A Similarity Measure Between Unordered Vector Sets with Application to Image Categorization

Yan Liu*and Florent Perronnin
Xerox Research Centre Europe (XRCE)
Textual and Visual Pattern Analysis (TVPA)
Firstname.Lastname@xrce.xerox.com
http://www.xrce.xerox.com

Abstract

We present a novel approach to compute the similarity between two unordered variable-sized vector sets. To solve this problem, several authors have proposed to model each vector set with a Gaussian mixture model (GMM) and to compute a probabilistic measure of similarity between the GMMs. The main contribution of this paper is to model each vector set with a GMM adapted from a common "universal" GMM using the maximum a posteriori (MAP) criterion. The advantages of this approach are twofold. MAP provides a more accurate estimate of the GMM parameters compared to standard maximum likelihood estimation (MLE) in the challenging case where the cardinality of the vector set is small. Moreover, there is a correspondence between the Gaussians of two GMMs adapted from a common distribution and one can take advantage of this fact to compute efficiently the probabilistic similarity. This work is applied to the image categorization problem: images are modeled as bags of low-level features and classification is performed using a kernel classifier based on the proposed similarity measure. Experimental results on the PASCAL VOC 2006 and VOC 2007 databases show the excellent performance of our approach.

1. Introduction

There exist several pattern analysis problems where the objects of interest can be represented by unordered vector sets of variable cardinality. For instance, in computer vision, images are often represented as bags of low-level local feature vectors. Performing such tasks as retrieval or kernel-based learning on these representations requires the definition of a suitable measure of similarity. The application of interest in this work is the categorization of images. While categorization systems based on the bag-of-features representation neglect the absolute or relative position of the feature vectors in the image, they have shown excellent performance on several benchmarks [6, 5, 4].

There are two broad classes of approaches to measure the similarity of vector sets: *model-free* approaches seek a direct measure of similarity while *model-based* approaches first estimate the distribution of a vector set and then measure the similarity between distributions.

A typical model-free approach is the pyramid match kernel of Grauman and Darrell [11, 12]. The idea is to partition the feature space in a hierarchical manner and to count the number of correspondences between the two vector sets at each level of the hierarchy.

Model-based approaches can themselves be divided into two sub-classes as a vector set can be modeled with a *discrete* or a *continuous* distribution. In the first case, one makes use of an intermediate representation, generally referred to as a visual vocabulary in the context of images, which is obtained offline through the clustering of a large number of vectors. Each image is characterized by a histogram of visual word frequencies [21, 2]. One of the main limitations of this approach is the assumption that the distribution of the features that can be encountered by the system is known a priori.

Of particular interest to us in this work are those approaches which model a vector set with a continuous distribution, generally a Gaussian mixture model (GMM) [13, 14, 10, 18, 22, 23]. The most commonly used measures of similarity between two GMMs are the Kullback-Leibler divergence (KLD) [10, 18, 22, 23] and the probability product kernel (PPK) [13, 14]. These methods have however two main shortcomings.

First, to model accurately a vector set, one needs to train a sufficiently large number of Gaussians. The robust estimation of the GMM parameters may be difficult if the cardinality of the vector set is small. For instance, the number

*Yan Liu is a Ph.D. student in the Laboratoire d'Informatique en Image et Systmes d'Information (LIRIS) at the Ecole Centrale de Lyon (ECL).

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

of local features extracted from an image typically varies from a few hundreds up to a few thousands. One could increase this number, e.g. by using a denser grid in the case of regular extraction or by lowering the detection threshold in the case of interest point detectors, but this would also significantly increase the cost of the feature extraction and the GMM estimation.

Second, as there is no closed form solution for the KLD or the PPK between two GMMs (except in special cases), approximate solutions have to be found. [18, 23] approximate the KLD using Monte Carlo (MC) sampling. However, the cost of this method is prohibitive as one has to randomly draw a large number of samples to obtain a reasonable estimate. [10] uses the the unscented transform, a deterministic approach which is reminiscent of MC sampling. As the cost of this method is quadratic in the number of Gaussians, it is impractical when the number of Gaussians is large. Goldberger *et al.* [10] and Vasconcelos [22] proposed two very similar approximations of the KLD. They are based on a two-step approach: first find a matching between the Gaussians of the two distributions and then compute the KLD between the pairs of matched Gaussians. The cost of these approaches is still quadratic in the number of Gaussians.

