Bayesian Localized Multiple Kernel Learning

Mario Christoudias
Raquel Urtasun
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-96
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-96.html

July 3, 2009

Bayesian Localized Multiple Kernel Learning

C. Mario Christoudias, Raquel Urtasun, and Trevor Darrell
UC Berkeley EECS & ICSI
{cmch,rurtasun,trevor}@eecs.berkeley.edu

Abstract

Multiple kernel learning approaches form a set of techniques for performing classification that can easily combine information from multiple data sources, e.g., by adding or multiplying kernels. Most methods, however, are limited by their assumption of a per-view kernel weighting. For many problems, the set of features important for discriminating between examples can vary locally. As a consequence these global techniques suffer in the presence of complex noise processes, such as heteroscedastic noise, or when the discriminative properties of each feature type varies across the input space. In this paper, we propose a localized multiple kernel learning approach with Gaussian Processes that learns a local weighting over each view and can obtain accurate classification performance and deal with insufficient views corrupted by complex noise, e.g., per-sample occlusion. We demonstrate our approach on the tasks of audio-visual gesture recognition and object category classification on the Caltech-101 benchmark.

1 Introduction

Many problems in machine learning involve datasets that are comprised of multiple views. The separate views can be defined over a single input (e.g., multiple image feature types), or from multiple information sources (e.g., audio and video). In this context, each view can provide a redundant indication of the underlying class or event of interest, useful for classification.

Multiple kernel learning approaches to multi-view learning [1, 12] have recently become very popular since they can easily combine information from multiple views, e.g., by adding or multiplying kernels. They are particularly effective when the views are class conditionally independent, since the errors committed by each view can be corrected by the other views. Most methods assume that a single set of kernel weights is sufficient for accurate classification, however, one can expect that the set of features important to discriminate between different examples can vary locally. As a result the performance of such global techniques can degrade in the presence of complex noise processes, e.g., heteroscedastic noise, missing data, or when the discriminative properties vary across the input space.

Recently, there have been several attempts at learning local feature importance. Frome et al. [5] proposed learning a sample-dependent feature weighting, and framed the problem as learning a per-sample distance that satisfies constraints over triplets

of examples. The problem was cast in a max-margin formalism, resulting in a convex optimization problem that is infeasible to solve exactly for large datasets; approximate sampling is typically employed. Lin et. al. [10] learn an ensemble of SVM classifiers defined on a per-example basis for coping with local variability. Similarly, Gonen and Alpaydin [6] proposed an SVM-based localized multiple kernel learning algorithm that learns a piecewise similarity function over the joint input space using a sample-dependent gating function.

In this paper we present a Bayesian approach to multiple kernel learning that can learn a local weighting over each view of the input space. In particular, we learn the covariance of a Gaussian process using a product of kernels: a parametric kernel computed over the input space and a non-parametric kernel whose covariance is rank-constrained and represents per-example similarities in each view. To make learning and inference tractable, we assume a piecewise smooth weighting of the input space that is estimated by clustering in the joint feature space. Unlike [5], in our framework learning can be done exactly for large datasets and is performed across multiple feature types. We exploit the properties of the covariance matrix and propose a simple optimization criteria, when compared to SVM-based approaches [13, 16], that allow us to efficiently learn multi-class problems.

We demonstrate our approach within two very different scenarios: audio-visual user agreement recognition in the presence of complex noise, and object recognition exploiting multiple image features. In audio-visual settings, the views are commonly corrupted by independent, complex noise processes (e.g., occlusions). Within this domain our experiments highlight our approaches ability to achieve accurate classification performance despite noisy audio-visual views containing per-sample occlusions. We also evaluate our approach on an object recognition task, and report improved performance compared to state-of-the-art single- and multi-view methods.

In the remaining sections we detail our approach to multiple kernel learning, present our experimental evaluation, conclude and give directions of future research.

