Co-training with Noisy Perceptual Observations

C. Mario Christoudias*, Raquel Urtasun*, Ashish Kapoor‡ and Trevor Darrell*
*UC Berkeley EECS & ICSI ‡ Microsoft Research

Abstract

Many perception problems involve datasets that are naturally comprised of multiple streams or modalities for which supervised training data is only sparsely available. In cases where there is a degree of conditional independence between such views, a class of semi-supervised learning techniques that are based on maximizing view agreement over unlabeled data has been proven successful in a wide range of machine learning domains. However, these 'co-training' or 'multi-view' learning methods have had relatively limited application in vision, due in part to the assumption of constant per-channel noise models. In this paper we propose a probabilistic heteroscedastic approach to co-training that simultaneously discovers the amount of noise on a per-sample basis, while solving the classification task. This results in high performance in the presence of occlusion or other complex observation noise processes. We demonstrate our approach in two domains, multi-view object recognition from low-fidelity sensor networks and audio-visual classification.

1. Introduction

Many perception problems inherently involve multiple 'views', where a view is broadly defined to mean any sensor stream of a scene or event. The different views can be formed from the same sensor type (e.g., multiple cameras overlooking a common scene), come from different modalities (e.g., audio-visual events, or joint observations from visual and infra-red cameras), and/or be defined by textual or other metadata (image captions, observation parameters).

It is also typical that labeled data is expensive to obtain while unlabeled data may be available in relatively large quantities; researchers have thus investigated semi-supervised and transductive learning techniques, which attempt to exploit the statistics of unlabeled data to improve performance. Semi-supervised learning in the presence of multiple views has received considerable recent interest in the machine learning community. A class of techniques based on the classic 'co-training' method [3], and the more general notion of maximizing agreement on unlabeled data while training classifiers to be optimally predictive of labeled data, has been successful in a range of tasks [7, 5, 11, 17].

With a few notable exceptions [5, 17, 11], however, co-training methods have had only limited success on visual tasks. We argue here that this is due in part to restrictive assumptions inherent in existing multi-view learning techniques. Classically, co-training assumes 'view sufficiency', which simply speaking means that either view is sufficient to predict the class label, and implies that whenever observations co-occur across views they must have the same label. In the presence of complex noise (e.g., occlusion), this assumption can be violated quite dramatically. A variety of approaches have been proposed to deal with simple forms of view insufficiency [17, 13, 18]. More complex forms of noise such as per-sample occlusion, however, have received less attention. We develop here a co-training algorithm that is robust to complex sample corruption and *view disagreement*, i.e., when the samples of each view do not belong to the same class due to occlusion or other view corruption.

Christoudias et al. [6] have reported a filtering approach to handle view disagreement, and develop a model suitable for the case where the view corruption is due to a background class. However, occlusion can occur with or without a dominant background, and as shown in our experiments below, their method performs poorly in the latter case. Yu et al. [18] recently presented a Bayesian approach to co-training, with a view-dependent noise term. We show here that the presence of complex noise can be tackled in a general and principled way by extending Bayesian co-training to incorporate sample-dependent noise. Our *heteroscedastic* Bayesian co-training algorithm simultaneously discovers the amount of noise while solving the classification task. Unlike previous multi-view learning approaches, our approach can cope with a variety of complex noises and per-sample occlusions that are common to many multi-sensory vision problems.

In this paper we demonstrate our approach on two

different multi-view perceptual learning tasks. The first task is multi-view object classification from multiple cameras on a low-fidelity network, where the object is often occluded in one or more views (e.g., as a result of network asynchrony or the presence of other objects). For a two-view multi-class object recognition problem we show that our approach is able to reliably perform recognition even in the presence of large amounts of view disagreement and partial occlusion. We also consider the task of audio-visual user agreement recognition from head gesture and speech, where view disagreement can be caused by view occlusions and/or uni-modal expression, and show that unlike existing approaches our method is able to successfully cope with large amounts of complex view corruption.

2. Background: Multi-view learning

Multi-view learning approaches [1, 3, 7, 10, 14, 16, 18] form a class of semi-supervised learning techniques that use multiple views to effectively learn from partially labeled data. Blum and Mitchell [3] introduced co-training which bootstraps a set of classifiers from high confidence labels. Nigam and Ghani [14] presented a co-EM algorithm that uses soft label assignment with EM to bootstrap classifiers from multiple views. Collins and Singer [7] proposed a co-boost approach that optimizes an objective that explicitly maximizes the agreement between each classifier, while Sindhwani et. al. [16] defined a co-regularization method that learns a multi-view classifier from partially labeled data using a view consensus-based regularization term.