The main contribution of this work is *to model each vector set with a GMM adapted from a common "universal" GMM* using the maximum a posteriori (MAP) criterion. This offers two main advantages:

- First, *MAP estimation is more accurate than MLE* in the challenging case where the training data is scarce as the universal model provides a priori information on the location of the parameters in the whole parameter space. We will show experimentally that this a priori information needs not to be exact: even if the universal model is learned on a set of images which is not directly related to the task at hand, excellent performance is obtained.

- Second, if two GMMs are adapted from a common distribution, there is a one-to-one correspondence between their Gaussians. We make use of this correspondence to derive approximations of the PPK and KLK with a *cost linear in the number of Gaussians*.

Note that the idea of learning visual vocabularies - modeled as GMMs - through the adaptation of a common universal vocabulary has already been used in [19]. However, in [19] the adapted vocabularies are *class-GMMs* and images are modeled with histograms of visual-word occurrences while in this article the adapted vocabularies are *image-GMMs*.

The remainder of the paper is organized as follows. In section 2 we describe the estimation of universal and adapted image models. In section 3 we present two similarity measures between distributions, the KLD and the PPK, and explain how they can be approximated in our case. Then in section 4 we provide experimental results on the PASCAL VOC 2006 and VOC 2007 databases and show the excellent performance of our system. Finally we draw conclusions in section 5.

2. Images as Adapted Mixtures of Gaussians

Let us first introduce our notation. The parameters of a GMM are denoted $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1...N\}$ where w_i, μ_i and Σ_i are respectively the weight, mean vector and covariance matrix of Gaussian i and N denotes the number of Gaussians. Let x be an observation vector and q its associated hidden variable. The likelihood that observation x was generated by the GMM is:

$$p(x|\lambda) = \sum_{i=1}^{N} w_i p_i(x|\lambda) \; . \tag{1}$$

where $p_i(x|\lambda) = p(x|q = i, \lambda)$. Finally, $\gamma_i(x) = p(q = i|x, \lambda)$ is the occupancy probability, i.e. the probability that observation x was generated by Gaussian i. It is computed using Bayes formula:

$$\gamma_i(x) = \frac{w_i p_i(x|\lambda)}{\sum_{j=1}^{N} w_j p_j(x|\lambda)} \; . \tag{2}$$

We now describe the training of the universal model and the adapted image models.

2.1. Training the universal model

The universal GMM is supposed to describe the content of any image and, therefore, it should be trained offline on a varied set of images. Let λ^u denote the parameters of the universal GMM. Let $X = \{x_t, t = 1...T\}$ be the set of training vectors. The estimation of λ^u may be performed by maximizing the log-likelihood function $\log p(X|\lambda^u)$. The standard procedure for MLE is the Expectation Maximization (EM) algorithm [3]. For the E-step, the values $\gamma_i(x_t)$ are computed. We provide here for completeness the M-step re-estimation equations [1]:

$$\hat{w}_i^u = \frac{1}{T} \sum_{t=1}^{T} \gamma_i(x_t) \; , \tag{3}$$

$$\hat{\mu}_i^u = \frac{\sum_{t=1}^{T} \gamma_i(x_t) x_t}{\sum_{t=1}^{T} \gamma_i(x_t)} \; , \tag{4}$$

$$\hat{\Sigma}_i^u = \frac{\sum_{t=1}^{T} \gamma_i(x_t) x_t x_t'}{\sum_{t=1}^{T} \gamma_i(x_t)} - \hat{\mu}_i^u \hat{\mu}_i^{u\prime} \; . \tag{5}$$

2.2. Training adapted image models

Our primary motivation for learning the image GMMs through the adaptation of a universal model is to overcome the scarcity of the training material. Indeed, only a small number of low-level feature vectors (typically from a few hundreds up to a few thousands) are extracted from one image. We will observe in section 4 that this is insufficient to train robustly a mixture with a large number of Gaussians (*e.g.* 100) for each image. In the following, λ^a denotes the parameters of an adapted model.