2 Local Multiple Kernel Learning via Gaussian Processes

In this section we present our approach to local multiple kernel learning. Let $\mathbf{X}_i = [\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(V)}]$ be a muti-view observation with V views, and let $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)} \cdots \mathbf{x}_N^{(v)}]^T$ be a set of N observations of view v. Let $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_N]^T$ be the set of labels, and let $\mathbf{f} = [\mathbf{f}_1 \cdots \mathbf{f}_N]^T$ be a set of latent functions. We assume a Gaussian Process (GP) prior over the latent functions such that

$$p(\mathbf{f}|\bar{\mathbf{X}}) = \mathcal{N}(0, \bar{\mathbf{K}}) \tag{1}$$

where $\bar{\mathbf{X}} = [\mathbf{X}^{(1)} \cdots \mathbf{X}^{(V)}]$ is the set of all observations, and $\mathbf{f}$ is the set of latent functions. We use a Gaussian noise model such that $p(\mathbf{Y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$. More sophisticated noise models, e.g., probit, could be used. However, for such models the marginalization of the latent functions $\mathbf{f}$ cannot be done in closed form and one would have to rely on analytical approximations or sampling methods.

Various strategies can be employed to combine the information from multiple observation types; the restriction being that the resulting covariance matrix $\bar{\mathbf{K}}$ has to be positive definite. We construct our covariance as a linear combination of covariance matrices

$$\bar{\mathbf{K}} = \sum_{v}^{V} \mathbf{K}^{(v)} + \sigma^2 \mathbf{I} \tag{2}$$

where $\mathbf{I}$ is the identity matrix. We only need to ensure that the $\mathbf{K}^{(v)}$ are positive definite, since then $\bar{\mathbf{K}}$ will also be positive definite. Note that one could parameterize $\bar{\mathbf{K}}$ as a weighted sum of $\mathbf{K}^{(v)}$, however, this parameterization is redundant since we have not yet placed any restrictions on the form of $\mathbf{K}^{(v)}$.

We are interested in learning a metric, which in our case is equivalent to learning the covariances $\{\mathbf{K}^{(1)}, \cdots, \mathbf{K}^{(V)}\}$. One can try to learn these covariances in a fully non-parametric way, however, this will not make use of the observations. Instead, we construct the covariance for each view using the product of a non parametric kernel, $k_{np}^{(v)}$, and a parametric kernel that is a function of the observations, $k_{p}^{(v)}$, such that

$$K_{ij}^{(v)} = k_{np}^{(v)}(i,j) \cdot k_{p}^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)}) \tag{3}$$

Learning in our framework consists of estimating the hyper-parameters of the parametric covariances $\mathbf{K}^{(v)} = \{k_{p}^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)})\}$, and the elements of the non-parametric covariances $\mathbf{K}_{np}^{(v)}$. The number of parameters to be estimated is $V \cdot (M + N^2)$, with M being the number of hyper-parameters for each parametric covariance. This is in general too large to be estimated in practice when dealing with large datasets. To make learning tractable, we assume low-rank approximations to the non-parametric covariances such that

$$\mathbf{K}_{np}^{(v)} = (\mathbf{g}^{(v)})^T \mathbf{g}^{(v)} \tag{4}$$

where $\mathbf{g} = [\mathbf{g}_1, \cdots, \mathbf{g}_N]^T \in \Re^{m \times N}$, and $m \ll N$. The number of parameters becomes $V \cdot (M + Nm)$. Note that if $m = N$ we have recovered the full non-parametric covariance. In our experiments we use $m = 1$. In this case $g_j^{(v)}$ becomes a scalar that can be interpreted as measuring the confidence of the sample, i.e., if the v-th view of the j-th training example is noisy, $g_j^{(v)}$ will be small.

