

---

# Bayesian Localized Multiple Kernel Learning

---

C. Mario Christoudias\*, Raquel Urtasun\*\* and Trevor Darrell\*

\*UC Berkeley EECS & ICSI    \*\*Toyota Technological Institute at Chicago

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, 11, 7] 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. [4] 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. [9] learn an ensemble of SVM classifiers defined on a per-example basis for coping with local variability. Similarly, Gonen and Alpaydin [5] 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 work we present a Bayesian approach to multiple kernel learning that can learn a local weighting over each view of the input space. Unlike [4], in our framework learning can be done exactly for large datasets. We exploit the properties of the covariance matrix and utilize a simple optimization criteria [7], when compared to SVM-based approaches [12, 14], that allow us to efficiently learn multi-class problems.

Let  $\mathbf{X}_i = [\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(V)}]$  be a multi-view observation with  $V$  views, and let  $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)} \dots \mathbf{x}_N^{(v)}]^T$  be a set of  $N$  observations of view  $v$ . Let  $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_N]^T$  be the set of labels, and let  $\mathbf{f} = [\mathbf{f}_1 \dots \mathbf{f}_N]^T$  be a set of latent functions. We assume a Gaussian Process (GP) prior over the latent functions,  $p(\mathbf{f}|\bar{\mathbf{X}}) = \mathcal{N}(0, \bar{\mathbf{K}})$ , where  $\bar{\mathbf{X}} = [\mathbf{X}^{(1)} \dots \mathbf{X}^{(V)}]$  is the set of all observations, and  $\mathbf{f}$  is the set of latent functions. We use a Gaussian noise model,  $p(\mathbf{Y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$ , although others such as the probit can be used, and construct our covariance as a linear combination of covariance matrices  $\bar{\mathbf{K}} = \sum_v \mathbf{K}^{(v)} + \sigma^2\mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix.

With our approach, the covariance for each view is defined 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

$$K_{ij}^{(v)} = k_{np}^{(v)}(i, j) \cdot k_p^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)}) \quad (1)$$

Note the non-parametric kernel performs a per-sample weighting of each feature channel.

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

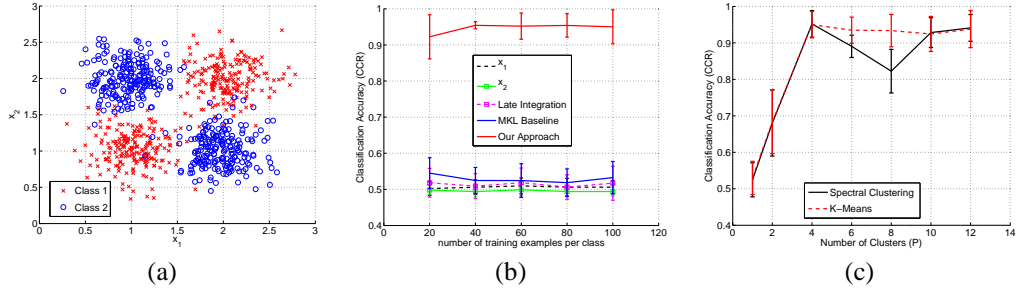


Figure 1: **Synthetic example with insufficient views.** (a) Synthetic example: two classes and views sampled from four normal distributions in the combined space with std. dev. 0.25 and means  $(1, 1), (1, 2), (2, 1), (2, 2)$ . (b) Classification performance of our approach with  $P = 4$  and baseline methods averaged over 50 splits of the data over different training set sizes, error bars indicate  $\pm 1$  std. deviation. (c) Influence of the number of clusters  $P$  (see text for discussion).

with large datasets. To make learning tractable, we assume low-rank approximations to the non-parametric covariances such that

$$\mathbf{K}_{np}^{(v)} = (\mathbf{g}^{(v)})^T \mathbf{g}^{(v)} \quad (2)$$

where  $\mathbf{g} = [\mathbf{g}_1, \dots, \mathbf{g}_N]^T \in \mathbb{R}^{m \times N}$ , 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,  $g_j^{(v)} = \boldsymbol{\alpha}^{(v)} \cdot \mathbf{e}_j$ , 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,  $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^{1 \times P}$ ,  $P$  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.

Learning is then performed by minimizing the negative log posterior where in the case of multiple classes we employ a 1-vs-all strategy and we jointly learn all classifiers by minimizing

$$\mathcal{L}_{multi} = \frac{C}{2} \ln |\bar{\mathbf{K}}| + \sum_{c=1}^C \frac{1}{2} \text{tr}(\bar{\mathbf{K}}^{-1} \mathbf{Y}^{(c)} \mathbf{Y}^{(c),T}) + \lambda C \sum_i \sum_j \frac{1}{(\alpha_j^{(i)})^2} \quad (3)$$

with respect to the set of parameters  $\bar{\boldsymbol{\alpha}} = [\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(V)}]$ , where  $C$  is the number of classes and  $\mathbf{Y}^{(c)}$  are the labels for discriminating class  $c$  from the rest. Note that the first two terms in Eq. (3) come from the negative log likelihood and the last term is a prior that favors non-zero solutions. 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, inspired by [7] we exploit the structure of the Gaussian process and employ a fast algorithm that shares the metric across classification tasks. The hyper-parameters of the parametric covariances are set by cross-validation.

For 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. 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.