Let $X = \{x_t, t = 1...T\}$ now denote the set of adaptation samples extracted from one image. We use the MAP criterion to adapt a GMM. The goal of MAP estimation is to maximize the posterior probability $p(\lambda^a|X)$ or equivalently $\log p(X|\lambda^a) + \log p(\lambda^a)$. Hence, the difference with MLE is in the assumption of a prior distribution $p(\lambda^a)$. To perform MAP learning, one has to (i) choose the prior distribution family and (ii) specify the parameters of the prior distribution.

It was shown in [9] that the prior densities for GMM parameters could be adequately represented as a product of Dirichlet (prior on weight parameters) and normal-Wishart densities (prior on Gaussian parameters). When adapting a universal model with MAP to more specific conditions, it is natural to use the parameters of the universal model as a priori information on the location of the adapted parameters in the parameter space. As shown in [9], one can also apply the EM procedure for MAP estimation. The M-step re-estimation equations are provided here for completeness:

$$\hat{w}_i^a = \frac{\sum_{t=1}^T \gamma_i(x_t) + \tau}{T + N \times \tau} , \quad (6)$$

$$\hat{\mu}_i^a = \frac{\sum_{t=1}^T \gamma_i(x_t) x_t + \tau \mu_i^u}{\sum_{t=1}^T \gamma_i(x_t) + \tau} , \quad (7)$$

$$\hat{\Sigma}_i^a = \frac{\sum_{t=1}^T \gamma_i(x_t) x_t x_t' + \tau[\Sigma_i^u + \mu_i^u \mu_i^{u\prime}]}{\sum_{t=1}^T \gamma_i(x_t) + \tau} - \hat{\mu}_i^a \hat{\mu}_i^{a\prime} . \quad (8)$$

The *relevance factor* τ keeps a balance between the a priori information contained in the generic model λ^u and the new evidence contained in X. If a mixture component i was estimated with a small number of observations $\sum_{t=1}^T \gamma_i(x_t)$, then more emphasis is put on the a priori information. On the other hand, if it was estimated with a large number of observations, more emphasis is put on the new evidence. Hence MAP provides a more robust estimate than MLE when little training data is available. The parameter τ is generally set manually [9, 20].

For a given number of Gaussians, the cost of one EM iteration is (almost) identical for MLE and MAP. The only difference is the addition in the M-step of MAP of the a priori information in the statistics (compare equations 3, 4 and 5 to 6, 7 and 8 resp.) However, as MAP uses some a priori information on the location of the parameters, it requires a smaller number of EM iterations to reach an accurate estimate. Therefore, it is *significantly faster* compared to MLE. This statement will be verified experimentally.

We finally note that an adapted model contains the same number of Gaussians as the universal model from which it is adapted.

3. Measuring the Similarity of GMMs

In the following, we present two measures of similarity between distributions and show how to approximate them in our case.

3.1. Probability Product Kernel

The probability product kernel (PPK) [14] between probability distributions p and q is defined as follows:

$$K_{ppk}^\rho(p, q) = \int_{x \in \Omega} p(x)^\rho q(x)^\rho dx . \quad (9)$$

The PPK has two special cases. When $\rho = 1$, the PPK takes the form of the expectation of one distribution under the other:

$$K_{elk}(p, q) = E_p[q(x)] = E_q[p(x)] . \quad (10)$$

This is referred to as the *Expected Likelihood Kernel* (ELK) [14]. When $\rho = 1/2$, it is known as the *Bhattacharyya Kernel* (BK).

There is a closed form solution for the PPK between two Gaussians:

$$K_{ppk}^\rho(p, q) = (2\pi)^{(1-2\rho)D/2} \mid \Sigma \mid^{1/2} \mid \Sigma_p \mid^{-\rho/2} \mid \Sigma_q \mid^{-\rho/2} \exp\left(-\frac{\rho}{2} \mu_p^\top \Sigma_p^{-1} \mu_p - \frac{\rho}{2} \mu_q^\top \Sigma_q^{-1} \mu_q + \frac{1}{2} \mu^\top \Sigma \mu \right) , \quad (11)$$

where $\Sigma = (\rho\Sigma_p^{-1} + \rho\Sigma_q^{-1})^{-1}$, $\mu = \rho(\Sigma_p^{-1}\mu_p + \Sigma_q^{-1}\mu_q)$ and D is the dimensionality of the feature vectors.