To further reduce the number of parameters we assume that the examples locally share the same weights and that the non-parametric covariance function, $k_{np}^{(v)}$, is therefore piecewise smooth over the input space. In particular, we perform a clustering of the data $\bar{\mathbf{X}}$ and approximate

$$g_j^{(v)} = \boldsymbol{\alpha}^{(v)} \cdot \mathbf{e}_j \tag{5}$$

where $\mathbf{e}_j \in \{0,1\}^{P \times 1}$ is an indicator of the cluster that example j belongs to, obtained by clustering the train and test data in the joint feature space, $\boldsymbol{\alpha}^{(i)} \in \Re^{1 \times P}$, P is the number of clusters. The number of parameters to estimate is now $V \cdot (M + P)$. We have experimented with various clustering methods; our approach has proven insensitive to over-clustering as described in our experiments.

We impose an additional constraint such that the resulting covariance $\bar{\mathbf{K}}$ is positive definite, and we incorporate a prior on the non-parametric covariances such that their

elements are non-zero. Learning is then performed by minimizing the negative log posterior

$$\mathcal{L} = \frac{1}{2}\ln|\bar{\mathbf{K}}| + \frac{1}{2}tr(\bar{\mathbf{K}}^{-1}\mathbf{Y}\mathbf{Y}^T) + \lambda\sum_i\sum_j\frac{1}{(\alpha_j^{(i)})^2} \tag{6}$$

with respect to the set of parameters $\bar{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}^{(1)}, \cdots, \boldsymbol{\alpha}^{(V)}]$. Note that the first two terms in Eq. (6) come from the negative log likelihood and the last term represents the prior. The hyper-parameters of the parametric covariances are set by cross-validation.

2.1 Multi-class learning

Our method can be easily extended to the multi-class case by combining binary classifiers with a 1-vs-1 or a 1-vs-all strategy. In principle one can learn a different metric for each classification task, however, the complexity of the problem will become intractable as the number of classes grow. Instead, we exploit the structure of the Gaussian process and develop a fast algorithm that shares the metric across classification tasks. We employ a 1-vs-all strategy and we jointly learn all classifiers by minimizing

$$\mathcal{L}_{multi} = \frac{C}{2}\ln|\bar{\mathbf{K}}| + \sum_{c=1}^{C}\frac{1}{2}tr(\bar{\mathbf{K}}^{-1}\mathbf{Y}^{(\mathbf{c})}\mathbf{Y}^{(c),T}) + \lambda C\sum_i\sum_j\frac{1}{(\alpha_j^{(i)})^2} \tag{7}$$

where C is the number of classes and $\mathbf{Y}^{(c)}$ are the labels for discriminating class c from the rest.

2.2 Inference

The mean prediction is an estimator of the distance to the margin, and thus one can choose the label for each test data point as the one with the largest mean prediction among all the 1-vs-all classifiers

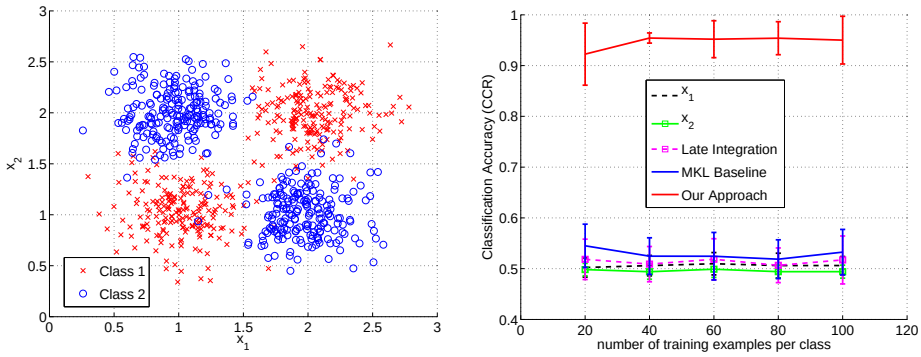
$$\bar{\mathbf{y}}_* = \max_c\{\bar{\mathbf{k}}_*^T\bar{\mathbf{K}}^{-1}\mathbf{Y}^{(c)}\} \tag{8}$$

where $\bar{\mathbf{k}}_*$ is the kernel computed between the training and test data using Eq. (5). Note that comparing the margins makes sense in this setting, since all the classifiers share the same covariance, and only $\mathbf{Y}^{(c)}$ depend on the class labels.