More formally, let $\mathbf{X}_i = [\mathbf{x}_i^1, \cdots \mathbf{x}_i^V]$ be a multi-view observation, and let $\mathbf{X}^i = [\mathbf{x}_1^i, \cdots \mathbf{x}_N^i]^T$ be a set of observations from a single view i. Multi-view learning approaches mutually train a set of classifiers, one per view, by maximizing their agreement on the unlabeled data, e.g., using the L_2 norm,

$$\min \sum_{\mathbf{x}_k \in U} \sum_{i \neq j} \| f_i(\mathbf{x}_k^i) - f_j(\mathbf{x}_k^j) \|_2^2 \qquad (1)$$

where f_i^k is the prediction from the classifier of view k for the unlabeled data point $\mathbf{x}_k$. Minimizing Eq. (1) is beneficial when the different views are conditionally independent given the class label and sufficient for classification, i.e., classification can be performed from either view alone.

Yu et al. [18] proposed a probabilistic approach to co-training, called *Bayesian Co-training*, that combines multiple views in a principled way. In particular, they introduced a latent variable $\mathbf{f}_j$ for each view and a consensus latent variable, $\mathbf{f}_c$, that models the agreement between the different classifiers. They assumed a Gaussian process prior [15] on the latent variables

$$\mathbf{f}_j \sim \mathcal{N}(0, \mathbf{K}_j) \ , \qquad (2)$$

where $\mathbf{f}_j = [f_j(\mathbf{x}_1^j), \cdots, f_j(\mathbf{x}_N^j)]^T$ is the set of latent variables for all observations of a single view j. For simplicity, in the discussion that follows we assume that the classification task is binary, $\mathbf{y} \in \{-1, 1\}$; this implies that $f_j(\mathbf{x}_i^j) \in \Re$.

Assuming conditional independence between the labels $\mathbf{y}$ and the latent variables in each view, $\mathbf{f}_i$, the joint probability can be factorized in the following form

$$p(\mathbf{y}, \mathbf{f}_c, \mathbf{f}_1, \cdots, \mathbf{f}_V) = \frac{1}{Z} \prod_{i=1}^{n} \psi(y_i, f_c(\mathbf{X}_i)) \prod_{j=1}^{m} \psi(\mathbf{f}_j) \psi(\mathbf{f}_j, \mathbf{f}_c) \qquad (3)$$

where Z is a normalization constant, V is the number of views, N the number of data points, and $\mathbf{X}_i = [\mathbf{x}_i^1, \cdots \mathbf{x}_i^V]$ is the i-th multi-view observation. The potential $\psi(\mathbf{f}_j) \sim \mathcal{N}(0, \mathbf{K}_j)$ arises due to the GP prior in Eq. (2) and specifies within-view constraints for the latent variables. Intuitively, this enforces that the latent variables in a particular view should co-vary according to the similarities specified by the kernel matrix $\mathbf{K}_j$.

The potential $\psi(y_i, f_c(\mathbf{X}_i))$ defines the dependence of the consensus variable and the final output. As with other GP models this can either be a Gaussian noise model or a classification likelihood defined via a link function (*e.g.*, probit or logistic function). For computational efficiency a Gaussian noise model was used in [18].

Finally, the potential $\psi(\mathbf{f}_j, \mathbf{f}_c)$ defines the compatibility between the j-th view and the consensus function and can be written as: $\psi(\mathbf{f}_j, \mathbf{f}_c) = \exp(-\frac{||\mathbf{f}_j - \mathbf{f}_c||^2}{2\sigma_j^2})$. The parameters σ_j act as reliability indicators and control the strength of interaction between the j-th view and the consensus latent variable. A small value of σ_j imposes a strong influence of the view on the final output, whereas a very large value allows the model to discount observations from that view.

It has been shown that Bayesian co-training improves performance with respect to other state-of-the-art co-training approaches [18]. However, it can only handle *per-view* noise, i.e., each sample of a given view is assumed to be corrupted by the same amount of noise. As a consequence, as shown below, its performance degrades significantly when dealing with complex non-stationary noise processes.

