

Co-training with Noisy Perceptual Observations

C. Mario Christoudias*, Raquel Urtasun*, Ashish Kapoor‡ and Trevor Darrell*

{cmch,rurtasun}@icsi.berkeley.edu, akapoor@microsoft.com, trevor@eecs.berkeley.edu

*UC Berkeley EECS & ICSI

‡ Microsoft Research

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). When labeled data is scarce, multi-view data can be exploited to learn more effectively under weak or partial supervision compared to learning from only a single view. Semi-supervised and transductive learning in the presence of multiple views has received considerable recent interest in the machine learning community, and a class of techniques based on the classic ‘co-training’ method [1] 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 [4, 2, 5, 8].

With a few notable exceptions [2, 8, 5], however, co-training-type methods have had only limited success on visual tasks. We argue here that this is due in part to restrictive assumptions inherent in existent 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 [8, 6, 9], however, more complex forms of noise such as per sample occlusion 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. [3] 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, their method performs poorly in the latter case.

Recently, Yu et al. [9] 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, and assume a Gaussian process prior [7] on the latent variables. To deal with noisy data, in this work 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 parametrized by per view noise covariances \mathbf{A}_j . The only restriction on \mathbf{A}_j is to be positive semi-definite so that the resulting matrix is a Mercer kernel and its inverse can be computed. Figure 1 depicts the undirected graphical model of our *Heteroscedastic Bayesian Co-training* approach.

Integrating out the latent functions \mathbf{f}_j in our model results in a GP prior over the consensus function that favors agreement between the latent functions \mathbf{f}_j [9]. The resulting multi-view *heteroscedastic co-training kernel* 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 handles semi-supervised and transductive settings as our kernel is non-stationary and

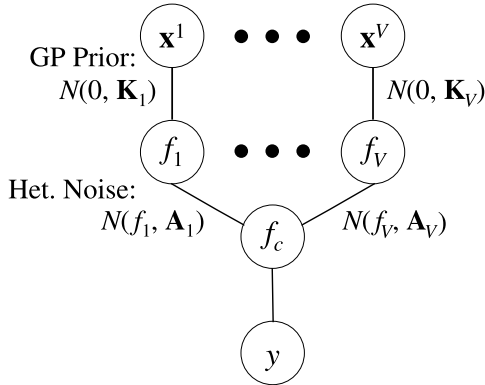


Figure 1: Graphical model of *Heteroscedastic Bayesian Co-training* (our approach). Our multi-view learning approach extends Bayesian co-training to incorporate sample-dependent noise modeled by the per view noise covariance matrices \mathbf{A}_j . Our method simultaneously discovers the amount of noise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

therefore depends on both the labeled and unlabeled data.

Learning the heteroscedastic model consists of solving for the kernel hyperparameters (e.g., RBF width) and the noise covariances \mathbf{A}_j defined in each view. Under this model, the number of parameters to estimate is very 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. In this work, we consider a *quantized-i.i.d.* sample-dependent noise model where we assume that the noise corrupting each sample is due to one of P noise processes. Such a model is useful, for example, when coping with different levels of per-sample occlusions (e.g., a sample is either occluded, partially occluded or un-occluded).

In our experiments 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.

References

- [1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.
- [2] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *ICMI*, November 2006.
- [3] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *UAI*, 2008.
- [4] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *SIGDAT*, 1999.
- [5] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using cotraining. In *ICCV*.
- [6] I. Muslea, S. Minton, and C. A. Knoblock. Adaptive view validation: A first step towards automatic view detection. In *ICML*, 2002.
- [7] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- [8] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *CVPR*, June 2005.
- [9] S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and R. B. Rao. Bayesian co-training. In *NIPS*, 2007.