3 Experimental Evaluation

We evaluate our approach on the tasks of audio-visual gesture recognition and object classification. In the audio-visual setting, the different sensory inputs are often corrupted by independent noise processes and can disagree on the class label (e.g., when recognizing head gesture a person can say 'yes' without nodding). Similarly, we explore object recognition using multiple image feature types, where the relevance of a feature type for the classification task can vary locally.

On both tasks we compare our approach to multi- and single-view GP classification baselines. In particular, we compare our approach both to global kernel combination, i.e.,



(a) (b)

Figure 1: **Synthetic example with insufficient views**. (a) The synthetic example consists of two classes and two views samples from four normal distributions in the combined space with std. deviation 0.25 and means $(1,1),(1,2),(2,1),(2,2)$. (b) Classification performance of our approach with $P = 4$ and baseline methods averaged over 50 splits of the data over different training set sizes, error bars indicate ± 1 std. deviation. Unlike the baselines, our approach achieves over 90% classification accuracy despite insufficient input views (see also Figure 5).

our approach with P=1, and to late integration of the single-view GP classifiers, whose output mean prediction is computed as, $\mathbf{y}_* = \sum_v \mathbf{y}_*^{(v)}$, where $\mathbf{y}_*^{(v)}$ is the mean prediction of the GP classifier in v-th view. In our experiments, we report performance using correct classification accuracy computed as the number of examples correctly classified over the total number of examples.

To perform clustering with our approach we use the self tuning spectral clustering algorithm of [14]. For our object classification experiments we set $\lambda = 10^5$. We found the performance of our algorithm to be fairly insensitive to the setting of this parameter. For the other datasets the prior is unused and we set $\lambda = 0$. For both our and the baseline approaches, we use RBF kernels in each view whose kernel widths are either computed with n-fold cross-validation or set proportional to the mean squared distance computed over the train and test samples as described below, and use $\sigma^2 = 0.01$.

3.1 Synthetic example

We first consider the two-view, two-class synthetic example depicted in Figure 1(a). Although classification can be easily performed in the joint space the view projections (x_1, x_2) form a poor representation for classification. Multi-view learning approaches suffer under such projections since the views are largely insufficient for classification—the distributions of each class mostly overlap in each view making it difficult to perform classification from either view alone.

We evaluate our approach on the synthetic example using a dataset consisting of 200 samples drawn from each of the four Gaussian distributions shown in Figure 1(a), each distribution having a std. deviation of 0.25 and means $(1,1),(1,2),(2,1),(2,2)$ respectively. Figure 1(b) displays the performance of our approach with $P = 4$ averaged over 50 splits as a function of the number of labeled samples per class, along with