Several approaches have been proposed in the multi-view learning literature to cope with view insufficiency or noise [7, 16, 17, 18, 13]. However, these approaches have mostly focused on relatively simple noise models, e.g., that assume per view noise corruption that

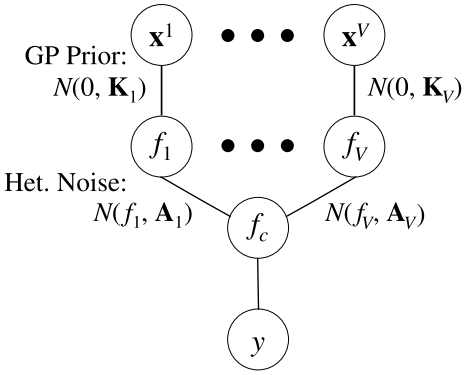


Figure 1. Graphical model of *Heteroscedastic Bayesian Co-training* (our approach). Our multi-view learning approach extends Bayesian co-training to incorporate sample-dependent noise modeled by the per view noise covariance matrices $\mathbf{A}_j$. Our method simultaneously discovers the amount of noise in each view while solving the classification task.

is uniform across samples [17, 18, 13], and have difficulty dealing with more complex view corruption such as that caused by per sample view occlusion. Christoudias et al. [6] addressed the problem of *view disagreement*, i.e., when the samples from each view do not always belong to the same class. They proposed a two step process, where first the samples with view disagreement are identified and filtered using an information theoretic criterion, and traditional co-training is applied to the remaining samples. Their method, however, is developed for the case where view corruption is due to a background class, and their algorithm suffers in the presence of more general noise, as we show below.

3. Heteroscedastic Bayesian Co-training

To deal with noisy data, in this paper we extend Bayesian co-training to the *heteroscedastic* case, where each observation can be corrupted by a different noise level. In particular, we assume that the latent functions can be corrupted with arbitrary Gaussian noise

$$\psi(\mathbf{f}_j, \mathbf{f}_c) = \mathcal{N}(\mathbf{f}_j, \mathbf{A}_j) \tag{4}$$

with $\mathbf{A}_j$ being the noise covariance matrix. The only restriction on $\mathbf{A}_j$ in our model is that it is positive semi-definite so that its inverse is well defined. Fig. 1 depicts the undirected graphical model of our *Heteroscedastic Bayesian Co-training* approach.

Integrating out the latent functions $\mathbf{f}_j$ in (3) results in a GP prior over the consensus function such that

$$p(\mathbf{f}_c) = \mathcal{N}(0, \mathbf{K}_c) \tag{5}$$

with covariance

$$\mathbf{K}_c = \left[\sum_j (\mathbf{K}_j + \mathbf{A}_j)^{-1}\right]^{-1} . \tag{6}$$

This implies that given a set of multi-view observations, the *heteroscedastic co-training kernel* $\mathbf{K}_c$ can be directly used for Gaussian process classification or regression. Unlike other co-training algorithms that require alternating optimizations, Bayesian co-training and our heteroscedastic extension can jointly optimize all the views. Furthermore, our approach naturally incorporates semi-supervised and transductive settings as the kernel $\mathbf{K}_c$ depends on both the labeled and unlabeled data.

For $\mathbf{K}_j$ we use an RBF kernel with parameter θ, i.e., $\exp(-\theta||\mathbf{x}-\mathbf{x}'||^2)$. Learning the heteroscedastic model consists of solving for the kernel hyper-parameters of $\mathbf{K}_j$ (i.e., RBF width) and the noise covariances $\mathbf{A}_j$ defined in each view. With no further assumptions the number of parameters to estimate is prohibitively large, $V(\frac{N(N-1)}{2}+1)$, with V being the number of views, and N the number of samples.

Additional assumptions on the type of noise can be imposed to reduce the number of parameters, facilitating learning and inference. When assuming *i.i.d. noise*, the covariance is restricted to be diagonal

$$\mathbf{A}_j = diag(\sigma^2_{1,j}, \cdots, \sigma^2_{N,j}) \tag{7}$$

where $\sigma^2_{i,j}$ is the estimate of the noise corrupting sample i in view j. The resulting i.i.d. noise model has $V(N+1)$ parameters, which is still too large to be manageable in practice.

To further reduce the computational complexity we assume that the noise is *quantized-i.i.d.*, i.e., there are only a finite number of noise levels that can corrupt a sample. The noise covariance for each view j can then be expressed in terms of an indicator matrix, $\mathbf{E}^{(j)}$, and a vector of P noise variances, $\phi_j = [\sigma^2_{1,j}, \cdots \sigma^2_{P,j}]^T \in \Re^{P\times 1}$ as

$$\mathbf{A}_j = diag(\mathbf{E}^{(j)} \cdot \phi_j) . \tag{8}$$

The indicator matrices, $\mathbf{E}^{(j)} = [\mathbf{e}^{(j)}_1, \cdots, \mathbf{e}^{(j)}_N]^T$ are matrices such that each row, $\mathbf{e}^{(j)}_i \in \{0,1\}^{P\times 1}$, is an indicator vector where one element has value one, indicating the noise level from which that sample was corrupted, and zero elsewhere. Note that if $P = 1$, we recover Bayesian co-training [18], and if $P = N$, and $E^{(j)}$ is full rank, we recover the full i.i.d. heteroscedastic case.

