Probabilistic Models for Multi-View Semi-Supervised Learning and Coding

by

C. Mario Christoudias

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author . . . . . .
Department of Electrical Engineering and Computer Science
July 6, 2009

Certified by . . .
Trevor J. Darrell
Associate Professor
Thesis Supervisor

Accepted by . . . . . .
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Probabilistic Models for Multi-View Semi-Supervised Learning and Coding

by

C. Mario Christoudias

Submitted to the Department of Electrical Engineering and Computer Science
on July 6, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

This thesis investigates the problem of classification from multiple noisy sensors or modalities. Examples include speech and gesture interfaces and multi-camera distributed sensor networks. Reliable recognition in such settings hinges upon the ability to learn accurate classification models in the face of limited supervision and to cope with the relatively large amount of potentially redundant information transmitted by each sensor or modality (i.e., view). We investigate and develop novel multi-view learning algorithms capable of learning from semi-supervised noisy sensor data, for automatically adapting to new users and working conditions, and for performing distributed feature selection on bandwidth limited sensor networks. We propose probabilistic models built upon multi-view Gaussian Processes (GPs) for solving this class of problems, and demonstrate our approaches for solving audio-visual speech and gesture, and multi-view object classification problems.

Multi-modal tasks are good candidates for multi-view learning, since each modality provides a potentially redundant view to the learning algorithm. On audio-visual speech unit classification, and user agreement recognition using spoken utterances and head gestures, we demonstrate that multi-modal co-training can be used to learn from only a few labeled examples in one or both of the audio-visual modalities. We also propose a co-adaptation algorithm, which adapts existing audio-visual classifiers to a particular user or noise condition by leveraging the redundancy in the unlabeled data. Existing methods typically assume constant per-channel noise models. In contrast we develop co-training algorithms that are able to learn from noisy sensor data corrupted by complex per-sample noise processes, e.g., occlusion common to multi-sensor classification problems. 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 accurate performance in the presence of occlusion or other complex noise processes. We also investigate an extension of this idea for supervised multi-view learning where we develop a Bayesian multiple kernel learning algorithm that can learn a local weighting over each view of the input space.

We additionally consider the problem of distributed object recognition or indexing from multiple cameras, where the computational power available at each camera sensor is limited and communication between cameras is prohibitively expensive. In this scenario, it is desirable to avoid sending redundant visual features from multiple views. Traditional supervised feature selection approaches are inapplicable as the class label is unknown at each camera. In this thesis, we propose an unsupervised multi-view feature selection algorithm based on a distributed coding approach. With our method, a Gaussian Process model of the joint view statistics is used at the receiver to obtain a joint encoding of the views *without* directly sharing information across encoders. We demonstrate our approach on recognition and indexing tasks with multi-view image databases and show that our method compares favorably to an independent encoding of the features from each camera.

Thesis Supervisor: Trevor J. Darrell
Title: Associate Professor

Acknowledgments

First and foremost I would like to thank my advisors Trevor Darrell and Raquel Urtasun for their endless support and guidance. They have played a central role in my development as a researcher, and without them none of this would have been possible for which I am forever grateful. I would also like to thank my thesis readers Regina Barzilay and Bill Freeman for their valuable feedback and improvements to this thesis.

I have had the opportunity to collaborate with many bright and gifted researchers from whom I have also learned a great deal. In particular, I would like to thank Louis-Philippe Morency, who has not only been a mentor but also a great friend, and Kate Saenko a valued friend and collaborator. I would like to thank Jacob Eisenstein, Ashish Kapoor, Subhransu Maji, and Allen Yang with whom I have also enjoyed working with and learning from.

Thanks to the past and current members of the vision interfaces and ICSI vision groups for creating a fun and productive working environment, including, Neal Checka, David Demirdjian, Carl Ek, Mario Fritz, Kristen Grauman, Naveen Goela, Brian Kulis, John Lee, Ariadna Quattoni, Ali Rahimi, Mathieu Salzmann, Gregory Shakhnarovich, Michael Siracusa, Patrick Sundberg, Kevin Wilson, and Tom Yeh. Also thanks to the many friends that I have made while at MIT for their friendship and encouragement over the years, specifically, Xavier Carreras, Rodney Daughtrey, Rick Frauton, David Fried, Karen Livescu, and Summer Sheremata.

Finally and most importantly, I would like to thank my parents John and Aphrodite Christoudias and my sister Tina Christoudias, their love and support has helped make this dream a reality and I could not have done it without them, I love you dearly.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in sensing technologies have made possible the development of computational systems that exploit multiple sensor streams or modalities to solve an end-goal or task. The abundance of cheap, diverse sensors now available has had far reaching implications in many facets of computer science and engineering. In computer vision and graphics, multi-camera systems have been employed for surveillance and scene understanding [55, 107, 50, 88, 108, 109, 92] and for building photo-realistic models of object appearance useful for animation and scene rendering [64, 10, 14, 15]. Speech recognition systems have gone beyond audio-only classification and also incorporate other modalities such as lip appearance to robustly recognize peoples' speech in noisy environments [78, 83]. The availability of digital cameras for everyday use has shaped the landscape for digital photography and the world-wide web, where it is now common-place that web documents are characterized both by textual and visual content. Sensor networks and robot navigation systems now utilize a wide array of modalities for understanding, navigating and manipulating their environment including infra-red, intensity and laser imagery [114, 18, 17].

In many of these areas the availability of multiple data streams has proven useful for classification problems [78, 83, 123, 88, 92, 89] where the observations from each stream act as redundant, but complementary instantiations of an underlying class or event. Classification from multiple sources or modalities offers many potential advantages over classification from a single source alone, and brings to light new and

promising solutions to difficult machine learning problems. In real-world scenarios it is often the case that labeled data is difficult to obtain, however, unlabeled data is readily available, and that there exists a mis-match between the distribution of the data available during training and that of the data observed during system deployment. As multi-sensor systems are becoming more common-place, learning algorithms that exploit multiple sources and can take advantage of unlabeled data are desirable.

In this thesis we investigate and develop algorithms that leverage multiple data streams to perform classification and learn from both labeled and unlabeled data, based on multi-view learning concepts from the machine learning community and multi-view coding concepts from the distributed coding literature. We first exploit cross-modal redundancy to develop novel semi-supervised learning schemes. The idea of utilizing multiple feature splits or views to learn from partially labeled data was first formalized by Blum and Mitchell [11] as the *co-training algorithm* that iteratively learns a multi-view classifier by having the classifier in one view provide training labels to the classifiers in the other views. In [11], Blum and Mitchell describe co-training within a PAC learning framework and formulate the modeling assumptions necessary for its successful application. Since its development, techniques have been proposed to relax these assumptions and a variety of *multi-view learning* methods have been formulated that function on the more general principle of maximizing classifier agreement over the unlabeled data [26, 97, 98, 128, 51, 6]. Techniques in multi-view learning can be generally categorized into iterative methods such as original co-training [11], and those that impose an agreement-based prior or regularization term over the unlabeled data [97, 128]; both variants are considered in this thesis.

Multi-view learning methods have been successfully applied to a variety of classification problems [26, 63, 123], however, they have been limited by their ability to cope with sensor noise. In real-world systems the observation noise often varies per view and classification from either view alone is not equally reliable. Similarly, for many datasets the noise varies per sample, e.g., as a result of occlusion or other non-stationary noise processes. 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 this assumption can be violated quite dramatically. A variety of approaches have been proposed to deal with simple forms of view insufficiency [123, 75, 128]. More complex forms of noise such as per-sample noise, however, have received less attention. We develop here co-training algorithms that are robust to complex sample corruption and *view disagreement*, i.e., when the samples from each view do not belong to the same class due to occlusion or other forms of view corruption.

A working assumption of many machine learning techniques is that the training and test data belong to the same underlying distribution. For many problems, however, there exists a mis-match between these distributions. For example, in speech and gesture interfaces the target distribution varies largely with user and/or environment, and although generalization can be achieved to a fair extent by a generic model, a large increase in performance can be seen if the model is adapted to the specific end-user or working condition [49]. Numerous techniques for *model adaptation* have been proposed particularly for the application areas of speech recognition and more generally human-computer interfaces [81, 121, 49]. Multi-view learning algorithms such as co-training can be formulated for the task of model adaptation. In Chapter 3, we develop a co-adaptation algorithm that bootstraps a generic classifier to a specific user or environment within a multi-view learning framework. We demonstrate this approach for learning user-specific speech and gesture classifiers and demonstrate favorable performance to single-view adaptation techniques.

Another challenge in multi-sensor classification is coping with the potentially large amount of information from each sensor. This is particularly a problem in distributed networks where classification is performed from a set of remote sensors transmitting information over a bandwidth limited network. In machine learning many feature selection algorithms have been proposed that form compact representations of the data by optimizing over some performance criteria such as minimizing redundancy [33, 67, 82] or maximizing classification discriminance [37, 117, 82]. Techniques have also been explored for the multi-view scenario [6, 85]. Yet, in many distributed sensor

networks, it is often the case that the computation available at each sensor is limited and communication between sensors is prohibitively expensive, for which many of these techniques are in-applicable.

In the presence of limited computation, unsupervised feature selection can be applied, however, without explicitly sharing information between sensors it can only be done naively at each sensor. Ideally one would like to exploit the mutual information between sensors to achieve a more compact feature representation of the joint feature space. Work in information theory has shown that encoding rates close to the joint entropy can be achieved for a set of sensors sharing a common receiver even *without* explicitly sharing information between sensors. This was first shown by Slepian and Wolf [103] and many distributed coding methods have since been formulated [119, 84, 1, 132, 19, 24, 91, 93, 127]. In what follows, we develop a novel distributed coding algorithm in the context of bag-of-words object recognition from multiple camera sensors on a distributed network and demonstrate comparable performance to multi-view classification while achieving a large compression rate.

The problems of classification and distributed coding from multiple noisy sensor streams can be naturally cast within a probabilistic framework. Gaussian Processes (GPs) form a class of non-parametric probabilistic models for performing regression and classification that have been shown to perform quite well on a large variety of tasks [52, 90, 111, 116, 13]. They define a principled, probabilistic framework for kernel machine classification and generalize existing techniques [86]. GPs are general and flexible models, that are also highly practical, as inference can often be performed in closed form[1] and model specification only involves the specification of a kernel function and its hyper-parameters which can be learned via maximum likelihood. Unlike other kernel machines, such as the Support Vector Machine (SVM), GPs are probabilistic and offer a measure of prediction uncertainty useful for performing classification and feature selection [52].

[1]Note this assumes a Gaussian noise model in the output which is typically what is used for regression problems. In the context of classification, a Gaussian noise model can also be used although more sophisticated models such as the logit function have been explored [86].

In this thesis, we propose probabilistic approaches with multi-view GPs for performing supervised and semi-supervised multi-view learning and distributed coding in realistic noisy environments. In Chapter 5, we propose a heteroscedastic Bayesian co-training algorithm that is an extension of the work of [128] to model per-sample noise processes, e.g., occlusion. With our approach the noise of each stream is simultaneously discovered while solving the classification task. We demonstrate our approach on the tasks of user agreement recognition from head gesture and speech and multi-view object classification and show that unlike state-of-the-art multi-view learning approaches our method is able to faithfully perform semi-supervised learning from noisy sensor data. An extension of this idea is explored in Chapter 6 for supervised multi-view learning where we develop a Bayesian multiple kernel learning algorithm that can learn a local weighting over each view of the input space. In Chapter 7, we also develop a similar model with multi-view GPs for performing distributed coding of visual feature histograms. With our approach the joint feature histogram distribution is modeled at the common receiver and the GP prediction uncertainty is used to reconstruct the histogram in each view given the reconstructed histograms from previous views. We evaluate our approach on the COIL-100 multi-view image dataset for the task of instance-level object recognition and demonstrate comparable performance to multi-view classification while achieving a large compression rate without explicitly sharing information across views.

1.1 Thesis Contributions

The contributions of this thesis are summarized below.

- **Application of multi-view learning to multi-modal human-computer interfaces:** Multi-modal classification is well suited for multi-view learning because each modality provides a potentially redundant view to the learning algorithm. While the concept of multi-modal co-training was mentioned as promising future work in the seminal Blum and Mitchell paper [11], it appears that there has been relatively little subsequent work on cross-modal co-

training. This thesis investigates the use of co-training for learning audio-visual speech and gesture classifiers and demonstrates its effectiveness for learning audio-visual classifiers from partially labeled data, and for performing model adaptation. It develops and evaluates a novel model adaptation algorithm, *co-adaptation*, that adapts a generic model to a specific user and/or environment with co-training. Our approach is demonstrated on the tasks of user agreement recognition from speech and gesture and audio-visual speech recognition. These results are reported in [20].

- **Investigation of view disagreement in multi-view learning:** This thesis identifies and investigates a new form of view insufficiency called *view disagreement*, i.e., when a sample belongs to a different class than the samples in the other views as a result of view corruption or noise. In multi-senor perceptual learning problems common examples of view disagreement include occlusion and uni-modal expression. We propose a filter-based co-training algorithm that utilizes an information theoretic criterion to detect and filter view disagreement samples during co-training. Our experiments demonstrate that unlike existing techniques, our filter-based co-training algorithm is able to learn accurate multi-view classifiers despite view disagreement. These results are reported in [21].

- **Development of a probabilistic co-training algorithm for learning from noisy data:** Perceptual learning problems often involve datasets corrupted by complex noise processes, such as per-sample occlusion. Multi-view learning algorithms have difficulty learning from such noisy data. In this thesis we extend the Bayesian co-training algorithm of Yu et. al. [128] to model per-sample noise and other complex noise processes. Our approach simultaneously discovers the noise while solving the classification task and can handle arbitrary view corruption processes including binary view disagreement. We demonstrate our approach for performing multi-view semi-supervised learning within a variety of perceptual learning tasks. These results are reported in [23].

- **Development of a Bayesian localized multiple kernel learning algorithm:** Most multiple kernel learning approaches are limited by their assumption of a per-view kernel weighting. We propose a Bayesian multiple kernel learning algorithm with Gaussian Processes that can learn a local weighting over each view and obtain accurate classification performance from insufficient views corrupted by complex noise, e.g., per-sample occlusion, containing missing data, and/or whose discriminative properties vary across the input space. We evaluate our approach on the tasks of audio-visual gesture recognition and object category classification with multiple feature types.

- **Development of a Gaussian Process distributed feature selection algorithm for multi-view object and scene recognition:** Feature selection is an important problem in machine learning that has close connections with data compression techniques in information theory. In the case of multi-sensor data, feature selection is of particular importance since the data from multiple sources is often high dimensional and highly redundant. This thesis develops a distributed feature selection algorithm with Gaussian Processes (GPs) borrowing concepts from distributed source coding in information theory. We demonstrate our approach for visual feature selection from distributed multi-camera systems for performing multi-view object recognition. Our approach is evaluated on both synthetic and real-world datasets, and achieves high distributed compression rates while maintaining accurate multi-view recognition performance. These results are reported in [22].

1.2 Thesis Outline

The thesis chapters are organized as follows.

- **Chapter 2**: Related Work and Background

  This chapter provides a brief overview of GP regression and classification, multi-view learning, and distributed coding and discusses related work.

- **Chapter 3**: Co-Adaptation of Audio-Visual Speech and Gesture Classifiers

  This chapter investigates the use of co-training for learning audio-visual speech and gesture classifiers from partially labeled data. It develops a multi-view model adaptation algorithm, *co-adaptation* that adapts a generic model to a specific user and/or environment with co-training. Both co-training and co-adaptation are evaluated on the tasks of audio-visual user agreement classification from speech and gesture and audio-visual speech recognition.

- **Chapter 4**: Multi-View Learning in the Presence of View Disagreement

  This chapter identifies and investigates a new form of view insufficiency called *view disagreement*, i.e., when a sample belongs to a different class than the samples from other views as a result of view corruption or noise. It develops a filter-based co-training algorithm that builds upon an information theoretic criterion for detecting and filtering view disagreement samples during co-training. Experiments are carried out for the task of audio-visual user agreement recognition from speech and gesture. Unlike other state-of-the-art multi-view learning methods the filter-based co-training approach is able to faithfully learn from partially labeled data despite view disagreement.

- **Chapter 5**: Co-training with Noisy Perceptual Observations

  This chapter develops a probabilistic co-training algorithm that is able to learn from noisy datasets corrupted by complex noise process such as per-sample occlusion, common to multi-sensor perceptual learning tasks. It extends the Bayesian co-training algorithm of Yu et. al. [128] to model per-sample noise and other complex noise processes. The resulting *heteroscedastic Bayesian co-training* approach simultaneously discovers the noise while solving the classification task and can handle a variety of complex noise processes including binary view disagreement. This approach is evaluated on the tasks of multi-view object and audio-visual user agreement classification.

- **Chapter 6**: Localized Multiple Kernel Learning with Gaussian Processes

This chapter presents a localized multiple kernel learning algorithm with Gaussian Processes that can learn a local weighting of the input space. Unlike, global approaches, we demonstrate that our approach can cope with insufficient input views corrupted heteroscedastic noise processes, missing data, and whose discriminative properties can vary across the input space. We demonstrate our approach on the tasks of audio-visual gesture recognition and object category classification on the Caltech-101 benchmark.