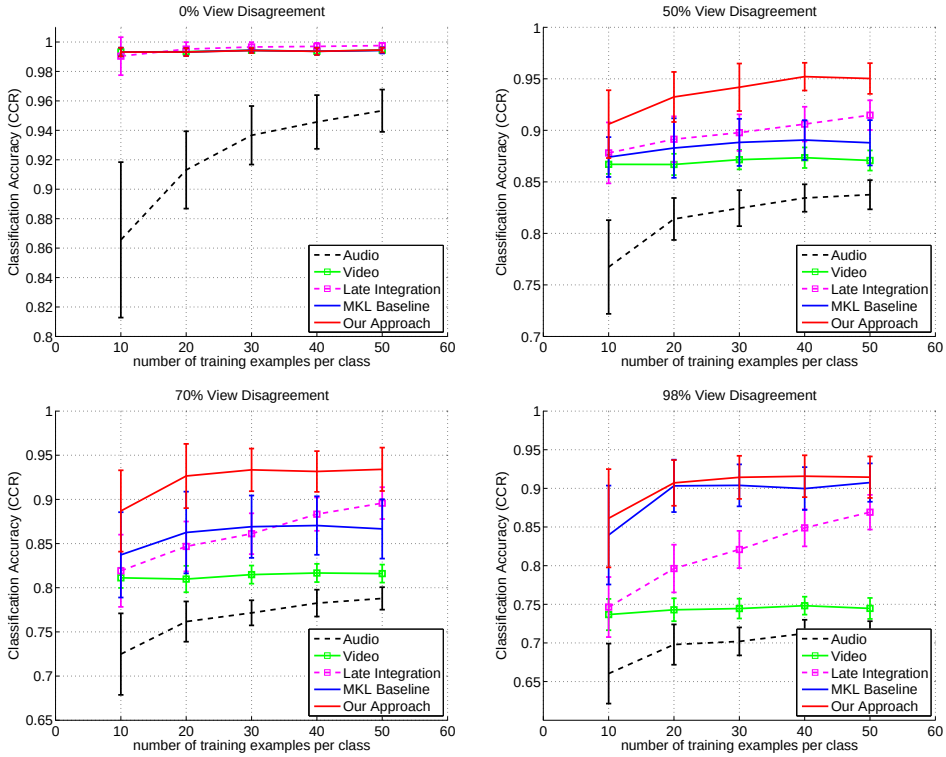
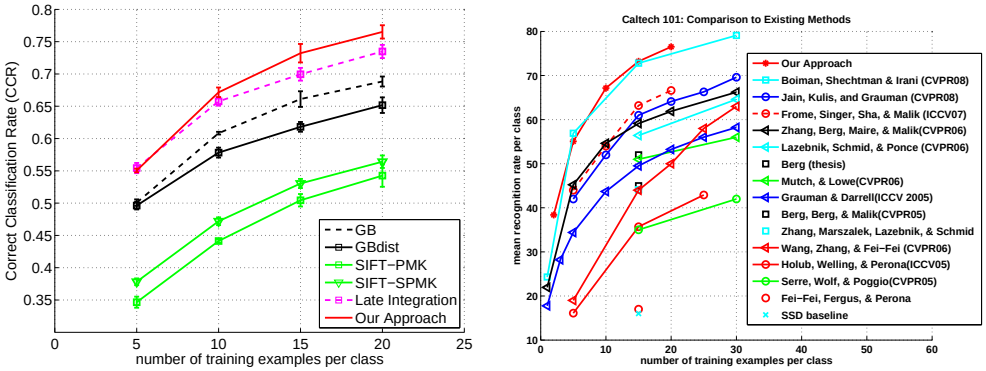


Figure 2: **Audio-visual user agreement experiments**. The performance of our approach is shown along with the baseline approaches averaged over 50 splits as a function of the number of training samples per class, error bars indicate ± 1 std. deviation. Unlike the baseline methods, our approach is able to achieve accurate classification performance despite the per-sample view corruption.

the baseline approaches (see Figure 5 for performance across P). We set the kernel width to half the mean squared distance for all approaches. Unlike the baselines, our approach achieves over 90% average performance across all training set sizes, whereas the baselines do near or slightly better than chance performance. Note that when using a global scaling in each view, it is difficult to recover the original structure apparent in the combined input space. Similarly, the late integration baseline is unable to achieve good performance given weak classifiers in each view. By applying a locally dependent combination of each view, our approach is able to learn an appropriate similarity function that can reliably discriminate each class and achieve good performance despite the view insufficiency.

3.2 Audio-visual user agreement in the present of view disagreement

Next we evaluate our approach on the task of audio-visual user agreement classification from noisy views. Examples of view corruption in this domain include per-sample occlu-



(a) (b)

Figure 3: **Caltech-101 benchmark comparison**. (a) Average performance is shown over 5 splits of the data, error bars indicate ±1 std. deviation. Our approach improves over single-view performance and outperforms the late integration baseline. (b) The performance of our approach is shown along with the most recently reported results the Caltech-101 dataset. In the plot, average performance is displayed.