However there is no closed form solution for the PPK in the case of mixtures of Gaussians (except for the special case $\rho = 1$). In the case of a mixture model, we have $p(x) = \sum_{i=1}^N \alpha_i p_i(x)$ and $q(x) = \sum_{j=1}^M \beta_j q_j(x)$. In [13] (section 4) the following approximation is suggested:

$$K_{ppk}^\rho(p, q) \approx \sum_{i=1}^N \sum_{j=1}^M \alpha_i \beta_j K_{ppk}^\rho(p_i, q_j) . \quad (12)$$

When $\rho \leq 1$ this approximation corresponds to an upper-bound on the true value of $K_{ppk}^\rho(p, q)$ and when $\rho \geq 1$ it is a lower-bound.

The evaluation of the PPK between two GMMs which contain respectively M and N Gaussians requires the computation of M×N PPKs between individual Gaussians. This

cost may be a handicap in the case of large values of M and N. We thus make use of the fact that two mixtures of Gaussians have been adapted from the same generic model to speed-up the computation. Indeed, Reynolds *et al*. [20] first noticed that there is a one-to-one correspondence between the i-th Gaussian of an adapted GMM and the i-th Gaussian of the GMM it is adapted from. By transitivity, it means that there is a one-to-one correspondence between the i-th Gaussians of two GMMs adapted from the same GMM (we recall that we necessarily have $M = N$ in our adaptation framework). Consequently, in our case, the terms $K^{\rho}_{ppk}(p_i, q_i)$ dominate the previous sum and the PPK may be further approximated as follows:

$$K^{\rho}_{ppk}(p, q) \approx \sum_{i=1}^{N} \alpha_i \beta_i K^{\rho}_{ppk}(p_i, q_i) \; . \tag{13}$$

This evaluation requires only the computation of N PPKs between individual Gaussians.

3.2. Kullback-Leibler Kernel

The Kullback-Leibler Divergence (KLD) between two continuous distributions is defined as follows:

$$KL(p||q) = \int_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} dx \; . \tag{14}$$

There is also a closed form solution for the KLD between two Gaussians:

$$KL(p||q) = \frac{1}{2} \Bigg[\log \frac{\mid \Sigma_q \mid}{\mid \Sigma_p \mid} + \mathrm{Tr}(\Sigma_q^{-1} \Sigma_p) + (\mu_p - \mu_q)^T \Sigma_q^{-1} (\mu_p - \mu_q) - D \Bigg] \; . \tag{15}$$

However, no closed-form expression exists for the KLD between two GMMs.

We follow the method of [10]. This approximation first consists in finding a mapping π from the Gaussians of p to the Gaussians of q as follows:

$$\pi(i) = \arg\min_j \left(KL(p_i||q_j) - \log \beta_j \right) \; . \tag{16}$$

Then π is used to approximate the KLD:

$$KL(p||q) \approx \sum_{i=1}^{N} \alpha_i \left(KL(p_i||q_{\pi(i)}) + \log \frac{\alpha_i}{\beta_{\pi(i)}} \right) \; . \tag{17}$$

This approximation is well motivated when Gaussians have little overlap, e.g. when the dimensionality D of the feature space is high. In our experiments, $D = 50$ (c.f. section 4.1).

If two GMMs contain respectively M and N Gaussians, computing the mapping function π requires the computation of M×N KLDs between individual Gaussians. Once again, we can make use of the fact that there is a one-to-one correspondence between the Gaussians of two GMMs adapted from the same model to perform the following approximation: $\pi(i) = i$. Under this assumption, the KLD can be rewritten:

$$KL(p||q) \approx \sum_{i=1}^{N} \alpha_i \left(KL(p_i||q_i) + \log \frac{\alpha_i}{\beta_i} \right) \; . \tag{18}$$

Hence, the computation of the KLD requires only N Gaussian computations in our case.