Learning our model consists of estimating the indicator matrices $\mathbf{E}^{(j)}$, the noise values ϕ_j for each view, and the kernel hyper-parameters θ_j. The number of parameters to estimate is now $V(K+1)$, with

$k \ll N$. We introduce a two-step process for learning the parameters. First, we learn the kernel hyperparameters $\Theta = \{\theta_1, \cdots, \theta_V\}$ and the noise values $\Phi = \{\phi_1, \cdots, \phi_V\}$ for each view using n-fold cross-validation, which as shown below, outperformed maximum likelihood. Note that we do not need to estimate the indicator matrices for the labeled data since they are known.

The indicator matrices for the unlabeled data are then estimated using Nearest Neighbor (NN) classification in each view independently (other classifiers are possible, e.g. GP classification). We compute the co-training covariance $\mathbf{K}_c$, which is non-stationary, using the labeled and unlabeled data.

Finally the labels for the unlabeled data are estimated using mean prediction

$$\bar{\mathbf{y}}_* = \mathbf{k}_c(\mathbf{X}_*)^T(\hat{\mathbf{K}}_c + \sigma^2\mathbf{I}_N)^{-1}\mathbf{y} \quad (9)$$

where $\mathbf{X}_*$ is a multi-view test sample, $\mathbf{k}_c(\mathbf{X}_*)$ is the kernel computed between the labeled and unlabeled data, and $\hat{\mathbf{K}}_c$ are the rows and columns of $\mathbf{K}_c$ corresponding to the training samples. The estimation of $\hat{\mathbf{K}}_c$ involves the computation of kernels formed using training and test data, since the kernel involves computing inverses. Here, we have assumed that the mapping between $\mathbf{f}_c$ and $\mathbf{y}$ is Gaussian with noise variance σ^2. In practice, a small value of σ is used, giving robustness to the inversion of $\mathbf{K}_c$.

Finally our method is easily extended to the multi-class case by combining binary classifiers with a 1 vs. 1 or 1 vs. all approach. In particular, in our experiments below we use 1 vs. all classifiers.

4. Experimental Evaluation

We demonstrate our approach on two different multi-view perceptual learning tasks: multi-view object classification and audio-visual gesture recognition.

We first consider the problem of multi-view object classification from cameras that lie on a low-fidelity sensor network, where one or more views are often corrupted by network asynchrony and/or occlusion. For this setting, we collected a database of 10 objects imaged from two camera sensor "motes" [4] placed at roughly 50 degrees apart. The objects were rotated from 0 to 350 degrees at 10 degree increments to give 36 views for each instance from each camera. We use a bag-of-words representation for classification, where SIFT features are extracted on a grid over a bounding box region surrounding the object in each image. These features are then quantized using a hierarchical feature vocabulary computed over the features of all the images across views and similarity between images is measured using the pyramid match similarity [8].

In this setting, we consider two forms of sample corruption, partial and complete view occlusion. In the latter case, we randomly replaced samples in each view with background images captured from each camera that do not contain any object. To simulate partial occlusions, we randomly selected a quadrant (i.e., 20% of the image) of each image and discarded the features from that quadrant.

For the second task, we evaluate our approach on the problem of audio-visual user agreement recognition from speech and head gesture. In this setting, sample corruption can occur in the form of view occlusion and uni-modal expression (e.g., a subject can say 'yes' without gesturing). We use the database of [5], that is comprised of 15 subjects interacting with an avatar in a conversational dialog task. The database contains segments of each subject answering a set of yes/no questions using both head gesture (i.e., head nod or shake) and speech (i.e., a 'yes' or 'no' utterance).

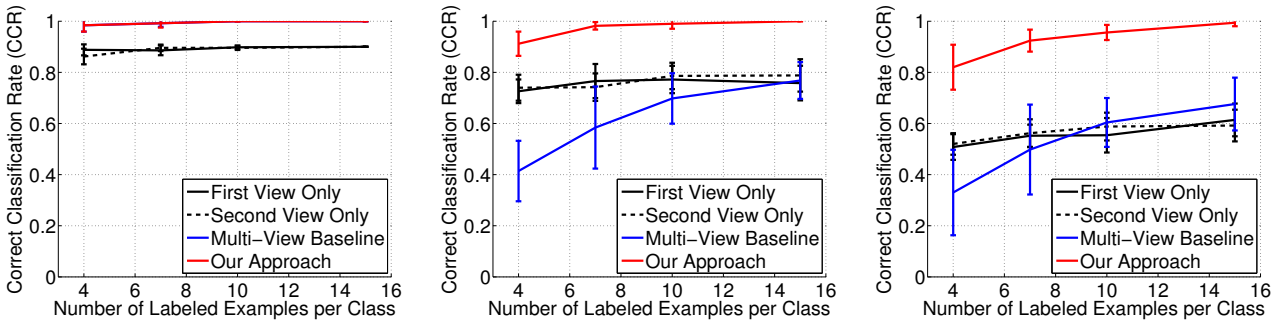
Following Christoudias et al. [6], we simulate view corruption by randomly replacing samples in the visual domain with random head motion segments taken from non-response portions of each user's interaction and in the audio domain with babble noise. The visual features are 3-D FFT-based features computed from the rotational velocities of a 6-D head tracker [12]. 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 [9]. For both the multi-view image and audio-visual databases we corrupt the samples such that for each corrupted multi-view sample at least one view is un-occluded.