- **Chapter 7**: Unsupervised Visual Feature Selection via Distributed Coding

  This chapter proposes a novel unsupervised feature selection algorithm with Gaussian Processes that borrows concepts from distributed source coding. This algorithm is demonstrated for the task of visual feature selection on a distributed multi-camera object recognition system. An evaluation is performed on both synthetic and real-world datasets. The proposed distributed feature selection strategy achieves high distributed compression rates while maintaining accurate multi-view recognition performance.

- **Chapter 8**: Conclusion

  This chapter summarizes the contributions of this thesis and discusses future work.

Chapter 2

Related Work and Background

In this chapter we discuss related work and background. We begin with an overview of multi-view learning methods, followed by a discussion of GP regression and classification, and techniques in distributed coding.

2.1 Multi-View Learning

For many machine learning problems acquiring labeled data is a costly and tedious process, however, unlabeled data is readily available. For example, unlabeled training data for the classification of web documents and/or images can be easily obtained in relatively large quantities from crawling the world wide web, however, labeling these images or documents can be a time consuming and difficult process. Techniques in semi-supervised learning aim to exploit both labeled and unlabeled data for learning a classifier and limit the need for human supervision [58, 5, 53, 11]. For many of these approaches, the unlabeled data is utilized to constrain or simplify the learning problem such that only a few labeled training examples are needed to construct an accurate classifier.

Multi-view learning methods form a class of semi-supervised learning techniques that exploit multiple views or feature splits of the data to learn under limited supervision [11, 26, 79, 128, 6, 51, 98]. One of the first instantiations of multi-view learning was explored in the work of Yarwosky for performing word sense disambiguation from

text documents [125] (e.g., the word 'plant' has multiple senses and, for instance, can mean a type of organism or a factory). In [125], Yarwosky split each document into two views for classifying word sense: 1) the dominant sense of the document, and 2) the local context surrounding the word of interest. The intuition is that the sense of the word remains predominantly the same within a given document. Also, the surrounding context of the word is indicative of its meaning (e.g., 'The *machines* at the plant are being upgraded today.'). By splitting each document into two views Yarwosky was able to obtain improved performance over single-view training from labeled and unlabeled data.[1]

The idea of using multiple views of the data to bootstrap and improve a set of weak classifiers from unlabeled data was later formalized by Blum and Mitchell as the *co-training algorithm* [11]. In their work Blum and Mitchell discuss how agreement between a set of target functions over the unlabeled data distribution can be used to constrain the solution space and simplify learning. The general notion of maximizing agreement between a set of classifiers to learn from partially labeled data was later introduced by Collins and Singer [26] and is the underlying principle behind many multi-view learning techniques [97, 128, 9, 96, 51]. In this light, multi-view learning algorithms can be thought of as semi-supervised learning methods that employ an agreement-based regularizer or prior over the unlabeled data to constrain the set of possible solutions [11]. This interpretation of multi-view learning has been seen in the co-regularization algorithm of Sindhwani et al. [98] that forms an agreement-based regularizer for regularized least squares and SVM classification and the Bayesian co-training algorithm of Yu et. al. [128] that utilizes an agreement-based prior within the context of a probabilistic model.

Co-training has been applied in a variety of application areas including natural language processing [26, 125], computer vision [63] and human-computer interfaces [123]. Collins and Singer learn named entity classifiers from partially labeled data with co-training [26]. Levin et. al. demonstrate the use of co-training for performing

[1] Abney [2] later provided a formal analysis of the single-view Yarwosky algorithm and further explored its connections to multi-view learning.

car detection from images [63]. Similarly, Yan and Naphade apply co-training for video-concept detection from speech transcripts and video [123].

As discussed by Blum and Mitchell the co-training algorithm makes the assumptions that each view is class conditionally independent and is sufficient for classification, i.e., classification can be performed from either view alone [11]. The first assumption guarantees that samples labeled by the classifiers across views are randomly selected, i.e., that there is no labeling bias. View sufficiency implies that the observations across views belong to the same class and that the target functions in each view do not disagree on any given input sample; loosely speaking view sufficiency is a necessary condition for enforcing agreement between the classifiers learned in each view.

The assumptions of class conditional independence and view sufficiency are generally hard to satisfy in practice, and although co-training has seen empirical success it limits the general application of co-training to new problem domains. A fair amount of work in the multi-view learning literature has focused on relaxing these assumptions [8, 51, 26]. Balcan et. al. [8] propose a relaxation of the class conditional independence assumption using graph-expansion theory. They show that a problem need only satisfy a weaker assumption of expansion in order for iterative co-training to be applicable. To overcome limitations due to view insufficiency techniques have been proposed that optimize over view agreement [26, 97] and those that model the noise of each view [123, 75, 128]. In this thesis we propose techniques of the later form that explicitly model the noise inherent in each view to cope with complex forms of view insufficiency and view corruption.

A formal description of the co-training algorithm and a brief summary of the findings of that work is provided in the following sub-section. Related multi-view learning methods are then outlined followed by a discussion of the view sufficiency assumption and state-of-the-art approaches for coping with view sufficiency.

2.1.1 Co-Training Algorithm

Co-training functions over separate feature splits or views of the data and trains a

Algorithm 1 Co-training Algorithm

Given a small labeled set S, a large unlabeled set U, k views, and parameters N and T:
Set $t = 1$
repeat
  for $i = 1$ to k **do**
    Train classifier f_i on view i of S
    Use f_i to label U, move N most confidently labeled samples to S
  end for
  Set $t = t + 1$
until $t = T$ or $|U| = 0$

set of classifiers one per view by mutually bootstrapping them from partially labeled data. Let $S = \{(\mathbf{x}_i, y_i)\}$ be a set of labeled examples, typically referred to as the data *seed set* and $U = \{\mathbf{x}_i\}$ the unlabeled data set. Furthermore, we decompose $\mathbf{x}$ into V feature splits or views such that $\mathbf{x} = (\mathbf{x}^1, \cdots, \mathbf{x}^V)$. Here, $\mathbf{x}^v$ can generally be thought of as a subset of the features of $\mathbf{x}$, however, can also correspond to a physical view of the data, e.g., an audio or video modality. Similarly, let $h^v(\mathbf{x}^v)$ be the hypothesis or classification function defined in view $v = 1, \cdots, V$ and $f^v(\mathbf{x}^v)$ be the underlying view target functions that we wish to learn from the partially labeled data, such that $h^{v*}(\mathbf{x}_i^v) = f^v(\mathbf{x}_i^v), \forall \mathbf{x}_i^v$. For brevity of notation we'll refer to these functions as h^v and f^v respectively.

The co-training algorithm is initialized from the seed set by training a weak classifier, h^v, from each view of the labeled data. This classifier is then improved from the unlabeled data U as follows. Each classifier is evaluated in turn on the unlabeled data and the N most confidently classified samples are added to the labeled set, S. After evaluation, the classifiers are then retrained on the expanded set. This process of evaluation and training is then repeated until all the data has been labeled or an early stopping criterion is met, e.g., the algorithm reaches a set number of iterations T. The co-training algorithm is summarized in Algorithm 1[2].

[2]In [11], Blum and Mitchell propose a slightly modified version of the algorithm that functions over subsets of the unlabeled data set U at each iteration to promote randomness in selection of examples and that is specified for only two views, however, the algorithm as presented in Algorithm 1 is its more general form. The algorithm proposed in [11] can be seen as employing a particular type of confidence measure.

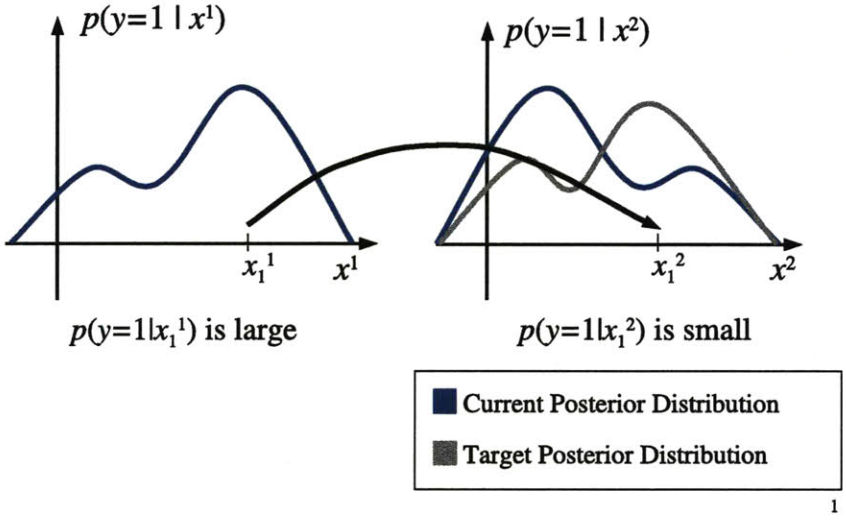


Figure 2-1: Illustration of the co-training algorithm. The classifier posterior and target distributions are shown for a two-view problem. The weak classifier in view two assigns a low probability to observation x_1^2 even though it has high probability under the ground-truth target distribution. Unlike single-view bootstrapping co-training can learn from both low and high probability observations. In this example, co-training labels the low probability observation x_1^1 using the corresponding high confidence observation, x_1^2, in the other view.

Co-training can be characterized as the multi-view equivalent of the bootstrapping algorithm for learning from partially labeled data [34]. It differs from single-view bootstrapping in that it has the ability to discover both *low* and high probability examples under the current hypothesis function and it can therefore learn more complex data distributions. As a high-level illustration of this point consider the two-view example depicted in Figure 2-1. In the Figure, the posterior distributions of two classifiers along with the target distributions is shown. Consider encountering the example $\mathbf{x}_i$ during bootstrapping. A single-view approach would have difficulty assigning $\mathbf{x}$ the correct label, however, co-training can more easily discover this low probability

example by classifying the observation in the other view.

In their paper Blum and Mitchell analyze co-training within a PAC learning framework and give intuitions as to how utilizing multiple data feature splits can be used to reduce problem complexity given both labeled and unlabeled data. As discussed in [11], the co-training model assumes that each data view, $\mathbf{x}^v$, is class conditionally independent, such that,

$$p(\mathbf{x}^1, \cdots, \mathbf{x}^V | l) \propto p(\mathbf{x}^1 | l) \cdots p(\mathbf{x}^V | l) \tag{2.1}$$

where l is the class. Under this assumption the labeled samples across views are randomly selected instances of the class, and co-training is able to achieve better generalization from the unlabeled data. In particular, Blum and Mitchell show that assuming that the target concepts are learn-able within the standard PAC model with classification noise and that each view satisfies class conditional independence, a set of weak hypothesis, h^v, can be improved to arbitrary precision from unlabeled data with co-training [11].

The second co-training assumption is that each view is sufficient for classification, i.e., that classification can be performed from either view alone, such that,

$$D(\mathbf{x}_i) = 0, \quad \forall \mathbf{x}_i \text{ s.t. } f^j(\mathbf{x}_i^j) \neq f^k(\mathbf{x}_i^k), \quad j \neq k \tag{2.2}$$

where $D(\mathbf{x})$ is the distribution of the random variable $\mathbf{x}$. Under this assumption of view *compatibility* or sufficiency Blum and Mitchell show that multiple feature splits can be used to limit the need for supervision in the presence of unlabeled data. In particular, they consider a two view problem and draw the multi-view samples as a bi-partite graph with one view per side of the graph and edges connecting related samples across views for which $D(\mathbf{x}^1, \mathbf{x}^2) > 0$, as depicted in Figure 2-2. In the Figure, dashed lines represent unlabeled data points and solid lines labeled data points.

Under this interpretation the data distribution D segments the data into a set of connected components. Assuming view sufficiency, these components outline the different concept classes and as discussed in [11] the number of labeled samples needed

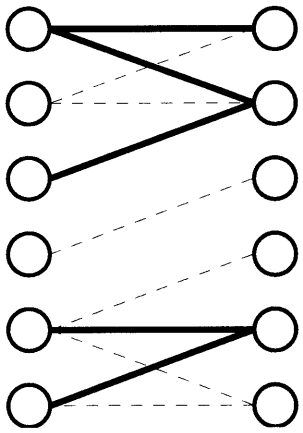


Figure 2-2: Bi-partite graph interpretation of co-training from [11], illustrated for a two view problem. In the graph, nodes are samples and each side of the graph is a different view. An edge exists between two samples if $D(\mathbf{x}^1, \mathbf{x}^2) > 0$. Labeled multi-view samples are displayed as solid edges. A concept class partitions the bi-partite graph into c components, with c the number of classes. The views are sufficient with respect to the concept class if the partitioning does not cut any edges in the graph (Figure taken from [11]).

given the unlabeled data is related to the number of connected components in this graph, which can be far less than $\|U\|$ and in the limit is one per connected component. An alternative, but equivalent viewpoint is that enforcing compatibility between concept classes constrains the set of possible solutions or labelings over U, since the set of solutions for which the views agree is a subset of all possible solutions and therefore learning can be performed from fewer labeled examples given unlabeled data with co-training. As discussed below, subsequent approaches explicitly optimize over the agreement between views and use this criterion to regularize the solution and learn from partially labeled data [98, 128, 51].

2.1.2 Related Methods

Co-training is an iterative semi-supervised learning approach in that it iteratively bootstraps a set of weak classifiers from confidently labeled examples on the unlabeled data. In doing so, it relies on having an accurate measure of classification confidence and greedily commits to a particular labeling of the data at each iteration, i.e., once

a sample is labeled it is added to the seed set and never revisited.

The Expectation Maximization (EM) algorithm provides an alternative optimization strategy for handling missing data [41]. In the semi-supervised setting EM consists of an E-step that computes class probabilities or "soft-label" assignments over the data and an M-step that re-learns the model parameters given the estimated probabilities. Nigham and Ghani [79] proposed *co-EM*, an extension of the Expectation Maximization (EM) algorithm to multiple views, where at each EM iteration the class probabilities from one view are used to train the classifier in the other view. Unlike traditional co-training, co-EM labels the entire unlabeled dataset at each iteration according to the current, improved model hypothesis and does not rely on a measure of classification confidence. Still, like other EM approaches, co-EM is sensitive to initialization and can converge to poor solutions.

Both conventional co-training and co-EM are iterative algorithms that implicitly maximize classifier agreement over the unlabeled data. An alternative approach is to use classifier agreement as an explicit form of regularization, i.e., the unlabeled data is used to regularize or constrain the set of possible solutions [26, 97, 98, 128]. The *co-boost* algorithm of Collins and Singer [26] was one of the first approaches to formulate co-training in this fashion, in which, a set of boosted classifiers are mutually trained by minimizing an objective that explicitly optimizes over classifier agreement. Sindhwani et. al. later proposed *co-regularization* [97, 98] that uses an agreement-based regularizer for regularized least squares and SVM classification. Similarly, Yu et. al. developed *Bayesian co-training* [128], a probabilistic co-training algorithm that utilizes the unlabeled data to form an agreement-based prior and regularize the solution. Regularization-based co-training methods are advantageous to iterative co-training techniques like co-EM in that they are less sensitive to view insufficiency or noise, since they use classifier agreement as more of a soft-constraint over the data, whose dominance, for example, can be controlled by a regularization parameter [97].

Multi-view manifold learning methods form a set of related techniques that use multiple views over the unlabeled data to learn a latent data representation [6, 51]. This representation is then used to learn a classifier from the labeled data samples.

In this way, similar to the regularization-based methods these techniques use the unlabeled data to simplify the learning problem and improve generalization. These methods also have improved theoretical properties with respect to the standard co-training assumptions. Examples of these techniques include the multi-view transfer learning approach of Ando and Zhang [6] and the Canonical Correlation Analysis (CCA) method proposed by Kakade and Foster [51] for performing multi-view regression.

The discussion thus far has focused on multi-view classification and regression. The use of multiple views has also been explored for the data clustering problem [31, 30, 9, 29]. Example domains include the clustering of text documents, where clusters over documents and over words are mutually informative [30], and clustering from both text and images. Similar to co-training, techniques in *co-clustering* exploit co-occurrence structure to reduce problem complexity and improve performance. In particular, these methods seek consistent clusterings across the different views such that the clusters are aligned with one another according to some agreement criterion. Sinkkonen et. al. [99] seek clusterings that maximize the mutual information between views. Similarly, Bickel et. al. [9] utilize co-EM to learn a mixture of Gaussians model and measure view consensus using a relative entropy measure. Dhilon [30] proposed a spectral co-clustering algorithm that finds clusters using a bipartite graph structure, which has been shown to be related to techniques in non-negative matrix factorization [32].

Similar to multi-view manifold learning methods, in the context of classification co-clustering approaches can be seen as finding informative representations of the data, useful for classification. Example approaches in this domain include manifold co-regularization of Sindhwani and Rosenberg [98] and regularized co-clustering with dual-supervision proposed by Sindhwani et. al. [98].

2.1.3 Multi-Modal Classification

For many classification problems the different views or feature splits are defined by separate input modalities, e.g., audio and video. There exists a large body of work on

multi-modal classification where information from multiple, potentially very different input sources are combined [56, 7, 105, 28, 76]. Kittler et al [56] provide a survey of various, commonly used classifier combination strategies from multi-modal inputs, and place them under a unifying framework. They show that many existing classifier combination schemes can be considered as special instances of compound classification over the joint feature space.