The Kullback-Leibler Kernel (KLK) can then defined by exponentiating the symmetric KLD $SKL(p, q)$:

$$SKL(p, q) = KL(p||q) + KL(q||p) \tag{19}$$

$$K_{klk}(p, q) = \exp\left(-\gamma SKL(p, q)\right) \; . \tag{20}$$

When choosing the value γ, one should take care that the kernel matrix is positive definite to ensure that K_{klk} is a true kernel.

4. Experimental Results

We first describe our experimental setup. We then report results on two challenging datasets: the PASCAL VOC 2006 and VOC 2007 databases.

4.1. Experimental setup

Low-level feature vectors are extracted on regular grids at multiple scales in our experiments. On the average, on the order of $1,000$ feature vectors are extracted per image per feature type. We make use of two types of low-level features. The first features are based on local histograms of orientations as described in [17] (128 dimensional features). The second ones are based on RGB statistics (96 dimensional features). In both cases, the dimensionality of the feature vectors is reduced to 50 through Principal Component Analysis (PCA).

The universal GMM is trained using the following iterative strategy inspired by HTK [24]. We first train a GMM with a single Gaussian. We then split it into two by introducing a small perturbation in the mean parameter and retrain the GMM using several iterations of EM. The process of splitting and retraining is repeated until the desired number of Gaussians is obtained. To train the adapted image GMMs with MAP, the default value for the relevance factor is $\tau = 10$.

For the PPK, we choose $\rho = 1/2$ (*i.e*. the Bhattacharyya Kernel) as this value lead to the best results in preliminary experiments. To set parameter γ for the KLK (c.f. equation 20) we followed [25]: γ is equal to the inverse of the mean of the symmetric KL divergence $SKL(p, q)$ between two GMMs (c.f. equation 19) as estimated on a subset of the whole training set.

For classification we experimented with kernel logistic regression (KLR) and support vector machines (SVM). As we obtained very similar performance with both classifiers, in the following we report only the KLR results. One linear classifier is trained per class in a one-against-all manner.

We have two separate systems: one for each feature type. The end result is the average of the scores of the two systems.

4.2. VOC 2006 database

The PASCAL VOC 2006 database [5] consists of 10 object classes: bicycle, bus, car, cat, cow, dog, horse, motorbike, person and sheep. There are 2,618 images for training and 2,686 for testing. During the VOC 2006 competition, the accuracy was primarily measured with the Area under the Curve (AUC). Therefore we also use the AUC (averaged over the 10 categories) to make our results easily comparable to the state-of-the-art.

In the following, we start with a comparative evaluation of the proposed approach. We then proceed with the analysis of the influence of parameter τ. We also carry out cross-database experiments showing that, even if the universal model is learned on a different database, the performance does not vary significantly. Finally, we analyze the computational cost of the proposed method on this database.

4.2.1 Comparative evaluation

We compare the performance of three systems:

(i) The proposed approach with MAP adaptation and the fast one-to-one mapping of Gaussian components (c.f. formula 13 for PPK and formula 18 for KLK). This system is later referred to as MAP_OTO.

(ii) A system which learns the image GMMs with MLE (using the same iterative strategy which was employed to train the universal model) and the slow one-to-many mapping of Gaussian components (c.f. formula 12 for PPK and formulae 16 and 17 or KLK). This system is later referred to as MLE_OTM.

(iii) An intermediate system which makes use of MAP adaptation as is the case of (i) but which uses the slow one-to-many scoring of (ii). This system is later referred to as MAP_OTM.

Hence, when comparing (ii) and (iii), we can measure the benefit of MAP compared to MLE. When comparing (i) and (iii) we can measure the impact on the accuracy of the fast one-to-one scoring versus the slow one-to-many scoring.

