Adapted Gaussian Models for Image Classification

Mandar Dixit Nikhil Rasiwasia Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
mdixit@ucsd.edu, nikux@ucsd.edu, nuno@ece.ucsd.edu

Abstract

A general formulation of "Bayesian Adaptation" for generative and discriminative classification in the topic model framework is proposed. A generic topic-independent Gaussian mixture model, known as the background GMM, is learned using all available training data and adapted to the individual topics. In the generative framework, a Gaussian variant of the spatial pyramid model is used with a Bayes classifier. For the discriminative case, a novel predictive histogram representation for an image is presented. This builds upon the adapted topic model structure, using the individual class dictionaries and Bayesian weighting. The resulting histogram representation is evaluated for classification using a Support Vector Machine (SVM). A comparative evaluation of the proposed image models with the standard ones in the image classification literature is provided on three benchmark datasets.

1. Introduction

Over the last decade, a substantial amount of computer vision research has been devoted to problems such as object recognition or image classification. One representation that has consistently achieved good performance is the so called *bag of visual features* (BoF), where images are represented as orderless collections of spatially localized features. This representation has been used to design two major types of image classifiers, which we denote by *generative* and *discriminative.* Both of these approaches rely on a *mid-level* generative representation, which summarizes the distribution of BoF extracted from the images. Under the generative classification strategy, this mid-level representation is a topic model, and image classification is based on the posterior probabilities of images under the models learned from the different classes [24, 10, 2, 22]. Under the discriminative strategy, the mid-level representation is the so called *bag of visual words* (BoW) model. This model represents an image by the histogram of occurrences of representative points, which are derived from the BoF representation [6, 10]. A support vector machine (SVM) is then used for image classification [6]. More recently, BoW has been shown to benefit from a weak encoding of spatial information, through the introduction of a *spatial pyramid* structure [17]. The globally orderless image representation is replaced by a collection of locally orderless bags of visual words, aggregated at different levels of spatial resolution.

Extensive experimental evaluation has shown that the success of either the generative or discriminant strategies is, in significant part, determined by the mid-level representation. Under the discriminative strategy, the lexicon of visual words was initially learned with the k-means algorithm. More recently, alternative techniques have been proposed to either learn the visual lexicon, or obtain a discriminative representation of the visual data from it. Some variations include the use of alternative clustering algorithms, e.g. kernel-based clustering with histogram intersection kernels (HIK) [26], methods replacing the hard quantization of k-means with soft weighting mechanisms [14], the use of sparse codes [27, 4] instead of histograms, or the replacement of the spatial averaging implemented by histograms with non-linear pooling operators [27, 4].

On the generative front, most emphasis has been given to the design of generative models learnable from weakly labeled data. Several researchers have investigated extensions of unsupervised modeling techniques from text classification such as latent Dirichlet allocation [1] and probabilistic latent semantic analysis [15], into supervised models for visual classification [10, 11]. Alternatively, topic distributions can be learned with supervision, modeling each topic with a Gaussian mixture model (GMM) [5], a Dirichelet mixture [22], or a kernel density estimate [2]. These approaches have been shown successful in problems such as image annotation, where class labels tend to be very noisy.

It could be argued that many of the enhancements proposed in the literature are attempts to fix the problems originated by a poorly learned mid-level representation. The problem is not so much that the techniques used are inherently poor, but that the available training data is not sufficient to guarantee probability estimates that generalize well.

Typically, the space is large (e.g. 128 dimensional for the ubiquitous SIFT descriptor [20]) and the number of training images per class, or topic, is small (usually in the hundreds). The class specific models, therefore, lack *generalization*. The BoW model, which is learnt over the entire corpus, achieves the needed *generalization* but sacrifices class *discrimination* ability in the process [2].

The problem of simultaneously achieving *class specificity* and *generalization*, has been extensively studied in speech processing. One of the most successful approaches is *model adaptation*. The idea is to learn a *global* or *background* model, with good *generalization*, from the entire corpus, and then *adapt it* to each image class, to guarantee good *discrimination*. The dominant approach, referred to as *Bayesian model adaptation* [25, 23], is to use the parameters of the global model to design a prior distribution on the model parameter space. Adaptation is then implemented with Bayesian inference techniques, which combine this prior with the data available *per* class, to obtain a *class-adapted* model, which can either be the model of maximum a posteriori probability (MAP) parameters, or the full Bayesian predictive distribution [8]. The process is illustrated in Figure 1.