We compare our approach against Bayesian co-training [18] and the approach of Christoudias et. al [6]. We also compare against single view performance using GP regression-based classifiers in each view and multi-view GP kernel combination. We evaluate each approach under the Correct Classification Rate (CCR) evaluation metric defined as

$$\text{CCR} = \frac{\#\text{ samples correctly classified}}{\#\text{ of samples}} \quad (10)$$

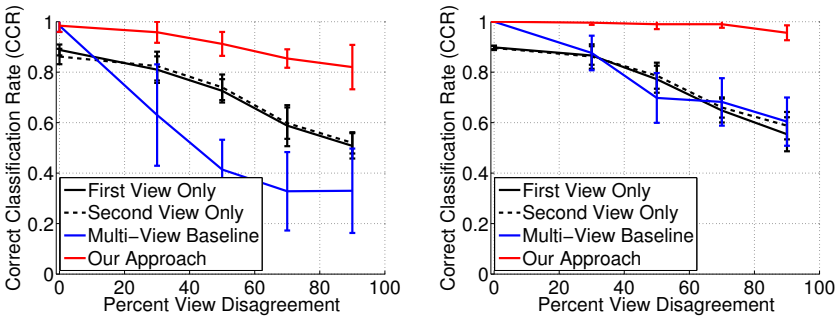
For learning the parameters to our model we used n-fold cross validation from the labeled examples, with $n = 2$ held-out examples per class.

In what follows we first demonstrate our approach for the case of binary view corruption under each of the above databases, where each view sample is either completely occluded or un-occluded. We then present results on a more general noise setting that also contains partial view occlusions.



(0% VD) (50% VD) (90% VD)

Figure 2. **Object recognition from multiple camera sensors with varying training set sizes**: Classification accuracy for a 10-class problem as a function of the number of training samples for 0%, 50% and 90% view disagreement. Performance is shown averaged over 10 splits, the error bars indicate ± 1 std. deviation. Our approach significantly outperforms the single-view and multi-view [18] baseline methods in the presence of view disagreement. Note for 0% view disagreement our approach and multi-view baseline perform the same and their curves overlay one-another.



($M = 4$) ($M = 10$)

Figure 3. **Object recognition from multiple camera sensors with varying levels of view disagreement**: Classification accuracy as a function of the level of view disagreement. Performance is shown averaged over 10 splits, error bars indicate ± 1 std. deviation. Our approach is able to achieve good performance across a full range of view disagreement levels, even when presented with a small number of labeled training samples ($M = 4$). Multi-view baseline performance is using the approach of [18].