More recently, copula-based models have been proposed for performing multi-modal classification [28, 76]. These approaches use copulas to approximate the distribution of the joint feature space and fuse information across modalities. Similarly, multiple kernel learning approaches have been explored for multi-modal learning [7, 105]. These can be seen as a related class of early integration approaches that use kernel combination to combine information from multiple modalities and model the joint feature space. A comparison and evaluation of the various multi-modal integration strategies remains an interesting and open area of research that is not addressed as part of this thesis, although the correct fusion strategy is likely to be problem dependent. Instead we develop multi-view learning approaches built upon multiple kernel combination and demonstrate the benefit of this class of approaches for learning from multiple information sources.

Multi-modal classification is well suited for multi-view semi-supervised learning because each modality provides a potentially redundant view to the learning algorithm. While the concept of multi-modal co-training was mentioned as promising future work in the seminal Blum and Mitchell paper [11], it appears that there has been relatively little subsequent work on cross-modal co-training. Li and Ogihara [65] use a multi-view learning algorithm applied to gene expression and phylogenetic data to perform gene function classification. Yan and Naphade [123] apply co-training to video and transcribed speech for video category classification. Maeireizo et al [72] co-train a user emotion classifier from speech prosodic features and text.

In this thesis, we investigate the use of multi-modal co-training for learning audio-visual speech and gesture classifiers, and for multi-view object classification, and demonstrate that co-training can be successfully applied to this class of problems.

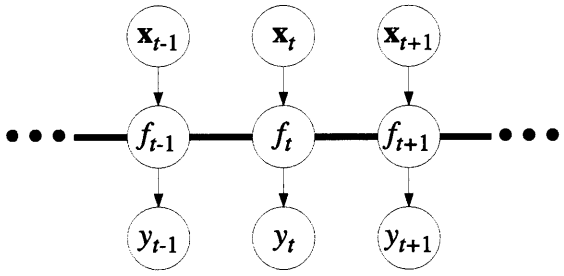


Figure 2-3: Graphical model for Gaussian Process (GP) regression and classification. $\mathbf{x}_t$ is the input observation, f_t is the latent mapping function, and y_t is the observed output at time index t. Thick bar indicates fully connected nodes. (Figure adopted from [86])

2.2 Gaussian Processes

A Gaussian Process (GP) is a stochastic process whose observations are jointly Gaussian and is the generalization of the multi-variate Gaussian distribution to samples obtained over time. In machine learning, GPs refer to a class of kernel-based techniques for performing regression and classification, and have close connections to other well known kernel machine methods such as the Support Vector Machine (SVM) and Relevance Vector Machine (RVM) [86]. They have been applied to a variety of problem areas including computer vision [52, 90, 111, 112, 116] and related problems in machine learning [59, 110, 57, 13, 40].

In what follows an overview of GP regression and classification is discussed both for the supervised and semi-supervised setting. A summary of the Bayesian co-training algorithm of Yu et. al. [128] which extends semi-supervised GP classification to multiple views is then provided.

2.2.1 GP Regression and Classification

The graphical model for GP regression and classification is provided in Figure 2-3. In the Figure[3], $\mathbf{x}_t$ is the input observation, f_t is the latent mapping function, and

[3]In this discussion we follow the notation of Rasmussen and Williams [86].

y_t is the observed output at time index t. A GP prior is assumed over the space of non-parametric functions $\mathbf{f}$,

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{2.3}$$

where $\mathbf{f} = (\cdots, f_{t-1}, f_t, f_{t+1}, \cdots)^T$ and the covariance, $\mathbf{K}$, is computed from the input space, $\mathbf{x}_t$.[4]

The class of functions modeled by the GP is determined by the chosen kernel or model *covariance function* $k(\cdot)$, where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ (e.g., the Radial Basis Function (RBF), also known as the Squared Exponential kernel, favors functions $\mathbf{f}$ that smoothly vary over time). The relationship between the output y and latent function f is task-dependent. For regression, y is modeled as a noisy version of f and the noise is typically modeled as additive Gaussian noise. Similarly, for classification the relationship between y and f is typically modeled with the logistic function, which maps f to a discrete valued output space.

GPs define a Bayesian formulation for regression and classification in which a GP prior is assumed over the space of non-parametric functions. Given a choice of covariance and output function, the parameters to the model are the hyper-parameters to these functions, e.g., assuming an RBF kernel and additive Gaussian noise these parameters include the length scale and kernel width of the RBF, and the output noise variance. Provided a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}$, $i = 1, \cdots, N$, the model hyper-parameters are learned using maximum likelihood. For regression this is done by maximizing the log marginal likelihood obtained by marginalizing out the latent f:

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_N^2 I| - \frac{N}{2}\log 2\pi, \tag{2.4}$$

where $\mathbf{y} = (y_1, \cdots, y_N)^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, and σ_N^2 is the output noise variance.

Given a new test point $\mathbf{x}_*$ inference is performed by computing the maximum a posteriori (MAP) estimate under the posterior distribution obtained by marginalizing f from the joint GP prior defined over f and f_*. Assuming an additive Gaussian

[4]In general $\mathbf{f}$ is infinite dimensional, however, as GPs are stationary they can be evaluated over a finite number of observations without having to consider the entire series, see [86] for a more detailed discussion.

noise model in the output, this can be done in closed-form to result in the *predictive distribution* for GP regression,

$$f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\bar{f}_*, \mathcal{V}[f_*]) \tag{2.5}$$

where f_* is the latent function defined over the test point $\mathbf{x}_*$ mean and variance are given by

$$\begin{aligned} \bar{f}_* &= \mathbf{k}_*^T(\mathbf{K} + \sigma_N^2 I)^{-1}\mathbf{y} \\ \mathcal{V}[f_*] &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_N^2 I)\mathbf{k}_* \end{aligned} \tag{2.6}$$

where $\mathbf{k}_* = (k(\mathbf{x}_1, \mathbf{x}_*), \cdots, k(\mathbf{x}_N, \mathbf{x}_*))^T$.

In Eq. 2.6, $\bar{f}_*$ is called the *mean prediction* and is the MAP estimate of the GP. Similarly, $\mathcal{V}(f_*)$ is the prediction uncertainty that can be used to gauge the reliability of the mean prediction estimate. The availability of an uncertainty over model predictions differentiates GPs from other kernel machines that are non-probabilistic, like the SVM, that is highly useful for many problem domains, e.g., see [52].

GP classification can be performed for the binary case by thresholding the output value to $\{-1, +1\}$, as is pursued in this thesis, although more sophisticated output functions can be used for classification, e.g., the logistic function. Under the logistic function learning and inference can no longer be carried out in closed form, however, variational approximations have been explored including the Laplace and Expectation Propagation (EP) methods for approximate inference, see [86] for details. In practice, the thresholded Gaussian and logistic models often perform similarly and the former model can be used for simplicity.

The computational complexity of GP regression and classification is dominated by the kernel inverse in the GP mean prediction equation, which is $\mathcal{O}(N^3)$. GPs have been shown to have good generalization properties and can often demonstrate good performance from small training datasets containing relatively few examples compared to other techniques [113]. For large datasets, however, learning and inference with GPs can become prohibitively expensive. GP *sparsification* techniques have been proposed to overcome this limitation and make learning and inference tractable

for large datasets [16, 4]. These techniques typically optimize over a set of latent input or inducing variables to find a representative subset of the larger dataset to formulate the GP [16]. Well known GP sparsification techniques include the FITC and PITC approaches proposed by Quinonero-Candela and Rasmussen [16]. Alvarez and Lawrence proposed a similar approach for multi-output GPs based on convolutional processes [4]. More recently, Lawrence and Urtasun [59] proposed a stochastic technique that exploits data sparsity to perform learning and inference over large datasets. They demonstrate their approach for performing collaborative filtering of user movie ratings from large web databases containing millions of examples.

Thus far we have considered supervised GPs that assume a fully labeled dataset, $\mathcal{D}$. There are many settings for which both labeled and unlabeled data are available. The extension of GPs to use partially labeled training datasets is non-trivial and several approaches have been explored for semi-supervised learning with GPs [134, 135, 58, 53, 95]. Lawrence and Jordan [58] develop a null-category model with GPs that utilizes the unlabeled data to steer the decision boundary away from high density regions of the data space. Zhu et. al. [134, 135] propose a graph-based regularization technique within a GP framework, in which similar points in input space are constrained to have the same labeling. Kapoor et. al. [53] explore a similar approach and propose an alternative algorithm for performing approximate inference based on the EP algorithm. Similarly, Sindhwani et. al. [95] exploit the data manifold to constrain the learning problem in the context of both labeled and unlabeled data.

Like other semi-supervised learning methods, the above methods utilize the unlabeled data as a form of regularization so as to constrain the space of solutions and simplify learning. These approaches optimize over partially labeled data within a single view. Yu et. al. [128] proposed a multi-view semi-supervised learning approach with GPs. We discuss this approach next.

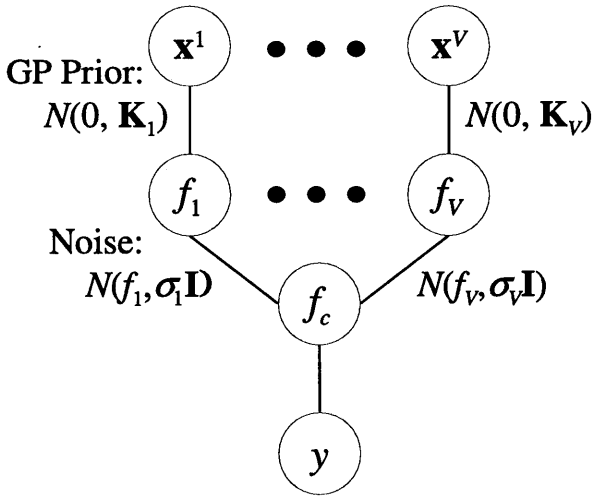


Figure 2-4: Graphical model for Bayesian co-training [128]. A latent predictor $\mathbf{f}_j$ is defined in each view, and a consensus latent variable, $\mathbf{f}_C$, models the agreement between the different views. The reliability of each view is modeled using a per-view noise variance σ_j^2.

2.2.2 Bayesian Co-Training

Yu et al. [128] proposed a probabilistic approach to co-training, called *Bayesian Co-training*, that combines multiple views in a principled way and generalizes previous approaches. 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 [86] on the latent variables

$$\mathbf{f}_j \sim \mathcal{N}(0, \mathbf{K}_j) \ , \tag{2.7}$$

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. The graphical model for Bayesian co-training is depicted in Figure 2-4.

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

$$p(\mathbf{y}, \mathbf{f}_c, \mathbf{f}_1, \cdots, \mathbf{f}_V) = \frac{1}{Z} \prod_{i=1}^{n} \psi(y_i, f_c(\mathbf{X}_i)) \prod_{j=1}^{m} \psi(\mathbf{f}_j) \psi(\mathbf{f}_j, \mathbf{f}_c) \tag{2.8}$$

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. (2.7) 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 [128].

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.

Integrating over the latent $\mathbf{f}_i$ results in a GP prior over the consensus function $\mathbf{f}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_c)$ with covariance function

$$\mathbf{K}_c = (\sum_j (K_j + \sigma_j^2 \mathbf{I})^{-1})^{-1} \tag{2.9}$$

The covariance function of Eq. 2.9 is called the *co-training kernel.* The resulting prior on $\mathbf{f}_c$ can be seen as an agreement-based prior that favors agreement between the views conditioned on σ_j^2. Using the co-training kernel standard GP regression and classification can be performed from multiple views. Unlike iterative-based co-

training techniques, Bayesian co-training jointly optimizes over all the views.

In [128], Yu et. al. demonstrate connections between Bayesian co-training and other existing techniques. Specifically, they show that marginalizing $\mathbf{f}_c$ in the Bayesian co-training model results in the co-regularization approach with least-squared loss for the regression problem,

$$p(\mathbf{y}, \mathbf{f}_1, \cdots, \mathbf{f}_m) = \frac{1}{Z} \exp\left\{ -\frac{1}{2}\sum_j \frac{\|\mathbf{f}_j - \mathbf{y}\|^2}{\sigma_j^2 + \sigma^2} - \frac{1}{2}\sum_{j=1}^{m} \mathbf{f}_j \mathbf{K}_j^{-1} \mathbf{f}_j - \frac{1}{2}\sum_{j<k} \frac{\|\mathbf{f}_j - \mathbf{f}_k\|^2}{\sigma_j^2 + \sigma_k^2} \right\}. \tag{2.10}$$

They also show that by marginalizing $\mathbf{f}_c$ and $\mathbf{f}_j$, $\forall j \neq k$ one obtains $\mathbf{f}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_k)$, with

$$\mathbf{C}_k = \left[\mathbf{K}_k^{-1} + \sum_{j \neq k} (\mathbf{K}_j + (\sigma_k^2 + \sigma_j^2)\mathbf{I})^{-1} \right]^{-1} \tag{2.11}$$

Yu et. al. refer to this result as individual view learning with side information and is another useful setting achievable within the Bayesian co-training model. Using Eq. 2.11 one can learn a classification function in view k while taking into account information from other views.

Another distinguishing characteristic of Bayesian co-training is the inclusion of a view reliability model within a principled, probabilistic framework. This is unique to other approaches that incorporate view reliability terms [123] in that the per view noise terms can be learned from partially labeled data using standard techniques such as maximum likelihood. Still, similar to other techniques Bayesian co-training assumes per view noise, and as we discuss in Chapter 5 such a model is limited in its ability to handle complex noise processes common to many real-world multi-sensor datasets, e.g., occlusions. In Chapter 5 we discuss how one can extend the Bayesian co-training model to incorporate a general noise model that can handle a variety of complex noise types.

2.3 Distributed Coding

Many multi-view learning and classification problems involve sensors that are not physically co-located, but instead are distributed across a wide sensing area. Examples in this domain are in *distributed networks*, such as [42, 3]. In distributed networks, it is often the case that communication between sensors is bandwidth constrained, as distributed transmission is required across possibly large distances, and that each sensor has limited computational power. In these systems, the question of efficient stream encoding is of central importance, and the development of compact signal representations is a critical part of multi-view classification algorithms.

The problem of finding compact feature representations from high-dimensional input signals is an active topic of research in both the fields of machine learning and information theory [82, 33, 27]. In machine learning, techniques in *feature selection* have been explored for finding maximally discriminative representations for a given classification task and finding compact feature representations that are minimally redundant [67, 82, 33]. Similarly, *data compression* methods in information theory and signal processing exploit redundancy in the data to formulate efficient feature encodings useful for efficient data transmission and storage [27]. Work in feature selection and data compression offer very different perspectives of a common underlying problem, whose findings can be mutually beneficial for deriving efficient data representations.

In a distributed network the sensors are typically bandwidth limited and communication between sensors is prohibitively expensive. A naive feature encoding algorithm would compress and transmit the signals from each sensor separately, however, a much more efficient encoding could be achieved by jointly compressing the views such that both dependencies within and across views are exploited. As sharing between sensors is prohibitively expensive, however, a joint encoding of the signals is not possible using standard feature selection and data compression techniques. Surprisingly, it has been shown that assuming a common receiver, encoding rates close to the joint entropy can be obtained even *without* explicitly sharing information between views. This re-

sult is known in the information theory community as the Slepian-Wolf theorem [103] and subsequently many distributed coding techniques have been pursued for the joint compression of signals transmitted on a distributed network [84, 122, 42, 93].

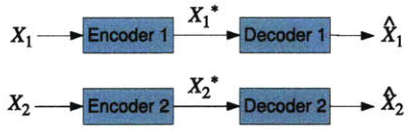
Distributed coding has been applied to a variety of areas including distributed image and video coding [70, 42]. Until recently [127], however, distributed coding approaches have not been explored for multi-camera computer vision systems. In Chapter 7, we demonstrate how distributed coding can be used to achieve an efficient feature selection algorithm for multi-view object recognition on a distributed network and propose a novel distributed coding approach with multi-view GPs. In the remainder of this section an overview of the Slepian-Wolf theorem is first provided, followed by a short survey of contemporary distributed coding algorithms.

2.3.1 Slepian-Wolf Theorem

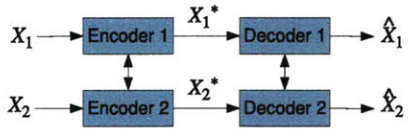
Let $X = (X_1, X_2)$ be a multi-view source signal that we wish to transmit over a noiseless channel to a central receiver[5]. Figure 2-5 depicts three different coding schemes that can be employed to compress the sources X_1 and X_2. Let R_1 and R_2 be the encoding rates of X_1 and X_2 respectively. Treating X_1 and X_2 as random variables drawn from the distribution $p(X_1, X_2)$, from information theory we know that under the independent coding scheme we have, $R_1 \geq H(X_1)$ and $R_2 \geq H(X_2)$, where $H(x)$ is the *entropy* of the random variable x [27]. Similarly, for the joint coding scheme, where both sources are available to each encoder and decoder, we have $R_1 + R_2 \geq H(X_1, X_2)$ and for correlated sources X_1, X_2, $H(X_1) + H(X_2) > H(X_1, X_2)$, i.e., a more efficient coding is achievable by a joint coding of correlated sources as one would expect.