While Bayesian model adaptation has received some attention in the vision literature, some of the previous efforts have discussed generative classification strategy. [28] addressed *image adaptation*, i.e. the adaptation of a background model to each image, so as to obtain a *Gaussian Super-Vector* representation, which comprises of the adapted model parameters and a Gaussian posterior map for the image. In [9], *class adaptation* was used to learn category models with very few examples for object detection.

In this work, we present a more general formulation of Bayesian adaptation, which targets *class adaptation* and is applicable to *both* the generative and discriminative strategies for the problem of image classification. In both cases, a global GMM is first adapted to *each class*, using a Bayesian extension of the EM algorithm [23]. For generative classification, this is combined with the generative equivalent of spatial pyramid coding, allowing adaptation to both class and spatial pyramid cell. For discrimintive classification, it is used to produce a novel representation, an histogram-based predictive distribution, denoted the *predictive histogram*. This consists of learning histograms from class-adapted GMMs, and combining them with a Bayesian weighting mechanism. Classification is then performed with an SVM, as is usually the case. Extensive experimental results are provided to support the the efficacy of the proposed Bayesian adaptation approach. It is shown that adapted GMMs outperform both standard GMMs and standard BoW as a mid-level representation. Classification performance is then investigated for the proposed generative and discriminative classifiers. These are shown to outperform all previous methods of comparable classification complexity.

2. Mid-level Representations

We start with a brief review of the generative topic models and the bag-of-words representation currently popular in vision. Both are based on modeling images as bags of features.

2.1. Bag-of-features

In recent computer vision, an image $\mathcal{I}$ is frequently represented as a bag of low-level visual features $\mathcal{I} = \{x_1, \ldots, x_M\}$. These are modeled as independent and identically distributed observations from a random variable X, defined on some feature space χ. A corpus is a collection of images $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$ annotated with respect to a vocabulary $\mathcal{L} = \{l_1, \ldots, l_N\}$ of N topics. The d^{th} image, I_d is annotated with a label vector $c_d \in \{0,1\}^N$, whose i^{th} entry, c_{di}, is an indicator variable for the i^{th} topic. In this work we consider the scenario where $c_{di} = 1$ for only one value of i, i.e. images are annotated with a single topic, which can also be seen as classes. Topics are assumed to be independently drawn from a random variable $T \in \{1, \ldots, N\}$.

2.2. Generative topic models

Under the generative classification framework [5, 10], images are assigned to topics with the Bayes decision rule

$$\begin{aligned} t^* &= \arg\max_t P_{T|X}(t|\mathcal{I}) & (1) \\ &= \arg\max_t \prod_{x_j \in \mathcal{I}} P_{X|T}(x_j|t) & (2) \end{aligned}$$

where we have assumed a uniform prior over topics, $P_T(t) = 1/N, \forall t$. This is a common assumption in the literature. The topic-conditional distributions $P_{X|T}(x|t)$ are learned from the set $\mathcal{D}_t$ of features extracted from all images labeled with the topic t. Various approaches have been proposed for this purpose [10, 5, 12, 6, 2].

In this work, we adopt the popular representation of these distributions as Gaussian mixture models (GMMs). Under these models, X can be sampled from a number of Gaussian clusters, according to the state of a hidden variable K

$$\begin{aligned} P_{X|\Theta}(x|\theta) &= \sum_{k=1}^{K} P_K(k) P_{X|K}(x|k) & (3) \\ &= \sum_{k=1}^{K} w_k \mathcal{G}(x; \mu_k, \Sigma_k) & (4) \end{aligned}$$

where $\sum_k w_k = 1$,

$$\mathcal{G}(x, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

is a Gaussian distribution of mean μ and covariance Σ, and $\Theta = \{(w_1, \mu_1, \Sigma_1), \ldots, (w_K, \mu_K, \Sigma_K)\}$ is the set of GMM parameters. These are learned so as to maximize the likelihood of the features extracted from images of the topic

$$\theta_t^* = \arg\max_\theta P_{X|\Theta}(\mathcal{D}_t|\theta), \tag{5}$$

using the expectation-maximization (EM) algorithm [7]. This algorithm iterates between the following steps