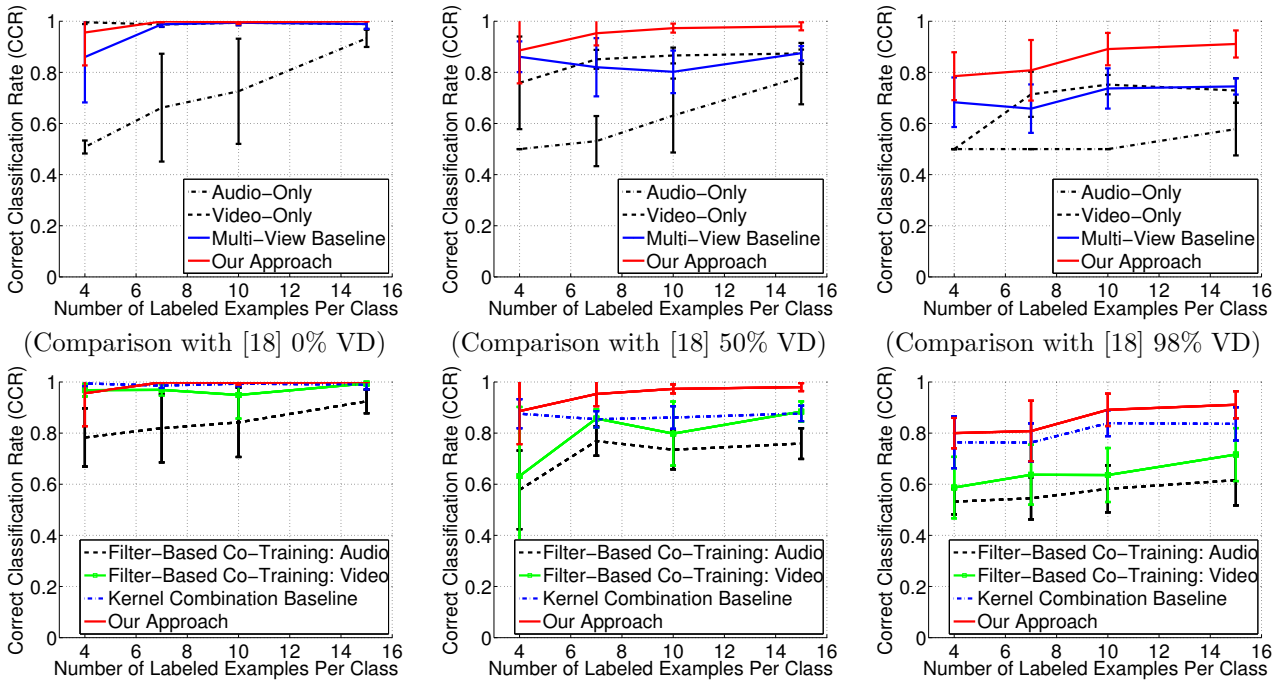
4.1. View disagreement

For the instance-level, multi-view object classification experiment we split the data into a labeled and unlabeled set by retaining M samples per object instance to comprise the training set and 5 samples per instance to form the unlabeled set. Figure 2 displays the results of our approach with $P = 2$ noise components averaged over 10 random splits of the data with labeled set sizes $M = 4, 7, 10, 15$ and with 0, 50 and 90 percent view disagreement. Single view GP regression-based classification performance and the performance of Bayesian co-training are also shown for comparison.

At zero percent view disagreement both Bayesian co-training and our approach give good performance, and improve over the single-view baselines. At non-zero view disagreement levels, however, Bayesian co-training is no longer able to improve over single-view performance and in fact often under-performs. The single-view baselines also degrade in the presence of view corruption since they are unable to reliably infer class labels over the occluded samples. In contrast, our approach is able to benefit from view combination and successfully infer the class labels even in the presence of gross view corruption (up to 90% view disagreement).

In Figure 3 the performance of our method compared to the single- and multi-view baselines on the multi-view image dataset is also shown for fixed training set sizes with varying view disagreement levels, averaged over the same splits used to generate Figure 2. In contrast to Bayesian co-training our approach is able to sustain good performance across all view disagreement levels, even with relatively few labeled training examples per class ($M = 4$).

Next we illustrate our approach on the audio-visual user agreement dataset from head gesture and speech. Similar to the previous experiments we separated the data into M samples per class for labeled set and 50 samples per class for the unlabeled dataset. Figure 4 shows the performance of our approach with $P = 2$ averaged over 10 random splits of the data over labeled set sizes M = 4,7,10,15 and with 0, 50 and 90 per-



(Comparison with [18] 0% VD) (Comparison with [18] 50% VD) (Comparison with [18] 98% VD)

(Comparison with [6] 0% VD) (Comparison with [6] 50% VD) (Comparison with [6] 98% VD)

Figure 4. **Audio-visual recognition with varying training set sizes**: Classification accuracy as a function of the number of training samples for 0%, 50% and 98% view disagreement. Performance is shown averaged across 10 splits, the error bars indicate ± 1 std. deviation. (top) Comparison with single-view and Bayesian co-training approaches, and (bottom) audio and video classifiers from filter-based co-training [6] and the results of multi-view GP kernel combination (see text for details). In contrast to the baseline approaches, our method is able successfully combine each view to achieve good classification accuracy even in the presence of gross view corruption (98% view disagreement).

cent view disagreement. As before, the performance of single view GP regression-based classification and Bayesian co-training are also shown. Figure 5 displays the same comparison over fixed training set sizes and for varying amounts of view disagreement.

Unlike the multi-view image database there is a clear imbalance between each of the modalities, where the audio modality is much weaker than the visual one. Yet, without any a priori knowledge of which is the more reliable modality both our approach and Bayesian co-training are able to effectively combine the views and retain the good performance of the visual modality in the presence of zero percent view disagreement. For non-zero view disagreement the performance of Bayesian co-training degrades and in contrast to all three baseline methods our approach is able to maintain relatively good performance even with up to 98% view disagreement.