sion and uni-modal expression, e.g., the user says 'yes' without nodding. We used a user agreement dataset that consisted of 15 subjects interacting with an avatar that answer a set of yes/no questions using head gesture and speech [3]. The head gesture consists of head nods and shakes and the speech data of 'yes' and 'no' utterances, with a total of 718 negative and 750 positive responses. Following Christoudias et al. [3], we simulate view corruption by randomly replacing samples in the visual domain with random head motion segments taken from non-response portions of each user's interaction and in the audio domain with babble noise. The visual features are 3-D FFT-based features computed from the rotational velocities of a 6-D head tracker. The audio features are 9-D observations computed from 13-D Mel Frequency Cepstral Coefficients (MFCCs) sampled at 100Hz over the segmented audio sequence corresponding to each user response using the method of [8]. We corrupt the samples such that for each multi-view sample at least one view is un-occluded. We set the kernel width to half the mean squared distance for all approaches.

Figure 2 displays the performance of our approach on the audio-visual gesture dataset with $P = 3$ over varying amounts of view corruption. In this domain we know $P >= 3$ since there are at least three forms of view corruption, i.e., occlusion in either view or no occlusion. In [3], an ad-hoc filtering method was proposed for solving for view corruption due to per-sample occlusion within a co-training framework. In this work, we learn the view corruption and demonstrate its importance within a supervised learning framework, and report results using multi-view classifiers that combine information from both views. Performance is shown averaged over 50 splits as a function of the number of labeled examples per class. The audio-visual dataset presents a skewed domain in which the visual modality is stronger than the audio modality.

In the absence of per-sample view corruption both our approach and the baselines are able to leverage the strong performance of the visual modality without having a priori knowledge of which is the more reliable view. As the amount of per-sample view

corruption increases the performance of the multi- and single-view baselines degrade significantly, whereas our approach maintains good performance. The corrupted samples in each view are entirely occluded and therefore classification from either view alone is not possible on the occluded samples and the performance of the single-view baselines degrades. Similarly, the late integration baseline degrades with per-sample view corruption given weak classification functions from each view. This is especially the case at the 98% view corruption level, where in contrast to kernel combination, late integration performs poorly.

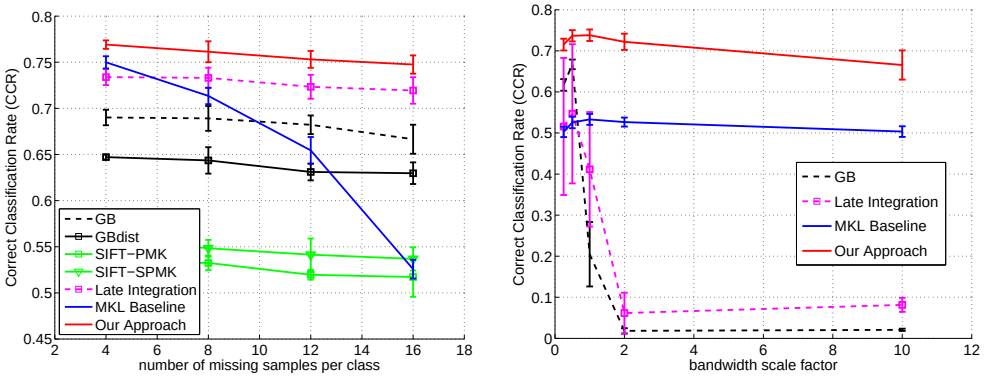
The global kernel combination baseline performs reasonably across the different view corruption levels, achieving the best performance from all the baselines. However, it does significantly worse than our approach in the presence of view corruption. In contrast, using a locally varying kernel we are able to faithfully combine the audio-visual views despite significant per-sample view corruption. At the 98% view corruption level our approach also begins to degrade in performance and the benefit of locally varying kernels reduces with respect to global kernel combination. Importantly our approach is not sensitive to over-clustering, (i.e., $P > 4$) as shown in Figure 5.