E Step:

$$\begin{aligned} n_k &= \sum_{i=1}^n P_{K|X}(k|x_i); && (6)\\ E_{K|X}(x) &= \frac{1}{n_k}\sum_{i=1}^n P_{K|X}(k|x_i)x_i && (7)\\ E_{K|X}(x^2) &= \frac{1}{n_k}\sum_{i=1}^n P_{K|X}(k|x_i)x_i^2 && (8)\end{aligned}$$

M Step:

$$\begin{aligned} \hat{w_k^t} &= n_k/n && (9)\\ \hat{\mu_k^t} &= E_{K|X}(x) && (10)\\ (\hat{\sigma^2})_k^t &= E_{K|X}(x^2) - (\hat{\mu_k^t})^2 && (11)\end{aligned}$$

The E step computes the statistics $n_k, E_{K|X}(x)$ and $E_{K|X}(x^2)$, based on the data and the model parameters, while the M step updates the parameters based on the statistics. It can be shown that these iterations converge to the ML estimate θ^*. The topic-conditional distributions are set to the GMMs of parameters estimated from the associated training data, i.e. $P_{X|T}(x|t) = P_{X|\Theta}(x|\theta_t^*)$.

2.3. Discriminant classifiers

An alternative classification approach is based on the popular "bag-of-words" image representation. A codebook $C = \{c_1, c_2, \ldots, c_K\}$ is first learned from the entire corpus $\mathcal{D}$. This is frequently done with clustering techniques, such as k-means. The resulting codebook is a simplified GMM, whose mixture components have identical weights, $w_k = 1/K$, $\forall k$ and covariances $\Sigma_k = \Sigma$, $\forall k$. It is also possible to use a full GMM, similar to that of the previous section, but now learned from the entire corpus, i.e. a GMM of parameters

$$\theta^* = \arg\max_\theta P_{X|\Theta}(\mathcal{D}|\theta). \tag{12}$$

The visual features extracted from each image are then mapped to the closest entries in the codebook. This is equivalent to mapping each visual descriptor to the mixture component of maximum posterior probability. Given a feature $x_i \in \mathcal{I}$, the closest codeword k_i^* is

$$\begin{aligned} k_i^* &= \arg\max_k P_{K|X}(k|x_i) && (13)\\ &= \arg\max_k P_K(k)P_{X|K}(x_i|k) && (14)\\ &= \arg\max_k w_k \mathcal{G}(x_i, \mu_k, \Sigma_k). && (15)\end{aligned}$$

This mapping is usually referred to as *hard quantization*. It has been shown that *soft quantization* schemes are likely to produce more discriminative image representations [14]. For GMMs, soft quantization corresponds to assigning features partially to each of the GMM clusters, according to their posterior probabilities

$$\mathbf{v}_i = \left[P_{K|X}(1|x_i), P_{K|X}(2|x_i), \ldots, P_{K|X}(K|x_i)\right] \tag{16}$$

where $\mathbf{v}_i$ is the vector of soft-counts associated with feature x_i. Hard quantization (15) is the limit case where $\mathbf{v}_i$ is a vector of all zeros and one '1' at position k_i^*. The counts/soft-weights of each codeword, contributed by all features in the image, are then pooled into a histogram

$$H(\mathcal{I}) = \mathcal{F}(\mathbf{v}_1, \ldots, \mathbf{v}_n) \tag{17}$$

which is the final image representation. The standard *average pooling* operator aggregates word counts into bins of $H(\mathcal{I})$ and normalizes

$$\mathcal{F}_{av}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \frac{1}{n}\sum_i \mathbf{v}_i$$

$H(\mathcal{I})$, thus, represents a histogram for the image $\mathcal{I}$. Apart from the average pooling operator, other operators have also been proposed. For example, [27] has shown that a max operator can sometimes produce better results.

$$\mathcal{F}_{max}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \left[\max_i(\mathbf{v}_{i1}), \ldots, \max_i(\mathbf{v}_{iK})\right]$$

The histogram $H(\mathcal{I})$, is fed to a support vector machine (SVM) with a suitable kernel, for the final image classification.