We also compared our approach to the filter-based co-training approach of Christoudias et. al. [6] on the audio-visual user agreement dataset. Figure 4 displays the performance of our approach and the performance of the naive Bayes audio and visual classifiers obtained from the filter-based co-training technique of [6] averaged over 10 splits of the data with training set sizes M = 4,7,10,15 and with 0, 50 and 98 percent view disagreement. Similarly, Figure 5 displays average performance over fixed training set sizes and with varying amounts of view disagreement.

The filter-based co-training baseline assumes that the conditional entropy formed by conditioning one view on a corrupted sample from another view is higher than that obtained by conditioning on an un-corrupted sample. In the absence of a dominant background class, this assumption does not hold for binary classification and filter-based co-training performs poorly. In contrast, our approach can model a wider range of view disagreement distributions and outperforms filter-based co-training on this task.

Supervised multi-view kernel combination approaches have recently received much attention in the machine learning and computer vision literature [2]. Kernel combination approaches can suffer in the presence of sample-dependent noise such as that caused by view disagreement, especially when there is an imbalance between each view or feature set as is the case in the audio-visual user agreement dataset. Figures 4 and 5 also display the performance of a kernel combination GP baseline whose covariance function is modeled as the weighted sum of the covariance functions from

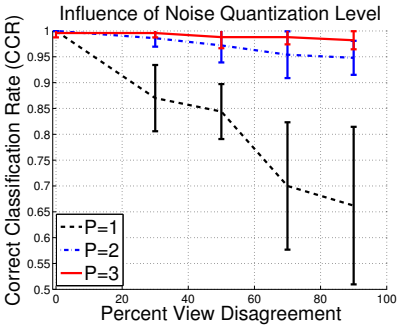


Figure 7. **Simultaneously coping with partial occlusion and view disagreement**: Influence of the number of noise levels P on classification accuracy when the multi-view image data is corrupted by view disagreement and partial occlusion. Performance is shown averaged over 10 splits with $M = 7$, error bars indicate $\pm$ 1 std. deviation. As expected performance improves with increasing model components. With $P = 1$ our model is equivalent to [18].

each view. The performance of the multi-view kernel combination baseline degrades in the presence of view disagreement on this audio-visual gesture recognition task.

Finally, we evaluated the performance of our approach using both maximum likelihood and n-fold cross-validation parameter learning. Figure 6 displays the performance under each technique evaluated with both datasets, with $M = 10$ and varying amounts of view disagreement, where maximum likelihood parameter learning was initialized from n-fold cross validation. Across both datasets n-fold cross-validation either matches or outperforms maximum likelihood performance.

4.2. General noise

Our multi-view learning approach can also cope with more general forms of view corruption or noise, beyond binary view disagreement. To illustrate this point we evaluated our approach on the multi-view object dataset with the views corrupted by two different noise processes, partial and complete occlusion.

Under this setting, we tested our approach with different noise quantization levels, P. Figure 7 displays the performance of our approach for P = 1,2,3. For P=1 our approach defaults to the Bayesian co-training baseline. For greater values of P our approach does increasingly better, since this gives our model greater flexibility to deal with the different types of noise present in the data. As expected P=3 does the best, since unlike P=2 it can further differentiate between partially and entirely occluded samples.

5. Conclusion

In this paper we have introduced *Heteroscedastic Bayesian Co-training*, a probabilistic approach to multi-view learning that simultaneously discovers the amount of noise on a per-sample basis, while solving the classification task. We have demonstrate the effectiveness of our approach in two domains, multi-view object recognition from low-fidelity sensor networks and audio-visual user agreement recognition. Our approach, unlike state-of-the-art co-training approaches, results in high performance when dealing with large amounts of partially occluded and view disagreement observations. Interesting avenues of future work include the generalization of our approach to non-i.i.d. sample-dependent noise models and the application of our approach to modeling sample dependent distances in multi-view kernel combination-based object category classification schemes.

Acknowledgements

Funding for this research was provided in part by the DOD and by NSF contract IIS-0704479.

References

[1] R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *ICML*, 2007.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[4] P. Chen, P. Ahammad, C. Boyer, H. S. Huang, L. Lin, E. Lobaton, M. Meingast, S. Oh, S. Wang, P. Yan, A. Y. Yang, C. Yeo, L.-C. Chang, J. Tygar, and S. S. Sastry. Citric: A low-bandwidth wireless camera network platform. In *ICDSC*, 2008.

[5] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *ICMI*, November 2006.