3.3 Object recognition

Finally we evaluate our approach on the Caltech-101 benchmark that is comprised of images from 101 object categories [4]. We use four different image features. For the first two feature types we used the geometric blur features described in [15]. The image similarities are computed over geometric blur features sampled at edge points in each image with and without a geometric distortion term. In the figures, we refer to these views as GB and GBdist respectively. The remaining kernels are computed from SIFT features using the PMK [7] and spatial PMK [9] similarity measures, referred to as SIFT-PMK and SIFT-SPMK. In this experiment, we cross-validated the kernel widths of the single-view and late integration baselines using n-fold cross validation with $n = 20$. As shown below, kernel combination is less sensitive to the bandwidth parameter and we approximate the kernel bandwidth using the mean squared distance criteria for both our approach and the multiple kernel learning baseline.

Figure 3(a) displays the performance of our approach with $P = 6$ averaged over 5 splits for varying number of training examples per-class, Figure 5(b) shows that the result is stable across P. The test samples were randomly chosen such that there were a total of 30 examples per-class in each split. Similarly, Figure 3(b) plots the performance of our approach compared to the most recently reported results on this dataset. Our approach obtains state-of-the-art performance. It improves over the single-view GP baselines and outperforms late integration, however, as discussed in the following subsection, it does not see a benefit compared to global kernel combination. Note that for this task the late integration baseline can be seen as a variant of the approach of [2], with a Gaussian Process used in place of the naive Bayes nearest-neighbor classifier in each view. Moreover, the method in [2] uses additional feature types including shape-context and self-similarity, and we anticipate increased performance with our approach provided more feature types.

An interesting property of our approach is its ability to cope with missing data. Missing data is simulated by removing at most one view per sample in the training



(a) (b)

Figure 4: **Caltech-101 with missing data.** (a) Unlike conventional kernel combination our approach can take advantage of partially observed multi-view samples. (b) Late integration is sensitive to bandwidth selection. The performance of our approach is relatively un-affected by the corrupted view and maintains stable performance as its bandwidth is drastically varied. Performance is shown averaged over 5 splits of the data with $N = 20$ and for (b) with 16 missing samples per class, error bars indicate ± 1 std. deviation. The kernel bandwidth is displayed as a multiple of the mean distance over the train and test samples.

set. For our approach, we use a per-sample $\{0, \alpha^j_{(i)}\}$ weighting according to the missing data. Under this setting, our approach can be seen as performing a variant of mean-imputation where the missing kernel value is computed from the other views as opposed to the samples within the same view [11]. In Figure 4(a) we report results fixing $\alpha^v_i = 1$ for the observed input streams and normalizing the weights of each sample so that their squares sum to one.

Figure 4(a) displays average performance across 5 splits of the data over varying amounts of missing data. As conventional kernel combination assumes fully observed views, it can only be trained on the fully observed data and is unable to take advantage of the partially observed examples; it exhibits poor performance compared to our approach and the other methods that are able to learn from both the fully and partially observed multi-view data samples. Our approach inherits the favorable performance of kernel combination while having the ability to utilize partially observed data samples. As in the fully observed case, it improves over single-view performance and outperforms late integration despite the missing data.

Experiments on the audio-visual and synthetic datasets demonstrated that unlike kernel combination, the late integration baseline suffers in the presence of weak per-view classifiers. A similar effect is seen in Figure 4(b) on the Caltech-101 dataset where we plot the performance of our approach and the late integration baseline as a function of the GB kernel bandwidth. The results are averaged over 5 splits of the data with $N = 20$ and with 4 samples removed per class and view, and the performance of the global kernel combination baseline and that of the affected view is also shown. The bandwidths of the other views are held constant. Note that in Figure 4, both the late integration baseline and our approach assume an equal weighting over the views. The performance of the