3. Improved representations based on Bayesian model adaptation

In this section, we discuss, in brief the theory of Bayesian model adaptation. Extensions to both the generative and discriminative approaches of the previous section are then proposed, based on model adaptation.

3.1. Bayesian model adaptation

Model adaptation, is a popular modeling approach in the speech and speaker recognition literatures [25], [23]. In the

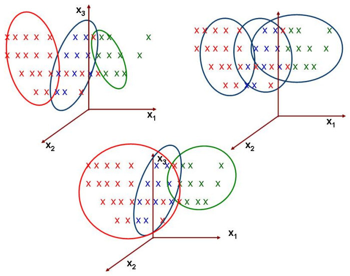

Figure 1. Schematic demonstrating the process of Bayesian model adaptation. The background model gradually adapts to the incoming topic specific data giving rise to the three topic models(top left). A generic model(top right) is not discriminant enough while the non adapted topic models(bottom) are too sensitive to outlier data.

vision context, it can be seen as means to increase the generalization ability of the topic models $P_{X|T}(x|t)$ of the generative approach. Rather than learning a model per topic, a *background model* is first learned from the entire corpus $\mathcal{D}$. This model is identical to that used by the discriminative approaches, i.e. that of (12), and is learned as discussed in Section 2.3. It is, however, not used as the basis for a histogram-based image representation. Instead, it provides *prior* knowledge about the parameters, which is combined with the data available per topic to learn individual topic models. This is a form of *regularization*, which guarantees improved generalization.

Given the set of parameters $\Theta = \{(w_1^b, \mu_1^b, \Sigma_1^b), \ldots, (w_K^b, \mu_K^b, \Sigma_K^b)\}$ of the background GMM, the prior $P_\Theta(\theta)$ for the training of topic model t is composed of a Dirichlet distribution for the weights w_i^t, a Gaussian for the means μ_i^t, and a Normal Wishart distribution for the covariances Σ_i^t [13],

$$\begin{aligned}(w_1^t, \ldots, w_K^t) &\sim Dir(Pw_1^b, \ldots, Pw_K^b) && (18)\\ \mu_k^t &\sim \mathcal{N}(\mu_k^b, \Sigma_k^b/r), \; k = 1, \ldots, K && (19)\\ \Sigma_k^t &\sim \mathcal{W}_p(R_k, h), \quad k = 1, \ldots, K. && (20)\end{aligned}$$

Here, P is a pseudocount for the Dirichlet distribution, r a smoothing parameter for the means, R_k a $d \times d$ positive definite symmetric matrix, and h a number of degrees of freedom for parameter Σ_K^t. The individual model for class t is obtained by maximizing the a-posteriori probability of the GMM parameters given the training data $\mathcal{D}_t$ available for the class and this prior,

$$\theta_t^* = \arg\max_\theta P_{X|\theta}(\mathcal{D}_t|\theta)P_\Theta(\theta). \tag{21}$$

As in section 2.2, this optimization is solved by an EM algorithm, which iterates between the following steps (see [23], [13] for details),

E Step:

$$\begin{aligned} n_k &= \sum_{i=1}^{n} P_{K|X}(k|x_i); && (22)\\ E_{K|X}(x) &= \frac{1}{n_k}\sum_{i=1}^{n} P_{K|X}(k|x_i)x_i && (23)\\ E_{K|X}(x^2) &= \frac{1}{n_k}\sum_{i=1}^{n} P_{K|X}(k|x_i)x_i^2 && (24)\end{aligned}$$

M Step:

$$\begin{aligned}\hat{w_k^t} &= \alpha_w^k(n_k/n) + (1-\alpha_w^k)w_k^b && (25)\\ \hat{\mu_k^t} &= \alpha_m^k E_{K|X}(x) + (1-\alpha_m^k)\mu_k^b && (26)\end{aligned}$$

$$(\hat{\sigma^2})_k^t = \alpha_v^k E_{K|X}(x^2) + (1-\alpha_v^k)((\sigma_k^b)^2 + (\mu_k^b)^2) - \hat{\mu_k}^2 \tag{27}$$

where

$$\begin{aligned}\alpha_w^k &= n/(n+P) && (28)\\ \alpha_m^k &= n_k/(n_k+r) && (29)\\ \alpha_v^k &= n_k/(n_k+r) && (30)\end{aligned}$$