Results are provided on figure 1. We can draw the following conclusions. First, MAP clearly outperforms MLE for both PPK and KLK. Especially the performance of the

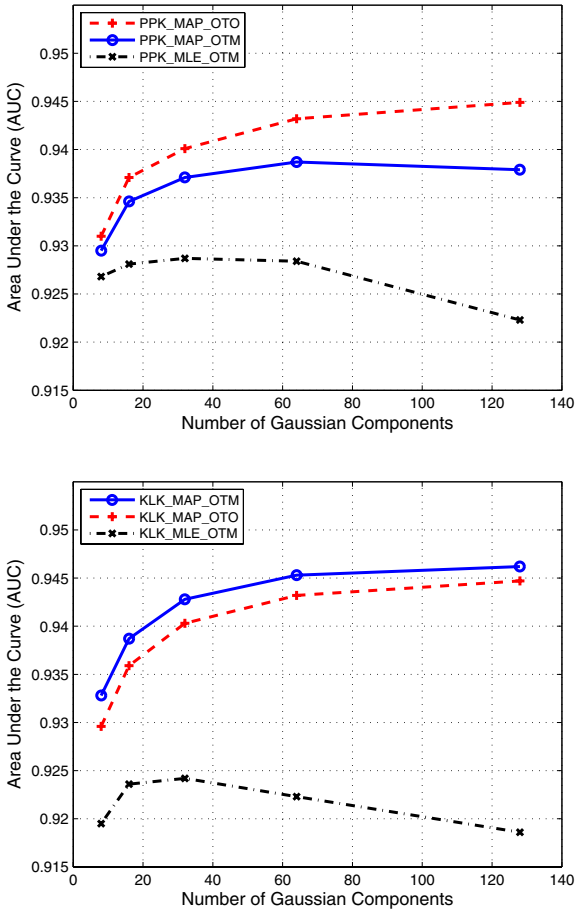

Figure 1. Performance on the PASCAL VOC 2006 database of the PPK (top) and KLK (bottom) for a varying number of Gaussian components.

MLE_OTM system starts to drop for more than 32 Gaussians while for MAP it continues to increase. This shows that we can learn robustly a larger number of Gaussians with MAP than with MLE. Second, the accuracy of PPK_OTO is superior to that of PPK_OTM. This observation came as a surprise as we first thought that by dropping terms in equation 12, we would lose information. Our best explanation is that the bound 12 is too coarse an approximation of the PPK. This suggests an alternative approach for computing the PPK similar to that used for KLK: first find a matching between the Gaussians of p and q and then approximate the PPK as a weighted sum of PPKs between the matched Gaussians. This approximation might be worth testing in the future. Third, the accuracy of KLK_OTO is inferior to that of KLK_OTM, but not significantly so, showing that our one-to-one approximation is a good one. Finally, PPK_OTO and KLK_OTO perform very similarly: the best results we obtained for both kernels was 0.945 for

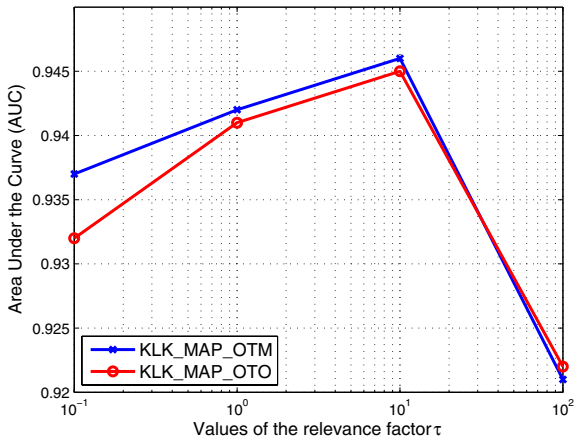

Figure 2. Average AUC on the VOC 2006 database as a function of parameter τ.

128 Gaussian components. To the best of our knowledge, the best average AUC reported so far on this database was 0.936 (this was the accuracy of the winning QMUL_LSPCH system [5]).

4.2.2 Influence of MAP parameter τ

We now analyze the influence of the relevance factor τ. τ impacts two competing aspects of our system:

- First τ *influences the robustness of the estimation*. We can consider two extreme cases. When $\tau = 0$, MAP turns into MLE and the parameters are not estimated robustly as was shown in the previous experiments. When $\tau = \infty$, the image GMMs remain equal to the universal model. As the distance between any pair of images is constant no kernel classifier can be learned. The best performance will thus be obtained when an intermediate value between these two extremes is chosen.