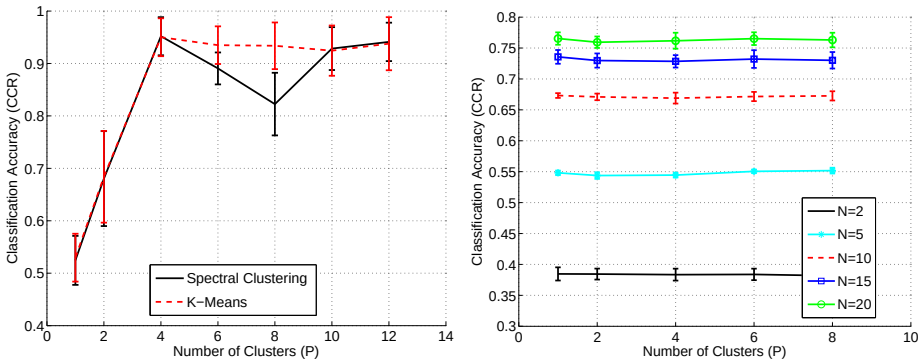


Figure 5: **Influence of the number of clusters**. Average performance is shown for each dataset, error bars indicate ± 1 std. deviation. (left) Influence on synthetic dataset. Performance is averaged over 50 splits with $N = 60$ and the rest test. Performance with spectral and k-means clustering is shown. A significant increase in performance is seen from $P = 1$ to $P = 4$ clusters and remains constant for larger cluster numbers. The decrease in performance at $P = 8$ with spectral clustering is the result of a poor clustering solution as seen by the steady performance found with k-means. (right) Influence on Caltech-101. Performance is averaged over 5 splits of the data. The number of clusters has little influence on Caltech-101, see text for details.

late integration baseline is sensitive to the kernel bandwidth parameter of each view and reflects the performance of the single view classifier. In contrast, our approach is relatively un-affected by the corrupted view and maintains stable performance across a wide range of bandwidth scale factors.

3.4 Influence of the Number of Clusters

As shown in the above experiments a local weighting of the views can lead to a large increase in performance when provided with insufficient or noisy views, or when coping with missing data. Figure 5 displays the performance of our approach with respect to number of clusters P on the synthetic and Caltech-101 datasets. For the synthetic dataset a large increase in performance is seen between $P = 1$ and $P = 4$, and the performance remains relatively constant for larger P values. A decrease in performance is seen with spectral clustering around $P = 8$ that is due to a poor clustering of the space. For comparison, performance obtained with k-means clustering is also shown and this effect is removed. The relatively stable performance for large P values suggests that P need only be roughly estimated with our approach and an over-clustering of the data space does not adversely affect our algorithm. The results on Caltech-101 show no change with varying P. We believe that this is due to the sparse nature of the Caltech-101 dataset; provided more training samples from each class or unlabeled data, we anticipate that a locally varying weighting of the space would also prove advantageous to a global weighting for the object classification task.

4 Conclusions

We have presented a Bayesian approach to multiple kernel learning where the weights can vary locally. Our approach learns the kernel matrix of a Gaussian Process using a product of a parametric covariance representing feature similarities and a rank-constrained non-parametric covariance that represents similarities in each view. We have proposed a simple optimization criteria that exploits the properties of the covariance to efficiently learn multi-class problems, and demonstrated our approach in the context of audio-visual user agreement recognition in the presence of complex noise processes, and object recognition from multiple image feature types. We plan to investigate soft clustering as well as the application of our approach to other domains, e.g., pose estimation.

References

[1] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, 2004.

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

[3] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *UAI*, 2008.

[4] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 2006.

[5] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.

[6] M. Gonen and E. Alpaydin. Localized multiple kernel learning. In *ICML*, 2008.

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

[8] A. Halberstadt. *Heterogeneous acoustic measurements and multiple classifiers for speech recognition*. PhD thesis, MIT, 1998.

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

[10] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In *CVPR*, 2007.

[11] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, 2005.

[12] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *JMLR*, 2006.

[13] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[14] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[15] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest-neighbor classification for visual category recognition. In *CVPR*, 2006.

[16] A. Zhen and C. S. Ong. Multiclass multiple kernel learning. In *ICML*, 2007.