We favor the update equations from [23] to the actual update equation [13]. The ratios α_p^k are relevance weights for the update of parameter p. They affect the extent of the adaptation, by controlling the influence of the background model on the parameter updates. If the Dirichlet pseudo-count P is large relative to the amount of training data n, the prior weights dominate (25). Similarly, if α_m^k and α_v^k are small, the statistics $E_{K|X}(x), E_{K|X}(x^2)$ from the topic data have small influence on the parameter estimates. The schematic in Fig. 1 demonstrates the process of Bayesian adaptation of the background by the topic data.

3.2. Adapted models for generative classification

The extension of Bayesian classification to Bayes adapted models is quite straightforward. The adapted topic models are simply used in the Bayes decision rule of (1).

3.3. Adapted models for discriminant classification

The availability of topic-adapted GMMs also enables a Bayesian treatment of the histogram-based representation. This builds on the fact that an image $\mathcal{I}$ has a different histogram under the GMM adapted to each topic $T = t$, and can thus be represented by N histograms $\{H^1(\mathcal{I}), \ldots, H^N(\mathcal{I})\}$. Each of these is a non-parametric representation of the topic-conditional distribution $P_{X|T}(x|t)$. Under Bayesian inference, these models are combined into the *predictive distribution*

$$P_{X'|X}(x|\mathcal{I}) = \sum_{t=1}^{T} P_{X'|T}(x|t)P_{T|X}(t|\mathcal{I}) \tag{31}$$

where X' is the random variable from which future observations are made, and X that from which the training observations in $\mathcal{I}$ have been drawn. Under the histogram representation this is written as

$$H(\mathcal{I}) = \sum_{t=1}^{T} P_{T|X}(t|\mathcal{I}) H^t(\mathcal{I}), \qquad (32)$$

where the topic posterior probabilities $P_{T|X}(t|\mathcal{I})$ are obtained from the adapted GMMs, i.e. those used in (1) for generative classification. As is usual in Bayesian inference, this predictive distribution is an average of all models, weighted by how much the observed data (the image $\mathcal{I}$) supports each of them [8]. The proposed extension of the discriminant classifier combined the histogram of (32), which we denote by *predictive histogram*, with an SVM classifier.

3.4. Spatial pyramid extensions

Various authors have shown that there is an advantage to augmenting the bag-of-features representation with a coarse coding of spatial feature location. In this work we rely on the popular spatial pyramid representation of [17]. A generic background GMM is adapted for each topic *per* spatial bin of the pyramidal structure. This produces a collection of models $P_{X|L,T}(x|l,t)$, where L is an index over the cells of the spatial pyramid. Given an image $\mathcal{I}$, a bag of features $\mathcal{I}_l, l = 1, \ldots, L$ is extracted from each cell. Under the generative classification approach, these cell log-likelihoods are averaged to get the image log likelihood.

$$\log P_{X|T}(\mathcal{I}|t) = \frac{1}{L} \sum_{l} \log P_{X|T,L}(\mathcal{I}_l|t,l). \qquad (33)$$

Under the discriminant classification approach, the histograms (32) corresponding to each cell $\mathcal{I}_l$ are concatenated to obtain a single vector from a set of local bags.

4. Experimental Evaluation

A number of experiments were performed to evaluate the classification performance of the proposed Bayesian extensions to generative and discriminative classification. These experiments are based on standard datasets, namely 15 Scenes[17, 10], Label Me [21] and UIUC Sports [18] so as to enable comparison with the state-of-the-art in image classification. In all cases, classification was based on the 128 dimensional SIFT descriptor of [20], sampled on a dense grid of spacing 8 on gray scale images.

4.1. Bayesian mid-level representation

The first set of experiments was designed to evaluate the effectiveness of the proposed representations, based on Bayesian adaptation, as mid-level image representations for image classification. For this, we compared class-adapted GMMs, here denoted as *adapted GMMs* (AGMM), to standard GMMs and the BoW representation.