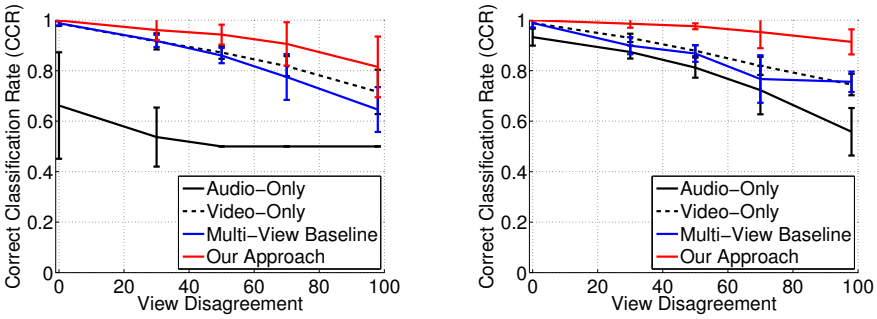
[6] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *UAI*, 2008.

[7] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *SIGDAT*, 1999.

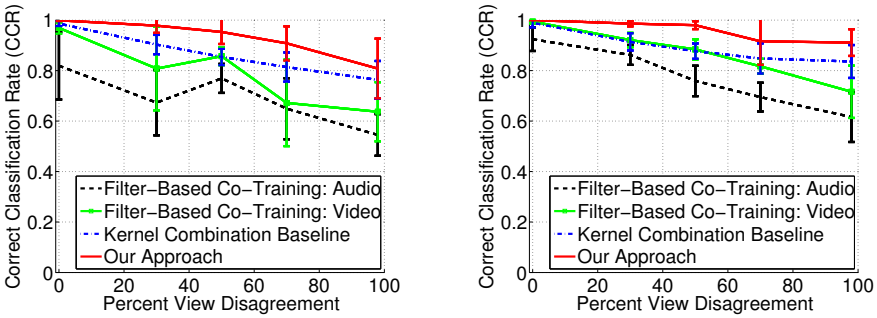
[8] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 2007.

[9] A. Halberstadt. *Heterogeneous Acoustic Measurements and Multiple Classiers for Speech Recognition.* PhD thesis, MIT, 1998.

[10] S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *COLT*, 2007.



(Comparison with [18] $M = 7$) (Comparison with [18] $M = 15$)



(Comparison with [6] $M = 7$) (Comparison with [6] $M = 15$)

Figure 5. **Audio-visual recognition with varying levels of view disagreement**: Classification accuracy as a function of the level of view disagreement. Performance is shown averaged over 10 splits, error bars indicate ± 1 std. deviation. (top) Comparison to single-view and Bayesian co-training approaches and (bottom) filter-based co-training and multi-view GP kernel combination. The audio-visual dataset contains imbalanced views which in the presence of per-sample view corruption adversely affects multi-view kernel combination. Unlike the baseline methods, our approach is robust to large amounts of view disagreement even when the views are imbalanced.

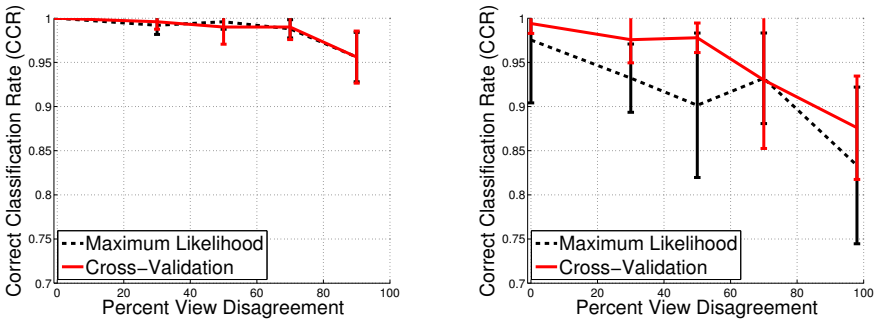


Figure 6. **Cross-Validation vs. Maximum Likelihood**: Average performance is shown over 10 splits with 10 labeled examples per class for (left) the multi-view image database and (right) the audio-visual gesture database. Cross-validation either matches or outperforms maximum likelihood across both datasets.

[11] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using cotraining. In *ICCV*, 2003.

[12] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *CVPR*, 2003.

[13] I. Muslea, S. Minton, and C. A. Knoblock. Adaptive view validation: A first step towards automatic view detection. In *ICML*, 2002.

[14] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of cotraining. In *Workshop on Information and Knowledge Management*, 2000.

[15] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[16] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *ICML*, 2005.

[17] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *CVPR*, 2005.

[18] S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and R. B. Rao. Bayesian co-training. In *NIPS*, 2007.