The final coding scheme is referred to as *distributed coding*, and is that typically encountered in distributed networks, where the sources are both available to the decoder, but are not shared between encoders, e.g., as a result of limited network bandwidth. Is clear that under the distributed coding scheme we can have $R_1 + R_2 \geq H(X_1) + H(X_2)$ as with the independent and joint coding approaches. The question

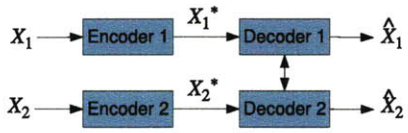
[5]We present distributed coding under the noiseless channel model following Slepian-Wolf [103]. For a discussion of coding with noisy channels see [27].



(a) Independent Source Coding



(b) Joint Source Coding



(c) Distributed Source Coding

Figure 2-5: Three different source coding schemes for compression of two noiseless sources X_1, X_2: (a) independent source coding, (b) joint source coding, and (c) distributed source coding. X_1^*,X_2^* are the encoded signals, and $\hat{X}_1$, $\hat{X}_2$ the decoded signals.

remains, however, as both sources are shared between decoders, can one do better? Surprisingly, even *without* sharing information across encoders, the answer is yes, and in fact one can achieve $R_1 + R_2 \geq H(X_1, X_2)$ with correlated sources for the distributed coding setting as was shown by Slepian and Wolf [103].

In [103], Slepian and Wolf generalize well-known information theoretic results on single source coding of a discrete noiseless signal to two correlated sources, where they consider various connection topologies between the different encoders and decoders. A central finding of their paper is the Slepian-Wolf distributed coding theorem [27], which states that for two distributed sources X_1, X_2 with a common decoder the

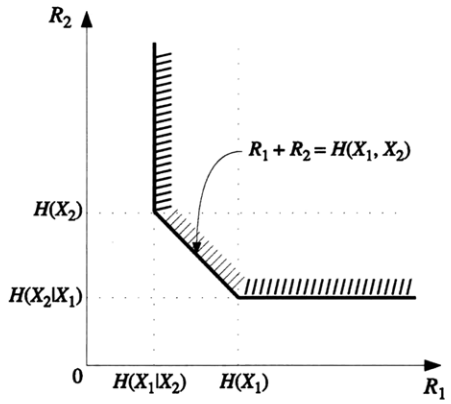


Figure 2-6: Achievable rate region for Slepian-Wolf distributed source coding.

achievable rate region is given by,

$$\begin{aligned} R_1 &\geq H(X_1|X_2), \\ R_2 &\geq H(X_2|X_1), \\ R_1 + R_2 &\geq H(X_1, X_2), \end{aligned} \tag{2.12}$$

which is also graphically displayed in Figure 2-6.

As an illustration of the Slepian-Wolf theorem consider the conceptual example depicted in Figure 2-7. The example consists of two sensors, one for temperature and another for rain, each represented using two bits of information. Figure 2-7 displays the sensor joint probability distribution as a correlation table. Summing over rows and/or columns of this table it is clear that a naive, independent coding of the sensor signals would require all two bits, since all bit combinations have non-zero marginal probability. Now, assume that instead we adopt a distributed coding scheme, in which the decoder first observes temperature. Note that conditioned on temperature, the conditional distribution over rain has only two entries with non-zero probability. Thus having knowledge of temperature and a model of the joint sensor statistics the decoder can reconstruct rain using only a single bit, even though the rain encoder has no knowledge of temperature.

Slepian and Wolf prove their theorem by considering the jointly typical sets over

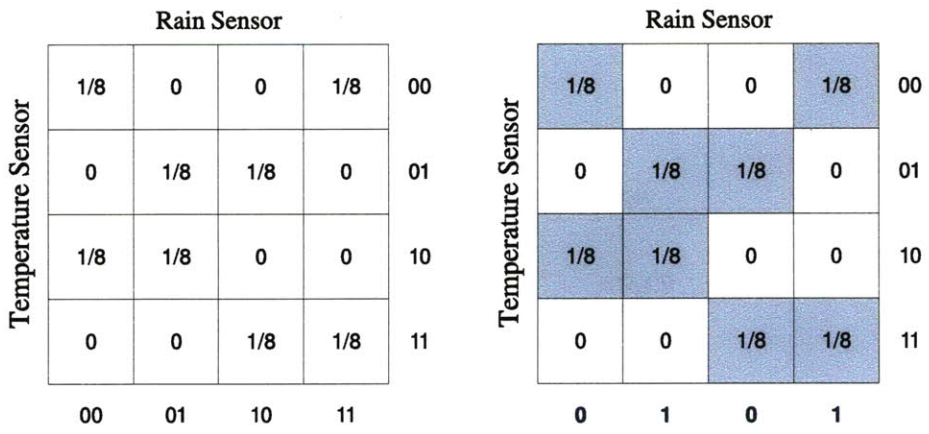


Figure 2-7: Conceptual example of Slepian-Wolf distributed coding. The joint distribution of two sensors, temperature and rain is shown as a correlation table. Independent coding of each view requires the full two bits per view. Having knowledge of temperature and a model of the joint sensor statistics the decoder can reconstruct rain using only a single bit, even though the rain encoder has no knowledge of temperature.

each source[6], for details see [103]. Cover and Thomas provide an informal explanation of Slepian-Wolf through graph coloring. Consider the two sources and their distributions as depicted in Figure 2-8. In the Figure, each dot is an instance of the source signal and the sample space of each source is shown as an oval. In this example, the decoder uses source X_1 to decode X_2. Imagine randomly coloring the instances of X_2 with 2^{R_2} colors, where R_2 is the encoding rate of X_2. If $R_2 \geq H(X_2|X_1)$ conditioning on X_1 would likely result in a distribution over X_2 containing $2^{H(X_2|X_1)}$ colors. Thus using an encoding rate of $R_2 = H(X_2|X_1)$ for X_2, the decoder can uniquely identify the correct value of X_2 given X_1 with high probability, even though the X_2 encoder has no knowledge of X_1. This point is illustrated graphically in Figure 2-8 by extending lines from an example X_1 to jointly typical examples in X_2 that have $p(X_2|X_1) \gg 0$.

[6]Although Slepian and Wolf only proved the theorem for two sources, it is easily extendible to multiple sources, see [27].

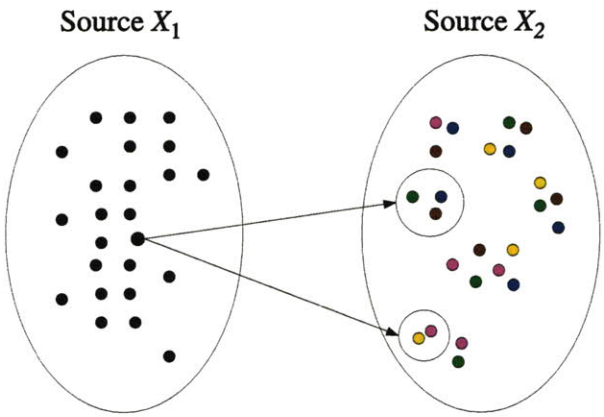


Figure 2-8: Illustration of Slepian-Wolf distributed source coding using graph coloring. A two view problem is shown. For each view, example instances are drawn as dots and the sample space of each view is shown as an oval. Consider coloring the examples in X_2 with R_2 colors. Setting $R_2 \geq H(X_2|X_1)$ conditioning on X_1 would likely result in a distribution over X_2 containing $2^{H(X_2|X_1)}$ colors, and the decoder can uniquely identify the correct value of X_2 given X_1 with high probability, even though the X_2 encoder has no knowledge of X_1.

The Slepian-Wolf theorem is an impressive result that has inspired the field of distributed coding. As is often the case, however, there is a large gap between practice and theory and the design of practical distributed coding algorithms that can achieve coding rates close to those supported by the Slepian-Wolf theorem is non-trivial and is an active area of research. In the following subsection, we provide a brief overview of some of the algorithms that have been developed in this field. For a more detailed survey of recent distributed coding methods see [122, 93].

2.3.2 Related Methods

One of the first practical distributed source coding algorithms was proposed by Wyner [120] using linear block codes that borrowed ideas from channel coding[7] Wyner also

[7]Channel coding methods utilize models of joint statistics between two sources (clean and noisy) in an analogous fashion to distributed source coding to decode from a noisy channel, for details see [27]. Many of the distributed coding methods to-date have borrowed ideas from channel coding [84, 93, 24, 106].

extended the Slepian-Wolf theorem to the lossy compression of continuous sources and showed that the achievable rate distortion for distributed source coding is analogous to that obtainable if sources are shared at each encoder [119]. Following Wyner and Ziv other *coset* based techniques have also been developed such as the DIstributed Source Coding Using Syndroms (DISCUS) method of Pradhan and Ramchandran [84] and others [93]. These methods exploit data cosets or syndromes modelled as linear channel codes to model jointly typical sets between sources.

More sophisticated linear block coding methods based on Low-Density Parity Check Codes (LDPCs) [39, 71] for distributed source coding were proposed by Schronberg et. al. [93] and others [70, 68, 91]. LDPCs define probabilistic decoding algorithms that find the most likely decoding of an input sequence given a set of parity constraints. An LDPC is typically specified as a factor graph and decoding is performed via loopy belief propagation [93]. For the channel coding problem they have shown to give good empirical performance close to the Shannon limit [71] and have also proven advantageous for distributed source coding [68, 91, 93]. Similarly to coset codes and LDPCs, Turbo coding techniques have also been adopted from the channel coding literature have for performing distributed coding [1, 19, 69, 132].

Most distributed source coding techniques, including the ones discussed thus far, are distributed coding with side information algorithms, i.e., they function at the corner points of the Slepian-Wolf rate region (recall Figure 2-6). It has been shown, however, that rate splitting methods can be used to achieve source rates at arbitrary rate points in the Slepian-Wolf rate region [118, 87, 129, 25]. One such technique is that of Coleman et. al. [25] that demonstrate rate-splitting with LDPC codes.

Although the distributed coding field has seen a fair amount of activity both in terms of theoretical and algorithmic development, distributed coding has seen limited application, and has been applied to relatively few application domains including distributed image and video coding [69, 42]. An application area for which distributed coding can have a large benefit is multi-camera object recognition in surveillance and scene understanding systems. With few exceptions, however, such as [127, 126], distributed coding has been relatively un-explored for this application domain. In

Chapter 7 we investigate the use of distributed coding for the coding of visual features in a multi-view object recognition scenario. We develop a novel distributed coding algorithm built upon GPs and demonstrate its performance on a publicly available multi-view object database.

Chapter 3

Co-Adaptation of Audio-Visual Speech and Gesture Classifiers

Multi-modal classification is well suited for multi-view learning because each modality provides a potentially redundant view to the learning algorithm[1]. While the concept of multi-modal co-training was mentioned as promising future work in the seminal Blum and Mitchell paper [11], it appears that there has been little subsequent work on cross-modal co-training. In this chapter, we investigate the use of multi-modal co-training for learning speech and gesture classifiers. To our knowledge, this is the first work to use co-training in the context of audio-visual speech and gesture. We also investigate the application of co-training for performing model adaptation and develop a co-adaptation approach that adapts a generic model to a specific user or environment with co-training. The application of co-training to multi-modal classification problems in this chapter leads to incites regarding the limitations of co-training when applied to multi-sensor domains with regard to satisfying view sufficiency. In subsequent chapters we address these limitations and develop multi-view learning algorithms that unlike existing state-of-the-art approaches can learn from challenging, noisy multi-sensor data.

[1]The work described in this chapter is published in the International Conference on Multimodal Interfaces, 2006, Christoudias, Saenko, Morency and Darrell [20].

3.1 Introduction

Human interaction relies on multiple redundant modalities to robustly convey information. Similarly, many human-computer interface (HCI) systems use multiple modes of input and output to increase robustness in the presence of noise (e.g. by performing audio-visual speech recognition) and to improve the naturalness of the interaction (e.g. by allowing gesture input in addition to speech). Such systems often employ classifiers based on supervised learning methods which require manually labeled data. However, obtaining large amounts of labeled data is costly, especially for systems that must handle multiple users and realistic (noisy) environments. In this chapter, we address the issue of learning multi-modal classifiers in a semi-supervised manner. We present a method that improves the performance of existing classifiers on new users and noise conditions without requiring any additional labeled data.

In the first part of the chapter, we explore co-training for two audio-visual tasks: speech unit classification and user agreement detection. The first task is to identify a sequence of acoustic and lip image features as a particular word or phoneme. The second task is to determine whether a user has expressed agreement or disagreement during a conversation, given a sequence of head gesture and acoustic features. Although we only deal with isolated sequences, the algorithm can be extended to continuous recognition. As the core classifier in the co-training paradigm, we use the hidden Markov model (HMM), which is common for speech and gesture sequence classification.

Co-training was originally proposed for the scenario in which labeled data is scarce but unlabeled data is easy to collect. In multi-modal HCI development, it may be feasible to collect enough labeled data from a set of users in a certain environment, but the resulting system may not generalize well to new users or environments. For example, a new user may gesture differently, or the room may become noisy when a fan is turned on. The semi-supervised learning problem then becomes one of adapting existing models to the particular condition. To solve this problem, we investigate a variant of co-training, which we call *co-adaptation*. Co-adaptation uses a generic su-

pervised classifier to produce an initial labeled training set for the new condition, from which a data-specific classifier is built. The algorithm then improves the resulting data-specific classifier with co-training, using the remaining unlabeled samples.

The development of user-adaptive multi-modal interfaces is a growing area of research. Adaptation to a user's multi-modal discourse patterns is known to be important, as users exhibit different interaction styles based on factors such as age and environment [121]. While we focus on improving the accuracy of low-level appearance, motion, and acoustic models, we believe our approach will also be useful in adapting timing and fusion parameters. A different approach to multi-modal adaptation is to design a system where the user adapts to the system's recognition capabilities while the system attempts to simultaneously adapt to the user [81]. In the context of audio-visual HMMs, maximum likelihood linear regression (MLLR) has been recently used for speaker adaptation [49]. Semi-supervised recognition of agreement and disagreement in meeting data using prosodic and word-based features was proposed in Hillard, Ostendorf, and Shriberg [48].

The remainder of this chapter is organized as follows. In the following section co-training is described in the context of audio-visual classification in Section 3.2. Our co-adaptation algorithm is presented in Section 3.3. Experiments and results are described in Section 3.4. Finally, a summary and a discussion of future work are given in Section 3.5.

3.2 Audio-Visual Co-training

In this chapter we consider the iterative co-training algorithm of Blum and Mitchell [11] described as Algorithm 1 of Chapter 2. As discussed in Chapter 2 the intuition behind the co-training algorithm is that classifiers operating on independent views of the data can help train each other by sharing their most confident labels. Its success of the algorithm depends on two assumptions: the conditional independence of the views, and the sufficiency of each view to learn the target function.

Although co-training has been applied to natural language [26] and other single-

modality tasks (e.g. [63]), it is unclear whether the assumptions required for its success will hold in the case of multi-modal HCI problems. We will now discuss what makes these problems different and how it may affect the training algorithm.

Co-training exploits the redundancy in the disjoint sets of features used to identify categories. Such redundancy is, in fact, what makes multi-modal tasks seem so well-suited to co-training: The spoken utterance "yes" and a head nod are redundant indications of user agreement; facial appearance and voice both convey user identity, etc. However, the assumption that each modality is sufficient for classification does not always hold. For example, the user can indicate agreement just by nodding and not providing any spoken feedback, or by nodding while saying something that does not explicitly state agreement. Another issue related to sufficiency is that the observations belonging to a particular category may not be aligned in time across modalities and may have variable-length segmentations. In this chapter, we make sure that for each segmented time period, each view in the training data is sufficient to identify the correct class. We present co-training algorithms that relax this assumption in subsequent chapters.

The other assumption made by the co-training paradigm is that of class-conditional independence of views. This seems like a reasonable assumption in the case of multiple modalities. In fact, the same assumption is made by many multi-modal fusion models which express the class-conditional likelihood of a multi-modal observation as the product of the observation likelihoods for each modality.

Finally, the original formulation of the co-training algorithm [11] relies on weak classifiers trained on a small quantity of labeled data to provide new labels at each iteration. To ensure that the quality of the labeled data does not deteriorate, the classifiers need to either have a low false positive rate, or reliable confidence estimates. While this may be possible for text classification tasks, it is harder to achieve for noisy multi-modal observations. In our formulation, which uses HMM classifiers, we compute confidence values as follows. Let x_i be an observation in modality i, and y

Algorithm 2 Co-Adaptation Algorithm

Given user-independent classifiers f_i^{UI}, $i = 1, ...k$, a user-dependent unlabeled set U and parameters N, M and T:
set $S = \emptyset$
for $i = 1$ to k **do**
  Use f_i^{UI} to label the M highest-confidence samples in U and move them to S
end for
Set $t = 1$
repeat
  for $i = 1$ to k **do**
    Train user-dependent classifier f_i on view i of S
    Use f_i to label N highest-confidence samples in U and move them to S
  end for
  Set $t = t + 1$
until $t = T$ or $|U| = 0$

be one of $1, ..., n$ labels. Then the posterior probability of y given x_i is

$$P(y|x_i) = \frac{P(x_i|y)P(y)}{\sum_{u=1}^{n} P(x_i|u)P(u)} \tag{3.1}$$

where the likelihood of x_i given the label is obtained from the HMM classifier f_i for each class. We use the posterior probability computed in (3.1) as the confidence value to assess the reliability of labels assigned to the unlabeled samples during co-training.