Table 1. Evaluation of mid-level representations.

Dataset	Model	Accuracy
15 Scenes	AGMM	**83.2**
	GMM	81.4
	BoW	79.3
Label Me	AGMM	**86.4**
	GMM	85.7
	BoW	85.6
8 Sports	AGMM	**82.5**
	GMM	80.4
	BoW	80.4

In all experiments, the background GMM of (12) was learnt from all training images, with the EM algorithm of Section 2.2. This model was then adapted to each of the classes, using the specific training sets and the EM algorithm of Section 3.1, resulting in one AGMM per class. Images were finally classified using the Bayes decision rule of (1). Classification performance is compared to those of two standard methods. The first is the equivalent classifier without model adaptation [5], where a GMM was learned per class, using the EM algorithm of Section 2.2 and the images were classified with the Bayes decision rule (1). The second used an SVM with the BoW model [6]. In this case, the Gaussian means of the learnt background GMM were used as codewords for image quantization. The histogram of the resulting visual words was then fed to an SVM classifier, using an histogram intersection kernel. It is worth emphasizing that the experiments of this section do not rely on spatial information.

The average (per-class) classification accuracy of the three methods is presented in Table 1, for 15 Scenes, Label Me, and Sports. The table supports a few conclusions. First, AGMM has superior performance than GMM and the BoW approach. Second, although popular, the BoW representation is not very effective. The use of a single global model seems to eliminate much of the discriminant information needed for accurate image classification. Third, the available amounts of training data are not sufficient to guarantee a GMM of sufficient generalization power, when learning is performed from each class individually. For all three datasets, model adaptation produces non-trivial gains in classification rate.

4.2. AGMM-SP Classification

We next considered the advantages, for generative classification, of combining model adaptation and encoding of spatial information. As discussed in Section 3.4, a spatial pyramid classifier was designed by adapting the background GMM to different spatial bins for each topic. This classi-

Table 2. Impact of spatial information coding on generative classification with AGMMs.

Dataset	AGMM	AGMM-SP
15 Scenes	83.2	**84.4**
Label Me	86.4	**87.4**
8 Sports	82.5	**82.9**

fier is denoted AGMM-SP, for *adapted GMM with spatial pyramid*. Table 2 presents a comparison of the classification accuracy of AGMM-SP, with three spatial levels, and that of AGMM (no spatial resolution). As previously found by many authors, the use of spatial information improves classification performance. We note, however, that in this case the gains of spatial encoding are not very large. This is probably due to the fact that the AGMM representation already combines generalization and class-specificity.

4.3. Image classification

The previous experiments establish AGMM as an interesting mid-level representation for image classification. We next performed an evaluation of image classifiers developed from this representation. For generative classification, we used the AGMM-SP classifier of the previous section. For discriminant classification, we used a classifier based on the histogram representation of Section 3.3 and an SVM with a spatial pyramid. The latter is denoted the PH-SVM-SP classifier, for *predictive histogram SVM with spatial pyramid*. Codeword lengths were set to 1024. This representation is similar to that of BoW. We experimented with histogram intersection and linear kernel SVMs, and attempted both average and max pooling strategies. On 15 Scenes and Label Me, we found that the combination of linear kernel and max pooling, proposed by [27], achieved the best results. On UIUC sports, the best results were obtained with an histogram intersection kernel and average pooling.

Tables 3 and 4, compare the classification accuracy of the proposed generative and discriminant classifiers to those of various recently proposed methods. Since not all previous methods have been applied to the three datasets, the accuracies of some of them are only reported for some of the datasets. In particular, Table 3 reports to UIUC Sports and 15 Scenes, and Table 4 to Label Me. A number of conclusions are possible. First, the PH-SVM-SP classifier outperforms that based on AGMM-SP. Second, both methods achieve the best results reported in the literature on Sports and Label Me. It should be noted that, on Sports, one of the closest competitors [26] uses an histogram intersection kernel to learn the dictionary of visterms. In the implementation of [26], this is combined with specialized features, such as CENTRIST or sPACT. For fairness, we compare to their implementation based on SIFT (denoted HIK-CBK in the table). With a single scale SIFT descriptor, we achieve an accuracy of 83.5% with PH-SVM-SP and 82.9% with AGMM-SP. If we sample SIFT descriptors at 5 scales as in [26], the performance of PH-SVM-SP increases to 84.4% as shown in Table 3.

