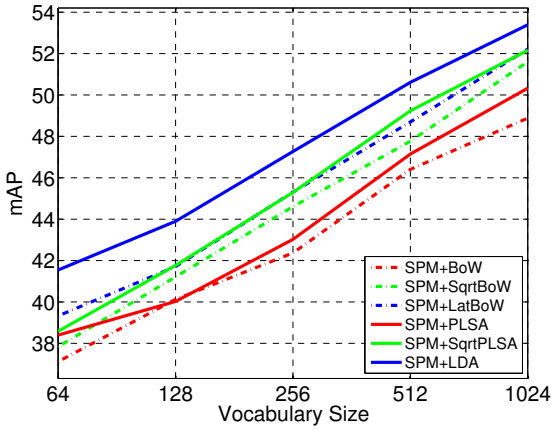

Figure 9. Topic models ($T = 2$, solid) compared with BoW models (dashed): BoW/PLSA (red), square-root BoW/PLSA (green), and Pólya/LDA (blue). SPM included in all experiments.

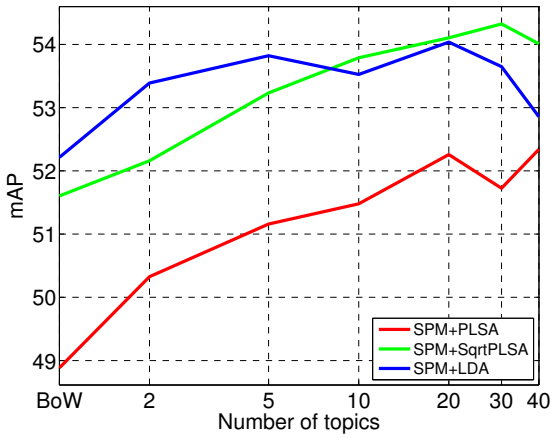

Figure 10. Performance when varying the number of topics: PLSA (red), square-root PLSA (green), and LDA (blue). BoW/Pólya model performance included as the left-most data point on each curve. All experiments use SPM, and $K = 1024$ visual words.

5. Conclusions

In this paper we have introduced latent variable models for local image descriptors, which avoid the common but unrealistic iid assumption. The Fisher vectors of our non-

iid models are functions computed from the same sufficient statistics as those used to compute Fisher vectors of the corresponding iid models. In fact, these functions are similar to transformations that have been used in earlier work in an ad-hoc manner, such as the square-root. Our models provide an explanation of the success of such transformations, since we derive them here by removing the unrealistic iid assumption from the popular BoW and MoG models. Second, we have shown that a variational free-energy bound on the log-likelihood can be successfully used to compute approximate Fisher vectors for intractable latent variable models, such as the latent MoG model, and the LDA topic model. Third, we have shown that the Fisher vectors of our non-iid models lead to image categorization performance that is comparable or superior to that obtained with current state-of-the-art representations based on iid models.

We believe that Fisher kernels combined with more advanced generative models, *e.g.* by modeling spatial temporal structure, is a promising direction of future research to derive more powerful image and video representations.

Acknowledgements. This work was partially funded by the QUAERO project supported by OSEO, French State agency for innovation and the EU integrated project AXES.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

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

[4] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop.

[5] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

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

[7] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

[8] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. PAMI*, 2012. to appear.

[9] M. Jordan, Z. Ghahramani, T. Jaakola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[10] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with Fisher vectors for image categorization. In *ICCV*, 2011.

[11] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image and Vision Computing*, 27(5):523–534, 2009.

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

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

[14] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. Free energy score space. In *NIPS*, 2009.

[15] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[16] F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.

[17] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

[18] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van-Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, pages 883–890, 2005.

[19] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

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

[21] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[22] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2):213–238, 2007.

A. Variational EM update equations

We use $q_{ik} = q(w_i = k)$ to define the sufficient statistics required for the variational update of the MoG parameters:

$$s_k^0 = \sum_i q_{ik}, \quad s_k^1 = \sum_i q_{ik} x_i, \quad s_k^2 = \sum_i q_{ik} x_i^2. \tag{20}$$

The parameters of the optimal variational distributions on the MoG parameters for a given image are then found as:

$$\alpha_k^* = \alpha_k + s_k^0, \tag{21}$$

$$\beta_k^* = \beta_k + s_k^0, \tag{22}$$

$$m_k^* = (s_k^1 + \beta_k m_k)/\beta_k^*, \tag{23}$$

$$a_k^* = a_k + s_k^0/2, \tag{24}$$

$$b_k^* = b_k + \frac{1}{2}(\beta_k m_k^2 + s_k^2) - \frac{1}{2}\beta_k^*(m_k^*)^2. \tag{25}$$

The $q(z_i)$ distributions are in turn updated from the variational distributions on the MoG parameters by setting:

$$\ln q_{ik} = \mathbb{E}_{q(\pi)q(\lambda_k,\mu_k)}\left[\ln \pi_k + \ln \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})\right] \tag{26}$$

$$= \psi(\alpha_k^*) - \psi(\hat{\alpha}^*) + \frac{1}{2}\Big[\psi(a_k^*) - \ln b_k^*\Big] \tag{27}$$

$$-\frac{1}{2}\Big[\frac{a_k^*}{b_k^*}(x_i - m_k^*)^2 + (\beta_k^*)^{-1}\Big]. \tag{28}$$