3.3 Co-Adaptation Algorithm

Co-training was proposed for the scenario where labeled data is scarce but unlabeled data is easy to collect. In certain multi-modal HCI applications, it may be feasible to collect a lot of labeled data to train a model on a particular set of users and environmental conditions (audio noise level, lighting, sensing equipment, etc.) However, such a model may not generalize well to new users and conditions.

To address this issue, we propose an adaptive version of the co-training algorithm that bootstraps a data-dependent model from a data-independent model trained on a large labeled dataset. Suppose we obtain unlabeled data from a new condition, such as a new user. We first use the user-independent model to specify a small seed set

of labeled examples using its most confident predictions. A user-dependent model is then trained on this initial seed set and improved with cross-modal learning on the rest of the unlabeled data. The resulting co-adaptation algorithm is summarized as Algorithm 2. The parameters to the algorithm are M, the number of examples added by each user-independent classifier to form the initial labeled set S, N the number of examples labeled by each user-specific classifier during each co-training iteration, and T the total number of iterations. The algorithm terminates when either all the data has been labeled or T iterations is reached.

The intuition behind the co-adaptation algorithm is that, while the overall performance of the generic model may be poor on new users or under new noise conditions, it can still be used to accurately label a small seed set of examples. The initial seed classifier can then be improved via co-training. Since the new classifier is trained using samples from the new working condition (i.e., new user and environment), it has the potential to out-perform the original generic classifier in the new setting, especially when user variation or difference in environment is large.

Note that, in Algorithm 2, a new user-dependent model is trained on the unlabeled data instead of adding the new labels to the user-independent labeled set. The advantage of this approach is that it is better suited to situations where there is a large imbalance between the amount of labeled and unlabeled data. Alternatively, we could use the new labels to adapt the parameters of the existing model using an HMM adaptation technique such as maximum likelihood linear regression (MLLR)[49]. The advantage of training a separate user-dependent model is that it enables us to use data-dependent features. For example, we can train a new model with higher-resolution visual observations, or apply data-dependent principal component analysis (PCA). We leave this as a future work direction.

3.4 Experiments

To evaluate our co-training framework, we apply it to two different multi-modal tasks: speech unit classification and agreement recognition in human-computer dialogue.

Both tasks exploit the audio and the visual modalities, and are typical examples of HCI applications.

In all experiments, we use correct classification rate (CCR) as the evaluation metric, defined as

$$\text{CCR} = \frac{\#\text{ sequences correctly classified}}{\text{total }\#\text{ of sequences}}.$$

We compare the co-adaptation algorithm to two other semi-supervised methods [34]. The first method uses the top N most confidently classified examples from one modality to train a classifier in the other modality. As we show in our experiments, this method is only beneficial when the relative performance of the classifiers on the unlabeled data is known a priori, so that stronger classifiers can be used to improve weaker ones. We show that co-adaptation can achieve the same or better improvements in performance without the need for such prior knowledge.

The second baseline we consider is single-modality bootstrapping, which does not use cross-modal learning, but rather learns a semi-supervised classifier separately in each modality. It is similar to co-adaptation (Algorithm 2), except that each classifier operates on its own copy of U and S, and classification labels are not shared across modalities. As we demonstrate in our experiments, cross-modal learning algorithms are better at improving weak classifiers than single-modality bootstrapping, especially when one modality is more reliable than the other.

In the following experiments, we use left-to-right HMMs with a mixture of Gaussians observation model.

3.4.1 Audio-Visual Agreement Recognition

In this section, we apply multi-modal co-training to the task of recognizing user agreement during multi-modal interaction with a conversational agent. In this setting, the user interacts with an agent using speech and head gestures. The agent uses recognized head nods (or head shakes) and agreement utterances in the user's speech to determine when the user is in agreement (or disagreement). In unconstrained speech,

Classifier	Seed	Co-training	Oracle
Audio	88.4 ± 9.9	91.7 ± 9.2 (p=0.03)	95.1 ± 5.4 (p<0.01)
Visual	95.5 ± 4.4	96.8 ± 3.6 (p=0.07)	97.5 ± 2.8 (p<0.01)

Table 3.1: Co-training of multi-modal agreement classifiers. Each column shows the mean CCR over 15 test subjects, ± the standard deviation. The p-value comparing the performance of the seed and co-trained classifiers, and the seed and oracle classifiers is also displayed.

there are a variety of utterances that can signify agreement, making recognition of agreement difficult with user-independent classifiers, as agreement utterances may vary per user. In this chapter, we focus on classifying "yes" and "no" utterances and nod and shake head gestures, and seek to improve these classifiers using unlabeled data.

Dataset

For our experiments on agreement recognition, we collected a database of 15 subjects interacting with a virtual avatar. In each interaction, the subject was presented with a set of 103 yes/no questions and was asked to respond with simultaneous speech and head gesture, and to use only "yes" and "no" responses along with head nods and shakes. Each interaction was recorded with a monocular video camera and lasted 10-12 minutes. A log file with the start and end times of each spoken utterance from the avatar was kept. During each interaction, a remote keyboard was used by the experimenter to trigger the dialogue manager after each subject's response. The end times of the subject's answers were also logged. The video sequences were then post-processed using the avatar's log file to extract the responses of each subject. The sequences were manually labeled to identify positive and negative responses and answers where subjects used extraneous speech or did not speak and gesture at the same time were discarded. To keep the responses to roughly the same length, any responses longer than 6 seconds were also discarded. The resulting data set consisted of 1468 agreement and disagreement audio-visual sequences.

To extract features for the visual classifiers we used a modified version of the 6-degree of freedom head tracker in [74] modified to perform monocular tracking.

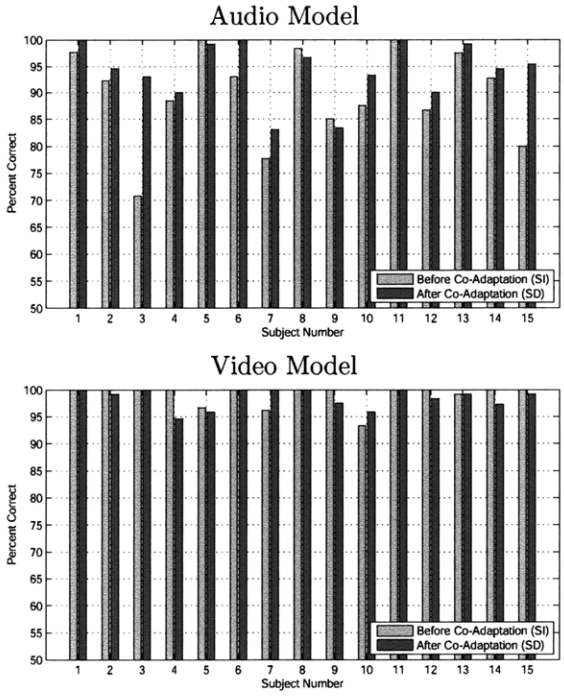


Figure 3-1: Detailed results for co-adaptation of multi-modal agreement classifiers (summarized in Table 3.2). The CCR rate of the user-dependent and user-independent classifiers are shown for each of the 15 test subjects. The light bars show the CCR of the user-independent classifiers and the dark bars show the CCR of the user-dependent classifiers found with co-adaptation.

This tracker was used to compute 3D head rotation velocities for each subject. For each answer segment we applied a 2-second, 64-sample, windowed fast-Fourier transform (FFT) to the x, y and z head rotation velocities computed at 0.1 second intervals within the segment. The x, y and z frequency responses at each time window were then concatenated into a single 99-dimensional observation vector. For the audio agreement classifiers, we used 13 dimensional Mel-frequency cepstral coefficients (MFCCs) computed at 100Hz from the audio of each answer segment.

Results

In this section we present our experiments on co-training audio and visual agreement classifiers. We first present results on co-training and then demonstrate our co-adaptation technique.

Classifier	User Independent	Co-Adaptation	Single-Modality Bootstrap
Audio	89.8 ± 8.8	94.2 ± 5.6 (p=0.023)	91.3 ± 8.7 (p=0.414)
Visual	99.0 ± 2.0	98.5 ± 1.8 (p=0.332)	98.5 ± 2.3 (p=0.411)

Table 3.2: Co-adaptation of multi-modal agreement classifiers. Each column shows the mean CCR over the 15 test subjects, ± the standard deviation. The p-value comparing the performance of each method to that of the user-independent model is also shown.

To begin we evaluate co-training for the construction of a user-dependent agreement classifier from a few labeled examples. For this experiment, we use Gaussian audio and visual classifiers (1-state HMMs with 1 mixture component). We evaluated the co-training algorithm using leave-n out cross-validation on each subject, where we split the data of each subject into 90 percent train and 10 percent test for each round of cross-validation. At each round the training data is split into an unlabeled training set and a labeled seed set of 3 positive and 3 negative examples. To remove bias due to a particular choice of seed set or unlabeled train and test set, co-training was evaluated over three cross-validation trials for each subject where the seed set as well as unlabeled train and test sets were chosen at random.

Table 3.1 displays the result of the agreement co-training experiment with $N = 4$ and iterating until all the unlabeled training data is labeled with co-training (see Algorithm 1). The table displays the average classification accuracy, averaged across all 15 subjects and three trials. In the table the performance of the co-trained audio and visual Gaussian classifiers are also compared to oracle performance, obtained by training the audio and visual agreement classifiers using a fully supervised paradigm, i.e., with ground truth labels on all the training data, and evaluating these classifiers on the test set. The table also gives the p-values of the difference in classifier performance before and after co-training computed using statistical t-tests. Through co-training we were able to increase overall performance of the audio classifier by 3.3 percent with a p-value of p=0.03, meaning that this increase is statistically significant. Similarly, we were able to gain a marginally significant increase in the performance of the visual classifier by 1.3 percent with a p-value of p=0.07.

Next we evaluate our co-adaptation algorithm. For this experiment, we used 5-

state HMMs with 2 mixture components, and ran our co-adaptation algorithm with $M = 4$, $N = 4$ and 3 iterations. We performed leave-one out experiments where we trained user-independent audio and visual classifiers on 14 out of the 15 subjects in our dataset and ran co-adaption on the left out subject. For each subject we ran co-adaptation on random splits of the data, 90 percent train and 10 percent test, and averaged the results over 10 trials. Figure 3-1 displays the classification accuracy of the user-independent and user-dependent audio and visual classifiers obtained with co-adaptation. The user-dependent HMM classifiers obtained with co-adaptation either matched or improved performance over the user-independent classifiers. As was the case in our previous experiment the main improvement of co-adaptation is seen in the audio modality as the user-independent visual classifiers are already performing quite well on each subject.

Table 3.2 displays the average classification accuracy of the user-independent and user-dependent classifiers obtained with co-adaptation, averaged over the 15 subjects. The user-dependent audio classifiers obtained with co-adaption do significantly better than the user-independent models, with an average improvement of 4.4 percent and a p-value of 0.023. In Table 3.2 we also compare our co-adaptation algorithm to single-modality bootstrapping with $M = 10$, $N = 10$ and 3 iterations, and found that unlike our approach the difference in performance between the user-independent and user-dependent audio HMM classifiers obtained with single-modality bootstrapping was not significant (p-value equal to 0.414). This is because co-adaptation, unlike single modality bootstrapping, was able to leverage the good performance of the visual classifiers to significantly improve the performance of the audio agreement classifier.

3.4.2 Audio-Visual Speech Classification

Audio-visual speech unit classification uses acoustic features extracted from the speech waveform and image features extracted from the speaker's lip region. It has been widely reported that visual input helps automatic speech recognition in the presence of acoustic noise (e.g. [47]). However, while recording audio-visual speech data is becoming easier, annotating it is still time-consuming. Therefore, we would like to

see whether co-training can help exploit unlabeled data for this task.

To satisfy the sufficiency assumption, it should be possible to distinguish between the speech units using only lipreading. This is possible if, for example, the units are digits recognized as whole words: "one", "two", etc. In this chapter, we evaluate our algorithm on the task of phoneme unit classification. To ensure sufficiency, we clustered several phonemes together so that the resulting "visemes" are visually distinguishable:

1: b, p, m, f, v

2: w, uw, oy, ao, ow, r

3: sh, zh, ch, jh, s, z

4: ae, aw, ay, ey, aa

Dataset

For evaluation, we used a subset of the multi-speaker audio-visual database of continuous English speech called AVTIMIT [47]. The database contains synchronized audio and video of 235 speakers reading phonetically balanced TIMIT sentences in a quiet office environment. There are 15 sentences per speaker, so the number of sequences in the dataset is between 20 and 60 per viseme, per speaker. To simulate noisy acoustic conditions, speech babble noise was added to the clean audio to achieve a 0 db signal-to-noise ratio. The result is similar to a noisy public place, such as a busy coffee shop. The database contained phonetic transcriptions produced by forced alignment, which we converted to viseme labels via the mapping shown in the previous section. Since the original database was labeled, we simulated unlabeled data sets omitting the labels.

For each label, the data sample consisted of a sequence of acoustic observations and a corresponding sequence of visual observations. The 42-dimensional acoustic feature vector, sampled at 100 Hz, contained 14 mel-frequency cepstral coefficients (MFCCs), their derivatives and double derivatives. Visual features were extracted

Classifier	Supervised	Co-training	Single-modality BS	Oracle
Audio	59.1 ± 5.6	67.0 ± 9.1 (p<<.01)	60.9 ± 7.7 (p=.10)	94.0 ± 1.1
Video	56.8 ± 10.5	66.2 ± 10.2 (p<<.01)	54.8 ± 12.2 (p=.10)	73.3 ± 4.5

Table 3.3: Co-training results on the speech dataset. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. "Supervised" refers to the seed classifier performance. In parentheses, we show p-values for co-training and the single-modality bootstrap baseline relative to the supervised classifier.

from a 32-by-32 region centered on the lips, and consisted of an 8-by-8 sub-grid of the discrete cosine transform (DCT) followed by a PCA transform to further reduce the dimensionality to 30 coefficients.

Results

In all of the following experiments, the number of HMM states was set such that the average sequence contained three frames per state, resulting in 3-4 states for the audio HMM and 1 state for the visual HMM. The number of Gaussian mixture components was set such that there was a minimum number of training samples per dimension for each component, up to a maximum of 20 components.

Again, our first goal is to show that we can improve classifiers that are poorly trained because of the lack of labeled training data by co-training them on unlabeled data. We thus look for the case when the amount of labeled data is too small, i.e., when adding more training data reduces the test error rate. For the speech dataset, this happens when the labeled set L contains 4 sample sequences per class. First, we train the supervised HMM classifiers on a randomly chosen L for each user, and test them on the remaining sequences. The results, averaged over all users, are shown in the first column of Table 3.3. Next, we co-train these initial classifiers, using N=4, M=2 and 9 iterations (after which the unlabeled set became depleted.) The results, in the second column of Table 3.3, show that co-training is able to significantly improve the performance in each modality, unlike single-modality bootstrapping (shown in the third column). For reference, the last column shows oracle performance, or what we would get if all of the labels added by co-training were correct. Note that, while the co-trained video classifier is approaching oracle performance, the audio is still far

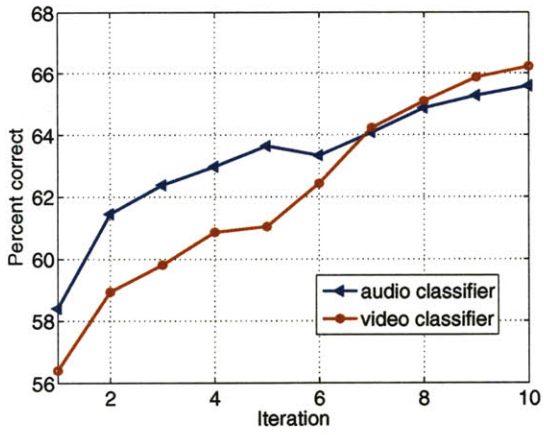


Figure 3-2: Learning rate of the co-training algorithm on the speech dataset. The plot shows the CCR after each iteration for the audio and video classifiers. The first iteration corresponds to the CCR of the seed classifier.

below that level. However, this dataset did not contain a lot of data per speaker. Perhaps, if more unlabeled data were available, the performance would continue to increase, following the trend shown in Figure 3-2.

Our second goal is to use our adaptive CT algorithm to improve existing user-independent (UI) models when new, unlabeled data becomes available. We train the initial UI audio and visual HMM classifiers on a large labeled dataset consisting of 50 users and approximately 20K samples. Then, for each of the users in the unlabeled dataset, we perform co-adaptation as described in Section 3.3, using N=25% of all samples, M=2, and 7 iterations. The UI and the final user-dependent (UD) co-trained classifiers are then tested on all of the data for each user.