Table 3. Classification accuracy of various methods on 15 scenes and Sports. HIK-CBK indicates the implementation of [26] using SIFT. Accuracy in [14] was obtained from graph.

Method	15 Scenes	Sports
PH-SVM-SP	**85.4**	**84.4**
AGMM-SP	84.4	82.9
Boureau *et al.* [4](Maco-SIFT)	85.6	-
Zhou *et al.* [28](HG)	85.2	-
Boureau *et al.* [4](Sp-SIFT)	84.1	-
SPMK [17]	81.2	-
ScSPM [27]	80.5	-
Kernel Codebook [14]	77	-
Wu *et al.* [26](HIK-CBK)	78.54	81.17
Fei-Fei [18]	-	73.4

Table 4. Classification accuracy of various methods on Label Me. Note that [3] reports higher performance, but using color-SIFT descriptors.

Method	Accuracy
PH-SVM-SP	**88.3**
AGMM-SP	87.4
HDP-HMT [16]	84.5
“gist” [21]	83.7
pLSA [3]	82.5
Contextual Ancestry [19]	82

Finally, on 15 Scenes, the performance of PH-SVM-SP and AGMM-SP is comparable to the current state-of-the-art. In [4] a discriminative dictionary is obtained, using an extension of the sparse coding method of [27]. Their result is slightly better, but while using macro-features. We have not pursued this type of extensions, although it could be incorporated easily in our approach. A comparison, therefore, to the best results using SIFT features and sparse coding (denoted Sp-SIFT) [4], which the proposed PH-SVM-SP considerably outperforms, is more fair. Furthermore, sparse coding requires the solution of an optimization problem to derive mid-level representations, and is substantially more complex than the schemes investigated in this work.

5. Conclusion

Many conclusions are possible from the results. We have shown that model adapted GMMs are superior in performance to ML-GMMs and globally orderless Bag-of-words models. This supports the previous claims that hard quantization in the feature space results in a loss of discriminative power [2]. The improvement over non adapted GMMs shows that a Bayesian approach to learning topic models provides the needed generalization beyond the training set. We have extended the idea of incorporating local spatial information, introduced in [17], to the Bayesian Classification framework and observed an improved performance with AGMMs. A novel predictive histogram representation that builds upon Bayesian inference of topic AGMMs was introduced and was shown to achieve even better classification performance. Both the proposed methods are shown to be competent with the previously published results in the literature.

References

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

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 2008.

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

[4] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559 –2566, jun. 2010.

[5] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 163 – 168 vol. 2, jun. 2005.

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, B, 39, 1-38*, 1977.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2 edition, November 2001.

[9] L. Fe-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1134 –1141 vol.2, 2003.

[10] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524 – 531 vol. 2, jun. 2005.

[11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proceedings of the 10th International Conference on Computer Vision, Beijing, China*, volume 2, pages 1816–1823, Oct. 2005.

[12] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *Int. J. Comput. Vision*, 71(3):273–303, 2007.

[13] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. In *Speech and Audio Processing, IEEE Transactions on*, volume 2, pages 291 –298, apr. 1994.

[14] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, pages 696–709, Berlin, Heidelberg, 2008. Springer-Verlag.

[15] T. Hofmann. Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, pages 289–296, 1999.

[16] J. Kivinen, E. Sudderth, and M. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, 2007.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169 – 2178, 2006.

[18] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[19] J. Lim, P. Arbela andez, C. Gu, and J. Malik. Context by region ancestry. In *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints, 2003.

[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] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1889 –1895, 2009.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19 – 41, 2000.

[24] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proc. Computer vision and pattern recognition*, pages 216–221, 2000.

[25] P. C. Woodland. Speaker adaptation for continuous density HMMs: A review. In *ITRW on Adaptation Methods for Speech Recognition*, pages 11–19, Aug. 2001.

[26] J. Wu and J. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630 –637, sep. 2009.

[27] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009.

[28] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang. Hierarchical gaussianization for image classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1971 – 1977, sep. 2009.