- Second τ *impacts the proposed fast scoring*. Indeed, our fast scoring is only possible if there is a one-to-one correspondence between the Gaussians of two adapted GMMs. The strength of the correspondence will depend on τ. If $\tau = \infty$, the correspondence is maximized and the one-to-one mapping is exact. When $\tau = 0$, the correspondence is minimal.

Hence, the τ which optimizes the robustness ($0 < \tau < \infty$) is necessarily different from the τ which optimizes the Gaussian correspondence ($\tau = \infty$).

We present the result in figure 2. This analysis was performed on the 128 Gaussians model using the KLK kernel. We can see that for small values of τ MAP_OTM outperforms MAP_OTO. This shows that, when τ is small the correspondence between the Gaussian of two adapted models is loose and that our one-to-one assumption is too naïve. However, as expected, as τ increases to more reasonable values, the difference between the two systems becomes narrower. For both systems the best performance is obtained for $\tau = 10$.

4.2.3 Cross database experiments

As the estimation of the image models with MAP relies on the a priori information contained in the universal model, it is important to understand how the performance of our approach is affected when the universal model is learned on another dataset. The alternate dataset we used to learn the visual vocabulary contains 120,000 unannotated images from a printing workflow of photo albums. We had a look at a small sample of these images to try to understand whether they were representative of the 10 categories found in the VOC 2006 database. While this set of images contains a very large number of photos of persons, it seems to contain very few (if no) occurrences of the 9 other classes. Hence, we believe that there is a strong mismatch between this dataset and VOC 2006. To learn a universal vocabulary, we took a random sub-sample of 2,000 images. This experiment was repeated 10 times with 10 different subsamples. We restricted this analysis to the case where we employ the fast scoring.

For both the PPK and the KLK kernels, the AUC (averaged over the 10 runs) did not change (0.945). What is interesting to notice is also that the AUC variation was very small from one sub-sample to another one: over the 10 runs, the worst performance obtained was 0.944 for PPK and 0.943 for KLK respectively.

Clearly, the proposed approach does not seem to be sensitive to the set of images used to train the universal model. Hence the same universal model can be used across different category sets. This is a clear advantage when one grows a category set incrementally as one does not need to relearn the universal GMMs, and thus the image GMMs, every time a new category is added.

4.2.4 Computational cost

We now perform a brief analysis of the computational cost of the proposed approach. For this analysis, we considered GMMs containing 128 Gaussians. The following durations were measured on a 2.4 GHz Opteron^{TM}machine.

The cost of training the GMM of one image with MLE using the iterative strategy of [24] is approximately 850 ms while it is only 30 ms for MAP. We recall that this difference is due to the greater number of EM iterations required for MLE compared to MAP. Note that, instead of the iterative

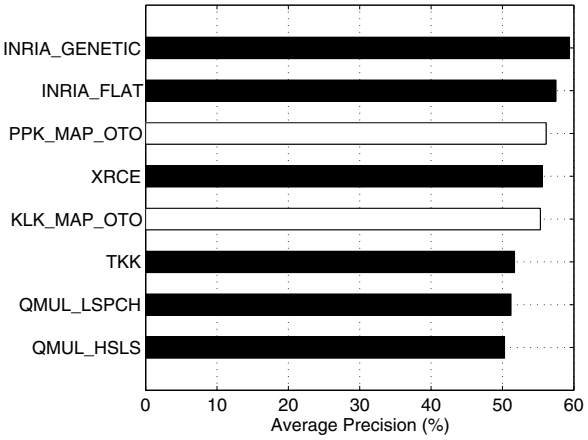

Figure 3. Comparison on the VOC 2007 database of the proposed approach (in white) with the leading participants (in black).

approach of [24], we could have used the alternative strategy which consists in starting from multiple random initializations of the parameters and picking the best one, *i.e*. the one which leads to the highest log-likelihood. However, the cost of this alternative would have been even greater.