First, we evaluate the case where the audio noise level in the labeled data matches the noise level in the unlabeled data. In this case, we are mostly adapting to the user. The results are shown in Table 3.4. The first observation is that the UI video classifier does not do much better than the UD supervised classifier (first column of Table 3.3.) Our co-adaptation algorithm improves the visual performance significantly, while the audio performance stays the same. One explanation for this is that audio is helping the video as the stronger of the two modalities. Therefore, we compare this to bootstrapping from the stronger audio modality (see "Audio-Bootstrap" in Table

Classifier	User Indep.	Co-Adaptation	Audio-BS	Video-BS
Audio	72.6 ± 4.5	72.0 ± 4.4 (p=.36)	70.2 ± 4.2 (p<<.01)	63.3 ± 11.8
Video	59.8 ± 11.3	68.1 ± 9.7 (p<<.01)	70.1 ± 6.2 (p<<.01)	62.4 ± 13.2

Table 3.4: User-adaptive co-training results on the speech data, matched labeled and unlabeled audio noise conditions. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. p-values are relative to the UI classifier.

3.4), and see that it has similar results, doing a little better on video, but a little worse on audio. However, bootstrapping from the video modality does much worse, actually degrading the audio classifier's performance.

Since it is usually not known what level of noise the system will encounter during its deployment, the labeled data collected for training the user-independent models is often clean. However, the case when the test environment is noisier than the training data is precisely when visual input helps the most. Therefore, a compelling application of our algorithm would be to adapt not only to a new user, but to noise in the audio. We repeat the previous experiment, but with UI audio models trained on clean data. The results are shown in Table 3.5. In this case, it is the audio modality that is "weaker", judging from the UI performance in the first column. This time, co-adaptation improves both modalities: the visual from 59.8% to 69.0%, and the audio from 52.8% to 69.9%. On the other hand, bootstrapping from either the video or the audio modalities does worse, with the latter significantly degrading UI visual performance. Finally, the last column shows that single-modality bootstrapping does worse than co-adaptation. The detailed CCR results obtained before and after co-training for each user are shown as bars in Figure 3-3. In most cases, our algorithm either improves the UI model performance (by as much as 134% in the case of user 8's visual model), or does not make it worse.

3.5 Chapter Summary

In this chapter, we investigated the multi-view semi-super-vised co-training algorithm as a means of utilizing unlabeled data in multi-modal HCI learning problems. Intuitively, the method uses single-modality classifiers to help train each other by it-

Clf.	User Indep.	Co-Adaptation	Audio-BS	Video-BS	Single-modality BS
Aud.	52.8 ± 4.8	69.9 ± 7.4 ($p \ll .01$)	55.4 ± 4.5	63.3 ± 11.8	58.6 ± 4.4 ($p \ll .01$)
Vid.	59.8 ± 11.3	69.0 ± 8.6 ($p \ll .01$)	51.5 ± 7.9	62.4 ± 13.2	60.7 ± 12.1 ($p = .03$)

Table 3.5: Co-adaptation results on the speech data, mis-matched audio noise conditions. Each column shows the mean CCR over 39 test speakers, ± the standard deviation. p-values are relative to the UI classifier.

eratively adding high-confidence labels to the common training set. We extended the confidence-based co-training method to HMM classifiers, and showed that it not only learns user-specific speech and gesture classifiers using just a few labeled examples, but it is more accurate than single-modality baselines.

We also proposed an adaptive co-training algorithm, co-adaptation, and showed that it can be used to improve upon existing models trained on a large amount of labeled data when a small amount of unlabeled data from new users or noise conditions becomes available. When either the audio or the visual classifier is more accurate, our method performs as well as bootstrapping from the stronger modality, however, it does not require such knowledge. When both modalities are weak, such as when the user-independent audio speech classifiers are trained on clean audio, but the new condition is noisy, our method improves significantly over single-modality baselines. Interesting avenues of future work include the investigation of sufficiency, the use of co-adaptation to perform high-level adaptation of audio-visual classifiers (e.g., adapting their language model), the use of user-dependent observations and the use of HMM adaptation techniques (MLLR, MAP) in our algorithm.

In this chapter we have assumed that the speech and gesture views are sufficient, e.g., for audio-visual agreement recognition we assumed that the user always says 'yes' when nodding. As one might expect, in real-world scenario this assumption can often be violated as a result of uni-modal expression or other forms of view corruption, e.g., occlusion. In the following Chapters we purse co-training algorithms that are able to learn in these realistic and complex working conditions, where view sufficiency is not always satisfied across all training samples. In the next Chapter we introduce the notion of view disagreement, i.e., when multi-view samples can be considered as

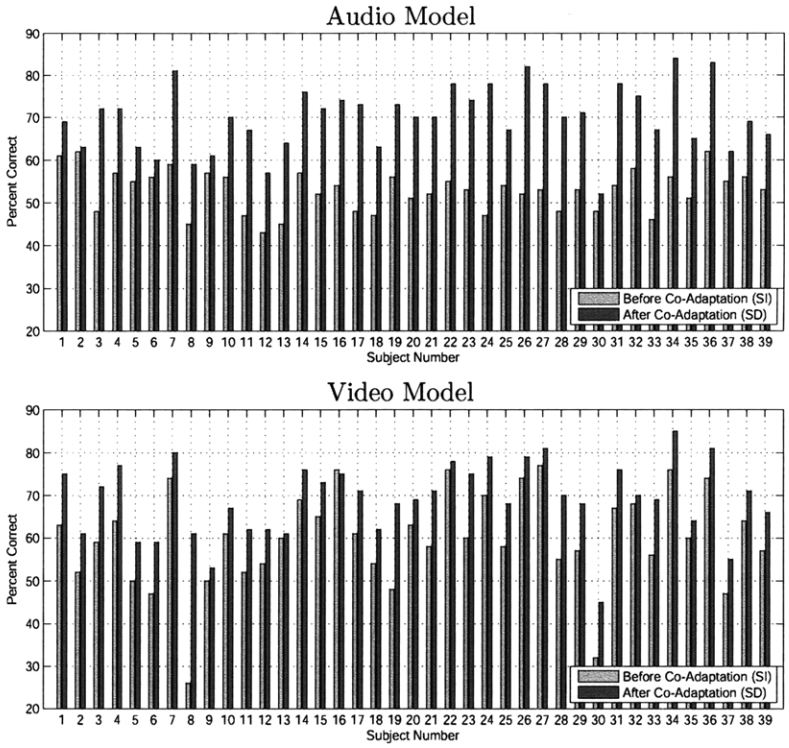


Figure 3-3: Detailed co-adaptation results for mismatched audio noise (summarized by column 2 of Table 3.5) for each of the 39 test speakers. The light bars show the UI models' CCR, the dark bars show the CCR after co-adaptation.

belonging to a separate class as a result of occlusion or other forms of view corruption, and pursue an information theoretic, filter-based co-training algorithm for coping with view disagreement. We then generalize this notion in Chapter 5 where we develop a heteroscedastic Bayesian co-training algorithm that models view sufficiency using a per-sample noise model within a probabilistic framework.

Chapter 4

Multi-View Learning in the Presence of View Disagreement

This chapter focuses on the problem of performing co-training from insufficient views that have been corrupted by complex noise, e.g., occlusion[1]. In Chapter 3 we looked at the problem of co-training from audio-visual data sources, and made the simplifying assumptions that users were always redundant in their expression (e.g., say 'yes' while nodding) and were functioning in a relatively constrained working environment (e.g., users were imaged frontal to the camera in the center of the frame). As one would expect, however, these assumptions can be violated quite drastically in real-world scenarios where users employ uni-modal expression and/or in the presence of view occlusions, and other forms of complex noise and missing data. As will be discussed in this and the following chapter, in fact, this problem of view disagreement is general to many multi-sensor classification scenarios and is a limitation of existing multi-view learning algorithms. In what follows, we present a filter-based multi-view learning approach for coping with view disagreement samples. In Chapter 5 we then generalize the notion of view disagreement to per-sample noise processes modeled within a probabilistic framework with GPs.

[1]The work described in this chapter is published in the Conference on Uncertainty in Artificial Intelligence, 2008, Christoudias, Urtasun and Darrell [21].

4.1 Introduction

A common assumption in multi-view learning is that the samples from each view always belong to the same class. In realistic settings, datasets are often corrupted by noise. Multi-view learning approaches have difficulty dealing with noisy observations, especially when each view is corrupted by an independent noise process. For example, in multi-sensory datasets it is common that an observation in one view is corrupted while the corresponding observations in other views remain unaffected (e.g., the sensor is temporarily in an erroneous condition before returning back to normal behavior). Indeed, if the corruption is severe, the class can no longer be reliably inferred from the corrupted sample.

These corrupted samples can be seen as belonging to a "neutral" or "background" class that co-occur with un-corrupted observations in other views. The view corruption problem is thus a source of *view disagreement*, i.e., the samples from each view do not always belong to the same class but sometimes belong to an additional background class as a result of view corruption or noise. In this chapter we present a method for performing multi-view learning in the presence of view disagreement caused by view corruption. Our approach treats each view as corrupted by a structured noise process and detects view disagreement by exploiting the joint view statistics using a conditional entropy measure.

We are particularly interested in inferring multi-modal semantics from weakly supervised audio-visual speech and gesture data. In audio-visual problems view disagreement often arises as a result of temporary view occlusion, or uni-modal expression (e.g., when expressing agreement a person may say 'yes' without head nodding).

The underlying assumption of our approach is that *foreground* samples can co-occur with samples of the same class or background, whereas background samples can co-occur with samples from any class, a reasonable assumption for many audio-visual problems. We define new multi-view bootstrapping approaches that use conditional entropy in a pre-filtering step to reliably learn in the presence of view disagreement. Experimental evaluation on audio-visual data demonstrates that the detection and

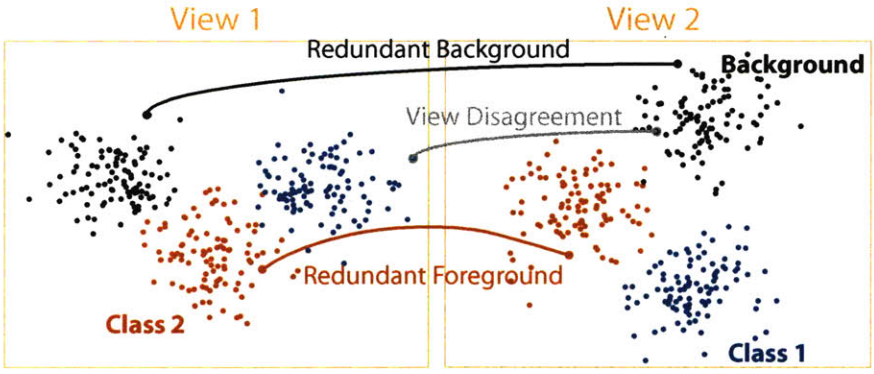


Figure 4-1: Synthetic two-view problem with normally distributed classes, two foreground and one background. Each view is 2-D; the two foreground classes are shown in red and blue. Corrupted samples form a separate background class (black samples) that co-occur with un-corrupted samples. For each point in the left view there is a corresponding point in the right view. Three point correspondences are shown: a redundant foreground sample, a redundant background sample and a sample with view disagreement where view 1 observed an instance of class 1, but view 2 for that sample was actually an observation of the background class. View disagreement occurs when one view is occluded and is incorrectly paired with background. Multi-view learning with these pairings leads to corrupted foreground class models.

filtering of view disagreement enables multi-view learning to succeed despite large amounts of view disagreement.

The remainder of this chapter is organized as follows. In the next section, a discussion of multi-view learning approaches and view disagreement is provided. Our conditional entropy based criterion for detecting view disagreement is then outlined in Section 4.3 and our multi-view bootstrapping approach is presented in Section 4.4. Experimental results are provided in Section 7.3. A discussion of related methods and connections between our work and other statistical techniques is given in Section 4.6. Finally, in Section 4.7 we provide a summary and discuss future work.

4.2 View Disagreement

Multi-view learning algorithms function on the common underlying principle of view agreement. More formally, let $\mathbf{x}_k = (x_k^1, ..., x_k^V)$ be a multi-view sample with V views, and let $f_i : x^i \rightarrow \mathcal{Y}$ be the classifier that we seek in each view. Multi-view

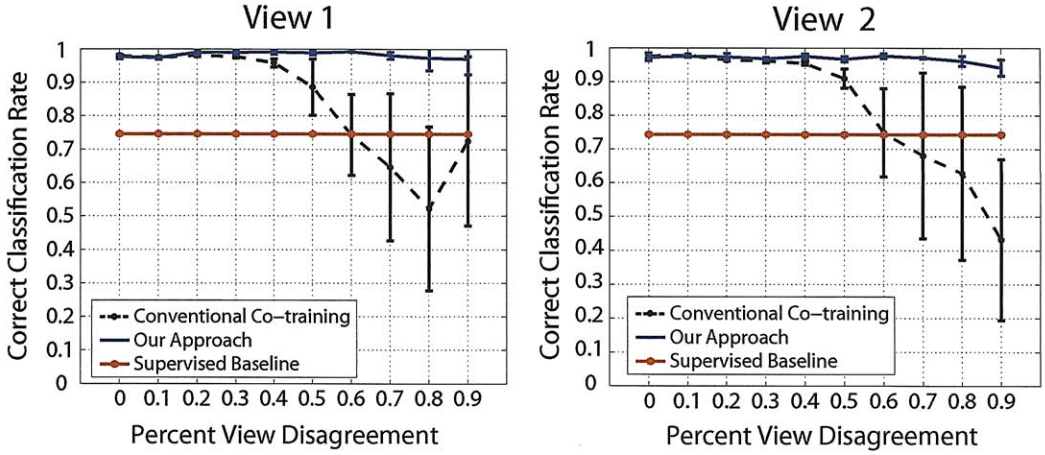


Figure 4-2: Multi-view learning for synthetic two-view example with varying amounts of view disagreement. Average performance is shown computed over 10 random splits of the training data into labeled and unlabeled sets; the error bars indicate ± 1 std. deviation. Our approach exhibits good performance at all view disagreement levels while conventional co-training begins to diverge for percent disagreement greater than 40%.

learning techniques train a set of classifiers $\{f_i\}$ by maximizing their consensus on the unlabeled data, $\mathbf{x}_k \in U$, for example by minimizing the L_2 norm [97],

$$\min \sum_{\mathbf{x_k} \in U} \sum_{i \neq j} \| f_i(x_k^i) - f_j(x_k^j) \|_2^2 \tag{4.1}$$

The minimization in Eq. (4.1) is only applicable to multi-view learning problems for which the views are *sufficient* for classification, i.e., that classification can be performed from either view alone. In practice, however, it is often difficult to define views that are fully sufficient. Previous methods for overcoming insufficiency have addressed the case where both views are necessary for classification [26, 9, 97]. These methods formulate multi-view learning as a global optimization problem that explicitly maximizes the consensus between views. Although these approaches allow for views with partial insufficiency, they still assume that each view is largely sufficient. In the presence of significant view disagreement these approaches would in general diverge and perform poorly.

In this chapter we identify and address a new form of insufficiency inherent to

many real-world datasets, caused by samples where each view potentially belongs to a different class, e.g., as a result of view corruption. We refer to this form of insufficiency as the *view disagreement* problem. The view disagreement problem is distinct from the forms of view insufficiency that have been addressed in the literature—previous methods for overcoming insufficiency have addressed the case where both views are necessary for classification [6, 26, 75, 97], but not the case where the samples from each view potentially belong to different classes.

The problem of view disagreement exists in many real-world datasets. In user agreement recognition from head gesture and speech [20], people often say 'yes' without head nodding and vice versa, and/or the subject can also become temporary occluded in either the audio or visual modalities by other speakers or objects in the scene. In semantic concept detection from text and video [123], it is possible for the text to describe a different event than what is being displayed in the video. Another example is web-page classification from page and hyper-link content [11], where the hyper-links can often point to extraneous web-pages not relevant to the classification task.

We illustrate the problem of view disagreement in multi-view learning with a toy example containing two views of two foreground classes and one background class. The samples of each class are drawn from Gaussian distributions with unit variance (see Figure 4-1). Figure 4-2 shows the degradation in performance of conventional co-training [11] for varying amounts of view disagreement. Here, co-training is evaluated using a left out test set and by randomly splitting the training set into labeled and unlabeled datasets. We report average performance across 10 random splits of the training data. As shown in Figure 4-2 co-training performs poorly when subject to significant amounts of view disagreement ($\geq 40\%$).

In what follows, we present a method for detecting view disagreement using a measure of conditional view entropy and demonstrate that when used as a pre-filtering step, our approach enables multi-view learning to succeed despite large amounts of view disagreement.

4.3 Detection and Filtering of View Disagreement

We consider an occlusion process where an additional class models background. We assume that this background class can co-occur with any of the $n+1$ classes in the other views[2], and that the n foreground classes only co-occur with samples that belong to the same class or background, as is common in audio-visual datasets [20].

In this chapter we propose a conditional entropy criterion for detecting samples with view disagreement. We further assume that background co-occurs with more than one foreground class; this is a reasonable assumption for many types of background (e.g., audio silence). In what follows, we treat each view x^i, $i = 1, ..., V$ as a random variable and detect view disagreement by examining the joint statistics of the different views. The entropy $H(x)$ of a random variable is a measure of its uncertainty [27]. Similarly, the conditional entropy $H(x|y)$ is a measure of the uncertainty in x given that we have observed y. In the multi-view setting, the conditional entropy between views, $H(x^i|x^j)$, can be used as a measure of agreement that indicates whether the views of a sample belong to the same class or event. In what follows, we call $H(x^i|x^j)$ the *conditional view entropy.*

Under our assumptions we expect the conditional view entropy to be larger when conditioning on background compared to foreground. Thus, we have $\forall p = 1, ...n$,

$$H(x^i|x^j_k \in C^{n+1}) > H(x^i|x^j_l \in C^p) \tag{4.2}$$

where C^i is the set of examples belonging to class i. A notional example of view corruption is illustrated in Figure 4-3. This example contains two, 1-D views of two foreground classes and one background class. As before, the samples of each class are drawn from a normal distribution with unit variance. The conditional view distributions of a multi-view sample with view disagreement is displayed. Note that the uncertainty of view i when conditioning on view j has greater uncertainty when view j is background.