## Experiments

We evaluate our approach on the tasks of audio-visual gesture recognition and object classification. Our approach is compared to multi- and single-view GP classification baselines. In particular, we compare our approach both to global kernel combination 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

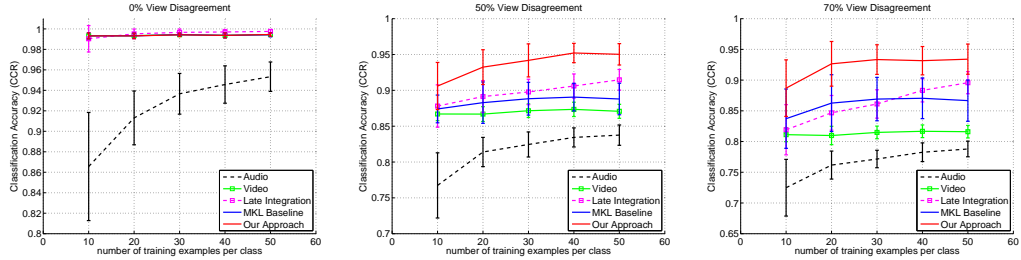


Figure 2: **Audio-visual user agreement experiments.** The performance of our approach is shown along with the baseline approaches averaged over 50 splits as a function of the number of training samples per class, error bars indicate  $\pm 1$  std. deviation. Unlike the baseline methods, our approach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

mean prediction of the GP classifier in the  $v$ -th view. For our object classification experiments we set  $\lambda = 10^5$ , and 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  $\sigma^2 = 0.01$ .

First we 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). 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 the baseline 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. Figure 1(c) displays the performance of our approach with respect to number of clusters  $P$  found with  $k$ -means and spectral clustering. A decrease in performance is seen with spectral clustering around  $P = 8$  that is due to a poor clustering of the space and that is avoided with  $k$ -means. Importantly our approach is not sensitive to over-clustering, (i.e.,  $P > 4$ ).

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 occlusion 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 [2]. 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, and corrupt the samples such that for each multi-view sample at least one view is un-occluded. Figure 2 displays the performance of our approach on the audio-visual gesture dataset with  $P = 3$  over varying amounts of view corruption (0%, 50%, and 70%) averaged over 50 splits of the data. The global kernel combination baseline performs reasonably across the different view corruption levels, however, does significantly worse than our approach in the presence of view corruption. Similarly, the late integration baseline degrades with per-sample view corruption given weak classification functions from each view. In contrast, using a locally varying kernel we are able to faithfully combine the audio-visual views despite significant per-sample view corruption and our approach maintains good performance.

Finally, we evaluate our approach on the Caltech-101 benchmark that is comprised of images from 101 object categories [3], with four different image feature kernel types: the geometric blur kernels described in [13] with and without a geometric distortion term, and PMK [6] and spatial PMK [8] kernel measures computed over SIFT features. Figure 3(a) plots the performance of our approach with  $P = 6$  compared to the most recently reported results on this dataset and our approach obtains state-of-the-art performance. Figure 3(b) displays the performance for varying  $P$ . The results show no change with varying  $P$  and we obtain similar results to those reported in [7] whose approach can be seen as special case of our model with  $P = 1$ . 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

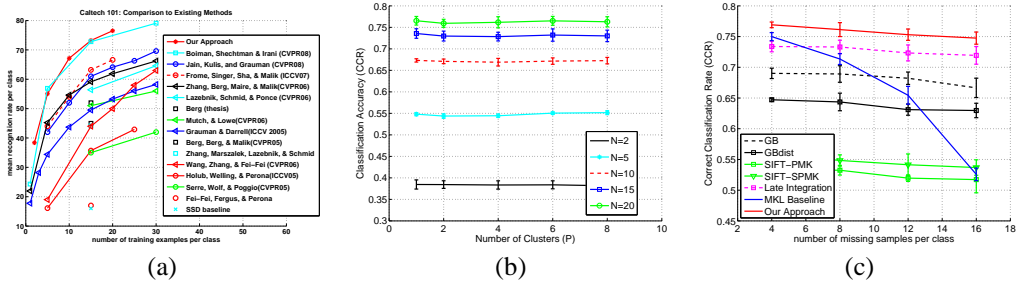


Figure 3: **Caltech-101**. (a) Comparison to state-of-the-art. (b) The number of clusters has little influence on Caltech-101, see text for details. (c) Unlike conventional kernel combination our approach can take advantage of partially observed multi-view samples. Performance is shown averaged over 5 splits of the data with  $N = 20$ , error bars indicate  $\pm 1$  std. deviation.

anticipate that a locally varying weighting of the space would also prove advantageous to a global weighting for the object classification task.

An interesting property of our approach is its ability to cope with missing data. We simulated missing data on Caltech-101 by removing at most one view per sample in the training set. For our approach, we use a per-sample  $\{0, \alpha_{(i)}^j\}$  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 [10]. In Figure 3(c) we report results fixing  $\alpha_i^v = 1$  for the observed input streams and normalizing the weights of each sample so that their squares sum to one. Unlike global kernel combination, our approach can benefit from both fully and partially observed examples and outperforms the multi- and single-view baseline approaches.

## Conclusion

We have presented a Bayesian approach to multiple kernel learning where the weights can vary locally that learns the kernel matrix of a GP using a product of a parametric and non-parametric covariance. We proposed a simple optimization criteria to efficiently learn multi-class problems, and demonstrated our approach on the tasks of audio-visual user agreement and object recognition. We plan to investigate soft clustering as well as the application of our approach to other domains.

## References

- [1] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, 2004.
- [2] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *UAI*, 2008.
- [3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 2006.
- [4] 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.
- [5] M. Gonen and E. Alpaydin. Localized multiple kernel learning. In *ICML*, 2008.
- [6] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *ICCV*, 2005.
- [7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *IJCV*, 2009.
- [8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [9] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In *CVPR*, 2007.
- [10] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, 2005.
- [11] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *JMLR*, 2006.
- [12] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.
- [13] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest-neighbor classification for visual category recognition. In *CVPR*, 2006.
- [14] A. Zhen and C. S. Ong. Multiclass multiple kernel learning. In *ICML*, 2007.