We now consider the cost of the kernel computations. On the VOC 2006 database, classifying one image takes approximately 140 s for the PPK and 30 s for the KLK using the one-to-many scoring. With the proposed fast scoring based on the one-to-one correspondence, the classification cost is reduced to 1.3 s for PPK and 0.4 s for KLK. These figures are consistent with the fact that, for both kernels, we expect the proposed one-to-one scoring to be 128 times faster than the one-to-many scoring when GMMs contain 128 Gaussians (linear versus quadratic cost). It is also interesting to note that for both the fast and slow approaches, KLK is almost 5 times faster to compute than PPK. We also carried-out small-scale experiments using MC sampling to approximate the PPK and KLK. However, using MC sampling with 1,000 samples (a rather modest number) we estimated it would take on this database approximately 240 s to classify one image from this database.

4.3. VOC 2007 database

The PASCAL VOC 2007 database [4] contains a total of 9,963 images: 5,011 images for training and 4,952 for testing. There are twenty object classes in this database: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa and tv monitor. During the VOC 2007 competition, the accuracy was primarily measured with the Average Precision (AP). Therefore, we also use the AP (averaged over the 20 categories) to make our results easily comparable to the state-of-the-art.

For these experiments, we used the proposed MAP_OTO approach. Figure 3 shows that the performance of our systems (0.561 for PPK and 0.553 for KLK) is comparable to the performance obtained by the leading participants (the best reported result is 0.594). More details on the competition can be found in [4].

5. Conclusion

In this article, we introduced a novel approach to compute the similarity between two unordered vector sets. The main contribution was to model each vector set with a generative model – a GMM in our case – adapted from a common universal model using MAP. We showed that this adaptation framework offers two major advantages compared to the case where the distributions are trained with MLE. First MAP provides a more accurate estimate compared to MLE when the cardinality of the vector sets is small. Second, there is a one-to-one correspondence between the components of adapted mixture models which may be used for fast scoring. This correspondence was used to derive efficient approximations for two kernels on distributions: the probability product kernel and the Kullback-Leibler kernel.

This approach was applied to the image categorization problem and it exhibited state-of-the-art results on the PASCAL VOC 2006 and VOC 2007 databases. We also showed that this approach is very practical. First, the classification cost is very reasonable. Second, the a priori information contained in the universal model needs not to be perfectly representative of the category set under consideration to obtain good results.

Future work could consider the use adaptation techniques other than MAP. Especially, techniques such as maximum likelihood linear regression (MLLR) [16, 7], cluster adaptive training (CAT) [8] or "eigenvoices" [15] have been shown to yield significantly better results than MAP in the speech recognition literature when the amount of adaptation data is extremely scarce.

References

[1] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical report, Department of Electrical Engineering and Computer Science, U.C. Berkeley, 1998.

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

[3] A. Dempster, N. Laird, and D. Ruin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society*, 39(Series B):1–38, 1977.

[4] M. Everingham, L. V. 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/index.html.

[5] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The PASCAL visual object classes challenge 2006 results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

[6] M. Everingham, A. Zisserman, C. Williams, L. V. Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 PASCAL visual object classes challenge. In *In Selected Proceedings of the First PASCAL Challenges Workshop*, 2006.

[7] M. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98, 1998.

[8] M. Gales. Cluster adaptive training of hidden Markov models. *IEEE Trans. on Speech and Audio Processing*, 8(4):417–428, 2000.

[9] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on speech and Audio Processing*, 2(2):291–298, 1994.

[10] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *ICCV*, volume 1, pages 487–493, 2003.

[11] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *ICCV*, 2005.

[12] K. Grauman and T. Darrell. Approximate correspondences in high dimensions. Technical Report MIT-CSAIL-TR-2006-045, MIT, 2006.

[13] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *COLT*, pages 57–73, 2003.

[14] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *JMLR, Special Topic on Learning Theory*, 5:819–944, 2004.

[15] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. on Speech and Audio Processing*, 8(6):695–707, 2000.

[16] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:806–814, 1995.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia application. In *Neural Information Proceeding Systems*, 2003.

[19] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006.

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

[21] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003.

[22] N. Vasconcelos. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Trans. on Information Theory*, 50(7):1482–1496, 2004.

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

[24] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, S. Povey, V. Valtchev, and P. Woodland. *The HTK book (version 3.2.1)*. Cambridge University Engineering Department, Dec 2002.

[25] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: an in-depth study. Technical Report RR-5737, INRIA, 2005.