[2]Note that background samples can co-occur with the any of the n foreground classes plus background.

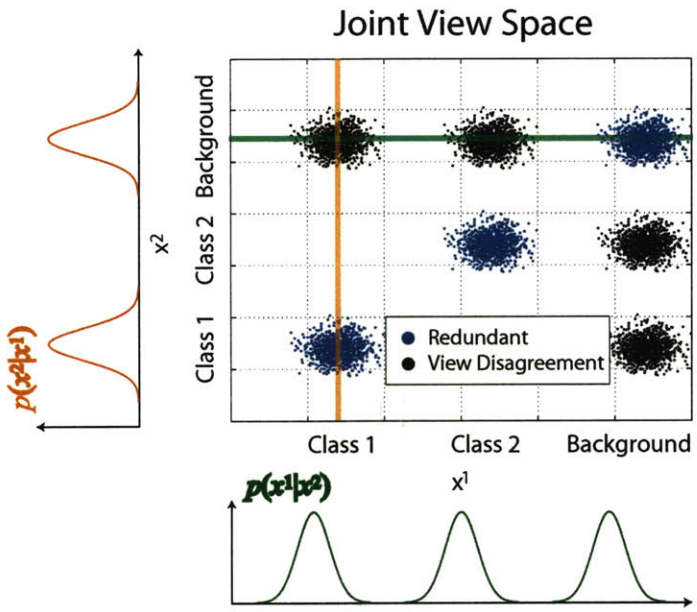


Figure 4-3: View disagreement caused by view corruption. The joint view space of a two-view problem with view disagreement is displayed. Redundant samples are highlighted in blue and samples with view disagreement in black. The conditional distributions for a sample with view disagreement are shown. The conditional distribution resulting from conditioning on background exhibits more peaks and therefore has a higher uncertainty than when conditioning on foreground.

We delineate foreground from background samples by thresholding the conditional view entropy. In particular, we define the threshold in each view using the mean conditional entropy computed over the unlabeled data. More formally, let (x_k^i, x_k^j) be two different views of a multi-view sample $\mathbf{x}_k = (x_k^1, ..., x_k^V)$. We define an indicator function, $m(\cdot)$, that operates over view pairs (x^i, x^j) and that is 1 if the conditional entropy of x^i conditioned on x_k^j is below the mean conditional entropy,

$$m(x^i, x_k^j) = \begin{cases} 1, & H(x^i|x_k^j) < \bar{H}_{ij} \\ 0, & \text{otherwise} \end{cases}, \tag{4.3}$$

with

$$H(x^i|x_k^j) = -\sum_{x^i \in U^i} p(x^i|x_k^j) \log p(x^i|x_k^j), \tag{4.4}$$

where U^i is the ith view of the unlabeled dataset, and $\bar{H}_{ij}$ is the mean conditional

entropy,

$$\bar{H}_{ij} = \frac{1}{M} \sum_{\mathbf{x}_k \in U} H(x^i | x_k^j), \tag{4.5}$$

where M is the number of samples in U. $m(x^i, x^j)$ is used to detect whether x^j belongs to foreground since under our model foreground samples have a low conditional view entropy.

A sample $\mathbf{x}_k$ is a *redundant foreground* sample if it satisfies

$$\prod_{i=1}^{V} \prod_{j \neq i} m(x^i, x_k^j) = 1. \tag{4.6}$$

Similarly, $\mathbf{x}_k$ is a *redundant background* sample if it satisfies

$$\sum_{i=1}^{V} \sum_{j \neq i} m(x^i, x_k^j) = 0. \tag{4.7}$$

A multi-view sample x_k is in *view disagreement* if it is neither a redundant foreground nor a redundant background sample.

Definition 1. Two views (x_k^i, x_k^j) of a multi-view sample $\mathbf{x}_k$ are in view disagreement if

$$m(x^i, x_k^j) \oplus m(x^j, x_k^i) = 1 \tag{4.8}$$

where $\oplus$ is the logical xor operator that has the property that $a \oplus b$ is 1 iff $a \neq b$ and 0 otherwise.

Eq. (4.8) defines our conditional entropy criterion for view disagreement detection between pairs of views of a multi-view sample.

In practice, we estimate the conditional probability of Eq. (4.4) as

$$p(x^i | x_k^j) = \frac{f(x^i, x_k^j)}{\sum_{x^i \in U^i} f(x^i, x_k^j)} \tag{4.9}$$

where $f(\mathbf{x})$ is a multivariate kernel density estimator[3]. In our experiments, the band-

[3]Note our approach is agnostic to the choice of probability model and more sophisticated condi-

Algorithm 3 Multi-View Bootstrapping in the Presence of View Disagreement

Given classifiers f_i and labeled seed sets S_i, $i = 1, ..., V$, unlabeled dataset U and parameters N and T:
Set $t = 1$.
repeat
  for $i = 1, ..., V$ **do**
    Train f_i on S_i
    Evaluate f_i on U^i
    Sort U in decreasing order by f_i confidence
    for each $\mathbf{x}_k \in U$, $k = 1, ..., N$ **do**
      for $j \neq i$ **do**
        if $\neg(m(x^i, x^j_k) \oplus m(x^j, x^i_k))$ **then**
          $U^j = U^j \backslash \{x^j_k\}$
          $S^j = S^j \cup \{x^j_k\}$
        end if
      end for
      $U^i = U^i \backslash \{x^i_k\}$
      $S^i = S^i \cup \{x^i_k\}$
    end for
  end for
  Set $t = t + 1$.
until $|U| = \emptyset$ or $t = T$

width of f is set using automatic bandwidth selection techniques [94].

4.4 Multi-view Bootstrapping in the Presence of View Disagreement

In this section we present a new multi-view bootstrapping algorithm that uses the conditional entropy measure of Eq. (4.8) in a pre-filtering step to learn from multi-view datasets with view disagreement.

Multi-view bootstrapping techniques, e.g., co-training, mutually train a set of classifiers, f_i, $i = 1, ..., V$, on an unlabeled dataset U by iteratively evaluating each classifier and re-training from confidently classified samples. The classifiers are initialized from a small set of labeled examples typically referred to as the *seed set*,

tional probability models can be used, such as [111], that perform better in high dimensional input spaces.

Algorithm 4 Cross-Modality Bootstrapping in the Presence of View Disagreement

Given existing classifier f_1 and initial classifier f_2, unlabeled dataset U and parameter N:

Initialization:
Sort U in decreasing order by f_1 confidence
Define $L = \{(y_k, x_k^2)\}$, $k = 1, ..., N$

Bootstrapping:
Set $S = \emptyset$
for each $(y_k, x^2) \in L$ **do**
  if $\neg(m(y, x_k^2) \oplus m(x^2, y_k))$ **then**
    $S = S \cup \{(y_k, x_k^2)\}$
    $L = L \backslash \{(y_k, x_k^2)\}$
  end if
end for
Train f_2 on S.

S. During bootstrapping, confidently classified samples in each view are used to label corresponding samples in the other views. It has been shown that multi-view bootstrapping is advantageous to self-training with only a single view [20].

We extend multi-view bootstrapping to function in the presence of view disagreement. A separate labeled set, S_i, is maintained for each view during bootstrapping and the conditional entropy measure of Eq. (4.8) is checked before labeling samples in the other views from labels in the current view. The parameters to the algorithm are N, the number of samples labeled by each classifier during each iteration of bootstrapping, and T the maximum number of multi-view bootstrapping iterations. The resulting algorithm self-trains each classifier using all of the unlabeled examples, and only enforces a consensus on the samples with view agreement (see Algorithm 3).

Figure 4-2 displays the result of multi-view bootstrapping for the toy example of Figure 4-1 using $N = 6$ and T was set such that all the unlabeled data was used. With our method, multi-view learning is able to proceed successfully despite the presence of severe view disagreement and is able to learn accurate classifiers in each view even when presented with datasets that contain up-to 90% view disagreement.

In audio-visual problems it is commonly the case that there is an imbalance between the classification difficulty in each view. In such cases, an accurate classifier can

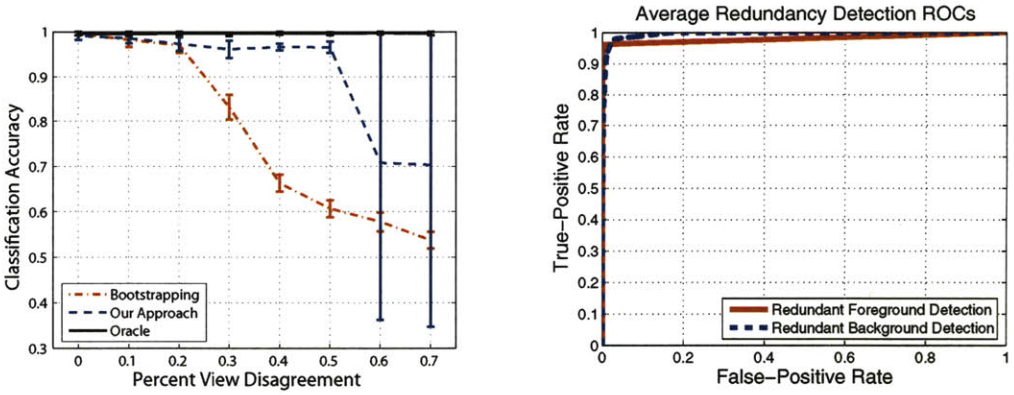
be learned in the weaker view using an unsupervised learning method that bootstraps from labels output by the classifier in the other view. Here, the class labels output by the classifier in the stronger view can be used as input to the conditional entropy measure as they provide a more structured input than the original input signal.

The resulting cross-modality bootstrapping algorithm trains a classifier f_2 in the second view from an existing classifier f_1 in the first view on a two-view unlabeled dataset U. The algorithm proceeds as follows. First f_1 is evaluated on U and the N most confidently classified examples are moved from U to the labeled set L. The conditional entropy measure is then evaluated over each label, sample pair $(y, x^2) \in L$, where $y = f_1(x^1)$. The final classifier f_2 results from training on the the samples in L that are detected as redundant foreground or redundant background (see Algorithm 4).

4.5 Experimental Evaluation

We evaluate the performance of multi-view bootstrapping techniques on the task of audio-visual user agreement recognition from speech and head gesture. Although users often use redundant expression of agreement, it is frequently the case that they say 'yes' without head gesturing and vice-versa. View disagreement can also be caused by noisy acoustic environments (e.g., a crowded room), temporary visual occlusions by other objects in the scene, or if the subject is temporarily out of the camera's field of view.

To evaluate our approach we used a dataset of 15 subjects interacting with an avatar in a conversational dialog task [20]. The interactions included portions where each subject answered a set of yes/no questions using head gesture and speech. The head gesture consisted of head nods and shakes and the speech data of 'yes' and 'no' utterances. In our experiments, we simulate view disagreement in the visual domain using both no motion (i.e., random noise) and real background head motion samples from non-response portions of the interaction. Similarly, background in the audio is simulated as babble noise.



(a) (b)

Figure 4-4: Bootstrapping a user agreement visual classifier from audio. (a) Performance is shown averaged over random splits of the data into 10 train and 5 test subjects over varying amounts of simulated view disagreement using a no motion background class; error bars indicate $\pm$ 1 std. deviation. Unlike conventional bootstrapping, our approach is able to cope with up-to 50% view disagreement. (b) Average view disagreement detection ROCs are also shown for redundant foreground and background detection. Our approach effectively detects view disagreement.

The visual features consist of 3-D head rotation velocities output by a 6-D head tracker [74]. For each subject, we post-process these observations by computing a 32 sample windowed Fast Fourier Transform (FFT) separately over each dimension, with a time window of 1 second corresponding to the expected length of a head gesture. The resulting sequence of FFT observations is then segmented using the avatar's transcript which marks the beginning and end of each user response.

The FFT spectra of each user response were amplitude normalized and blurred in space and time to remove variability to location, duration and rate of head motion. Principle Components Analysis (PCA) was then performed over the vector space resulting from flattening the FFT spectra corresponding to each response into a single vector. The resulting 3-D PCA space captured over 90% of the variance and was computed over the unlabeled samples of the training set.

The audio features consist of 13-D Mel Frequency Cepstral Coefficients (MFCCs) sampled at 100Hz over the segmented audio sequences corresponding to each user response, obtained from the avatar's transcript. The audio sequences were then converted into single frame observations using the technique of [46]. In this rep-

resentation, an audio sequence is divided into portions and an average MFCC vector is computed over each portion. In our experiments, we used proportions equal to $(0.3, 0.4, 0.3)$. The concatenated averages along with first derivatives and log duration define a 61-D observation vector. To reduce the dimensionality of this space, PCA was applied retaining 98% of the variance which resulted in a 9-D, single-frame audio observation space.

In our experiments we use correct classification rate as the evaluation metric, defined as:

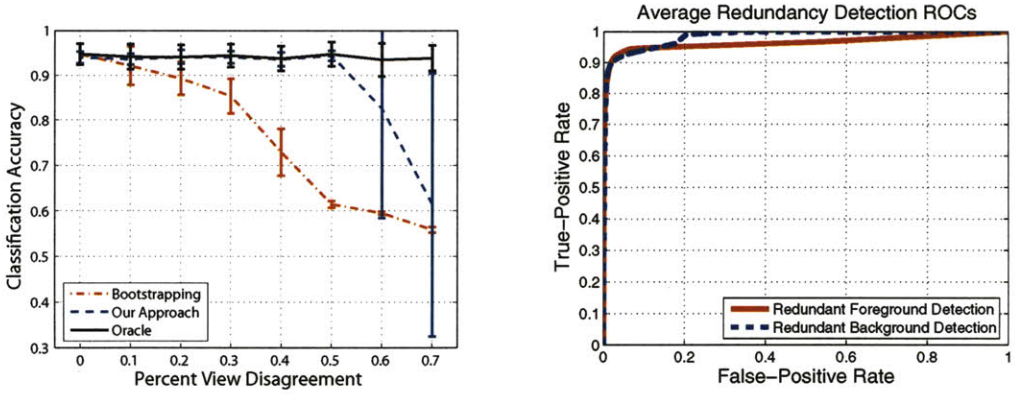
$$\text{CCR} = \frac{\text{\# of examples correctly classified}}{\text{total \# of examples}} \tag{4.10}$$

We used Bayes classifiers for audio and visual gesture recognition defined as $p(y|x) = \frac{p(x|y)}{\sum_y p(x|y)}$, where $p(x|y)$ is Gaussian. Specifically, Bayes classifiers for $p(y|x^a)$ and $p(y|x^v)$ are bootstrapped from semi-supervised audio-visual data; x^a and x^v correspond to audio and visual observations respectively.

4.5.1 Cross-Modality Bootstrapping

First, we evaluate our cross-modality bootstrapping approach. For this task, we are interested in performing semi-supervised learning of visual head gesture by bootstrapping from labels in the speech (e.g., those output by an off-the-shelf speech recognizer). We simulated view disagreement by randomly replacing observations in the visual modality with background sequences, and replacing labels in the audio with the background label. Redundant background was also added such that there were an equal number of redundant background samples as there were redundant foreground samples per class.

We first show results using a "no motion" visual background modeled as zero mean Gaussian noise in the 3-D head rotational velocity space with $\sigma = 0.1$. Figure 4-4 displays the result of evaluating the performance of multi-view bootstrapping (Algorithm 4) with varying amounts of view disagreement. Performance is shown averaged over 5 random splits of the data into 10 train and 5 test subjects. At small amounts of view disagreement ($\leq 20\%$) conventional bootstrapping and our approach



(a) (b)

Figure 4-5: Bootstrapping a user agreement visual classifier from audio with real visual background. Performance is shown averaged over random splits of the data into 10 train and 5 test subjects over varying amounts of simulated view disagreement; error bars indicate $\pm$ 1 std. deviation. The conventional bootstrapping baseline performs poorly in the presence of view disagreement. In contrast, our approach is able to (a) successfully learn a visual classifier and (b) classify views in the presence of significant view disagreement (up to 50%).

exhibit similar good performance. When the view disagreement is small the error can be viewed as classification noise in the audio. For larger amounts of view disagreement (up to 50%), conventional multi-view bootstrapping diverges and our algorithm still succeeds in learning an accurate head gesture recognizer from the audio-visual data. For $>$ 50% view disagreement, our approach begins to degrade and exhibits a large variance in performance. This high variability can be a result of poor bandwidth selection, or a poor choice of threshold. We plan to investigate alternative methods for modeling the conditional probability and more sophisticated threshold selection techniques as part of future work.

Figure 4-4(b) displays average receiver-operator curves (ROCs) for redundant foreground and background class detection that result from varying the entropy threshold of the conditional entropy measure. The mean conditional entropy defines a point on these curves. As illustrated by the figure, overall our approach does fairly well in detecting view disagreement.

Next, we consider a more realistic occluding visual background class generated from randomly selecting head motion sequences from non-response portions of each

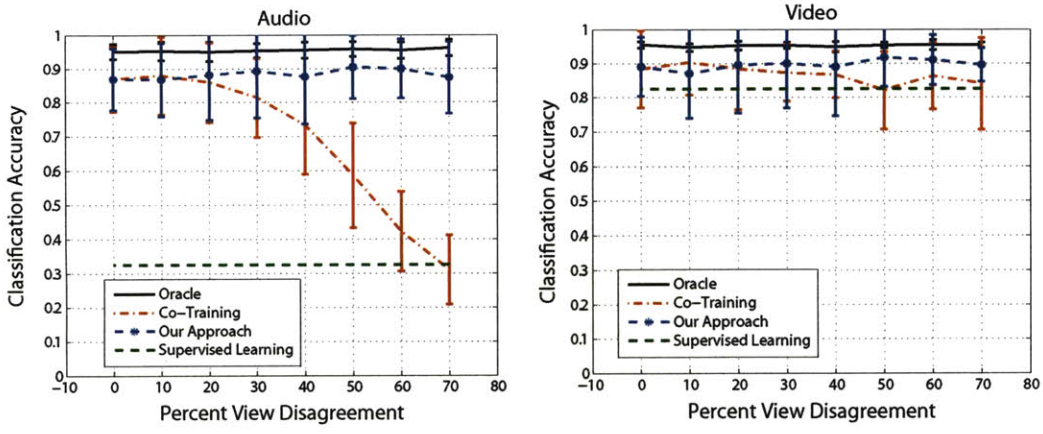


Figure 4-6: Multi-view bootstrapping of audio-visual user agreement classifiers. Performance of (a) audio and (b) video is displayed averaged over 5 random splits of the data into 10 train and 5 test subjects and over 10 random splits of each training set into labeled seed set and unlabeled dataset; error bars show ± 1 std. deviation. Conventional co-training performs poorly in the presence of significant view disagreement. In contrast, our approach performs well across all view disagreement levels.

user interaction. In contrast to the "no motion" class considered above, these segments contain miscellaneous head motion in addition to no motion.

Our view disagreement detection approach (Algorithm 4) performs equally well in the presence of the more challenging real background as is shown in Figure 4-5. As before, conventional bootstrapping performs poorly in the presence of view disagreement. In contrast, our approach is able to successfully learn a visual classifier in the presence of significant view disagreement (up to 50%).

4.5.2 Multi-View Bootstrapping

We evaluated the performance of multi-view bootstrapping (Algorithm 3) for the task of semi-supervised learning of audio-visual user agreement classifiers from speech and head gesture. Figure 4-6 displays the result of audio-visual co-training for varying amounts of view disagreement. Performance is shown averaged over 5 random splits of the data into 10 train and 5 test subjects and over 10 random splits of the training data into labeled seed set and unlabeled training set, with 15 labeled samples, 5 per

class. Conventional co-training and our approach were then evaluated using $N = 6$ and $T = 100$. We chose N such that the classes are balanced.

For this problem, the initial visual classifier trained from the seed set is much more accurate than the initial audio classifier that performs near chance. The goal of co-training is to learn accurate classifiers in both the audio and visual modalities. Note, that in contrast to cross-modality bootstrapping, this is done without any a priori knowledge as to which modality is more reliable. For small amounts of view disagreement ($\geq 20\%$), both conventional co-training and our approach (Algorithm 3) are able to exploit the strong performance in the visual modality to train an accurate classifier in the audio. For larger amounts of view disagreement, conventional co-training begins to diverge and at the 70% view disagreement level is not able to improve over the supervised baseline in both the audio and visual modalities. In contrast, our approach reliably learns accurate audio-visual classifiers across all view disagreement levels.

4.6 Discussion

Recently, Ando and Zhang [6] presented a multi-view learning approach that instead of assuming a consensus over classification functions assume that the views share the same low dimensional manifold. This has the advantage that it can cope with insufficient views where classification cannot be performed from either view alone. Still, their approach defines a consensus between views, and therefore assumes that the samples in each view are of the same class. View disagreement will violate this assumption and we expect their method to degrade as multi-view bootstrapping.

Our approach treats each view as corrupted by a structured noise process and detects view disagreement by exploiting the joint view statistics. An alternative method to coping with view disagreement is to treat each view as belonging to a stochastic process and use a measure such as mutual information to test for view dependency [101, 100]. In [100], Siracusa and Fisher use hypothesis testing with a hidden factorization Markov model to infer dependency between audio-visual streams.

It would be interesting to apply such techniques for performing multi-view learning despite view disagreement, which we leave as part of future work.

Our work bears similarity to co-clustering approaches which use co-occurrence statistics to perform multi-view clustering [31, 30, 29]. These techniques, however, do not explore the relationship between co-occurrence and view sufficiency and would suffer in the presence of view disagreement since the occluding background would potentially cause foreground clusters to collapse into a single cluster.

We demonstrated our view disagreement detection and filtering approach for multi-view bootstrapping techniques (e.g., [11, 79, 20]). However, our algorithm is generally applicable to any multi-view learning method and we believe it will be straightforward to adapt it for use with other approaches (e.g., [6, 26, 97]). Multi-view learning methods either implicitly or explicitly maximize the consensus between views to learn from unlabeled data; view disagreement adversely affects multi-view learning techniques since they encourage agreement between views.

In our experiments, our approach performs well on a realistic dataset with noisy observations. The success of our approach on this dataset is predicated on the fact that foreground and background classes exhibit distinct co-occurrence patterns, which our algorithm exploits to reliably detect view disagreement.

4.7 Chapter Summary

In this chapter we have identified a new multi-view learning problem, view disagreement, inherent to many real-word multi-view datasets. We presented a multi-view learning framework for performing semi-supervised learning from multi-view datasets in the presence of view disagreement and demonstrated that a conditional entropy criterion was able to detect view disagreement caused by view corruption or noise. As shown in our experiments, for the task of audio-visual user agreement our method was able to successfully perform multi-view learning even in the presence of gross view disagreement ($50 - 70\%$). Interesting avenues for future work include the investigation of alternative entropy threshold selection techniques, the use of alternative proba-

bility models for computing conditional entropy and modeling redundancy between non-stationary stochastic processes using measures such as mutual information.

View disagreement noise can be seen as a form of binary per-sample view corruption in which a sample is either corrupted or un-corrupted. In the more general setting, however, the per-sample noise is continuous with certain samples being more corrupted than others; also a sample can still be informative even if it is corrupted by noise, and a filter-based approach would in effect ignore such samples. In the next chapter we pursue a more general setting of the view disagreement problem. We model view sufficiency due to noise using heteroscedastic noise models within a probabilistic framework, in which the per-sample noise is simultaneously discovered while solving the classification task. As is shown in the following chapter, the proposed multi-view learning approach is general and can handle a variety of complex noise processes including binary view disagreement.

Chapter 5

Co-training with Noisy Perceptual Observations

In the previous couple chapters, we demonstrated the use of co-training for learning audio-visual classifiers and performing model adaptation, and addressed the problem of view disagreement, an important limitation of multi-view learning algorithms[1]. In this chapter we continue the investigation of view insufficiency due to complex noise processes such as per-sample occlusion and uni-modal expression, that are commonly encountered in multi-sensor, perceptual learning problems. In particular we propose a heteroscedastic Bayesian co-training algorithm that extends the Bayesian co-training algorithm of Yu et. al. [128] to model per-sample noise processes. Unlike the filter-based co-training method of Chapter 4, our heteroscedastic Bayesian co-training approach can handle arbitrary view corruption processes including binary view disagreement. We demonstrate our approach for performing multi-view semi-supervised learning within a variety of perceptual learning tasks.

[1]The work described in this chapter is published in the Conference on Computer Vision and Patter Recognition, 2009, Christoudias, Urtasun, Kapoor and Darrell [23].

5.1 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 meta-data (image captions, observation parameters).

With a few notable exceptions [20, 123, 63], 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 [123, 75, 128]. 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.

The previous Chapter 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. As reviewed in Chapter 2, Yu et al. [128] 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 chapter 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.

5.2 Heteroscedastic Bayesian Co-training

Bayesian co-training is a probabilistic co-training framework proposed by Yu et. al. [128] that generalizes existing multi-view learning techniques. The Bayesian co-training algorithm was reviewed in Chapter 2. Recall, that in Bayesian co-training latent predictor variables $\mathbf{f}_j$ are defined in each input view, $\mathbf{x}^j$, $j = 1, \cdots, V$, whose decisions are fused by a latent consensus function $\mathbf{f}_c$ using Gaussian potentials in each view with variance σ_j^2. A GP prior is assumed on the latent functions $\mathbf{f}_j$ and σ_j^2, $i = 1, \cdots, V$ model the global reliability of each view.

To deal with noisy data, in this chapter 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

$$\psi(\mathbf{f}_j, \mathbf{f}_c) = \mathcal{N}(\mathbf{f}_j, \mathbf{A}_j) \tag{5.1}$$

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. 5-1 depicts the

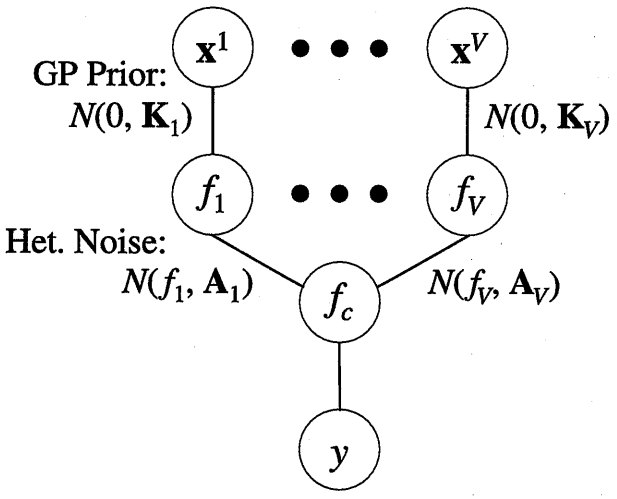


Figure 5-1: Graphical model of *Heteroscedastic Bayesian Co-training* (our approach). Our multi-view learning approach extends Bayesian co-training to incorporate sample-dependent noise modeled by the per view noise covariance matrices $\mathbf{A}_j$. This contrasts the original Bayesian co-training model depicted in Figure 2-4 that incorporates sample-independent noise terms per view σ_i^2 and is a special case of our more general framework. Our method simultaneously discovers the amount of noise in each view while solving the classification task.

undirected graphical model of our *Heteroscedastic Bayesian Co-training* approach.

Integrating out the latent functions $\mathbf{f}_j$ in (7.1) results in a GP prior over the consensus function such that

$$p(\mathbf{f}_c) = \mathcal{N}(0, \mathbf{K}_c) \tag{5.2}$$

with covariance

$$\mathbf{K}_c = \left[\sum_j (\mathbf{K}_j + \mathbf{A}_j)^{-1} \right]^{-1} . \tag{5.3}$$

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[2]. 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.

[2]See Appendix A for a derivation of the heteroscedastic co-training kernel.

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 independent per-sample noise, the covariance is restricted to be diagonal

$$\mathbf{A}_j = diag(\sigma_{1,j}^2, \cdots, \sigma_{N,j}^2) \tag{5.4}$$

where $\sigma_{i,j}^2$ is the estimate of the noise corrupting sample i in view j. The resulting 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.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, $\phi_j = [\sigma_{1,j}^2, \cdots \sigma_{P,j}^2]^T \in \Re^{P\times 1}$ as

$$\mathbf{A}_j = diag(\mathbf{E}^{(j)} \cdot \phi_j) \ . \tag{5.5}$$

The indicator matrices, $\mathbf{E}^{(j)} = [\mathbf{e}_1^{(j)}, \cdots, \mathbf{e}_N^{(j)}]^T$ are matrices such that each row, $\mathbf{e}_i^{(j)} \in \{0,1\}^{P\times 1}$, 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 [128], and if $P=N$, and $E^{(j)}$ is full rank, we recover the full 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 hyper-parameters $\Theta = \{\theta_1, \cdots, \theta_V\}$ 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

$$\bar{\mathbf{y}}_* = \mathbf{k}_c(\mathbf{X}_*)^T(\hat{\mathbf{K}}_c + \sigma^2\mathbf{I}_N)^{-1}\mathbf{y} \tag{5.6}$$

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.

5.3 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" [18] 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 [45].

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 [20], 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. [21], 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 [74]. 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 [46]. 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.

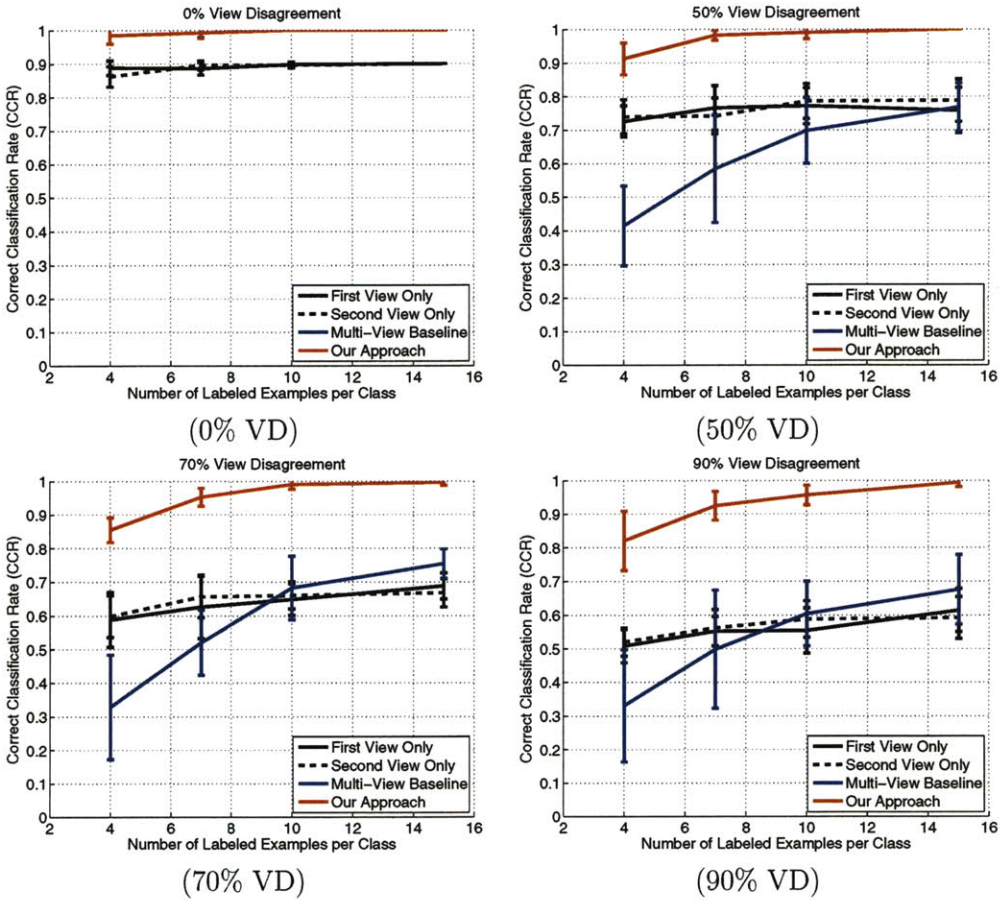
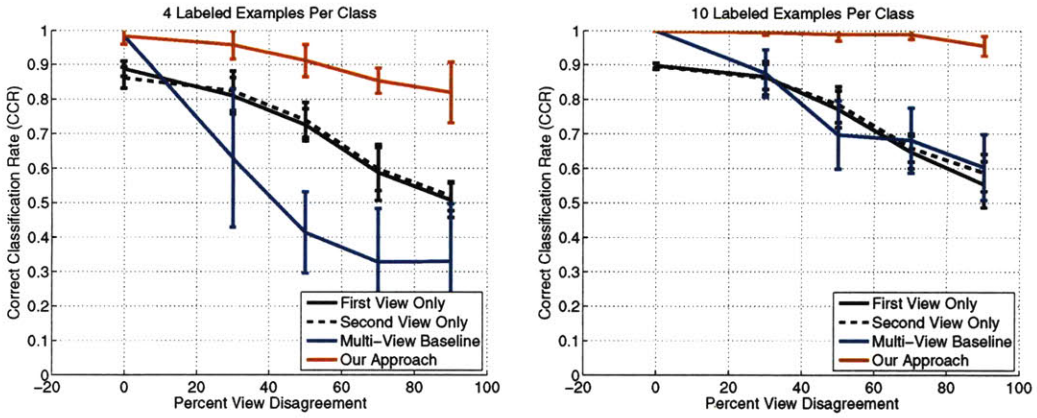


Figure 5-2: **Object recognition from multiple camera sensors with varying training set sizes**: Classification accuracy for a 10-class problem as a function of the number of training samples for different amounts of view disagreement. Performance is shown averaged over 10 splits, the error bars indicate ± 1 std. deviation. Our approach significantly outperforms the single-view and multi-view [128] 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 5-3: **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 $\pm$ 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 [128].

We compare our approach against Bayesian co-training [128] and the approach of Christoudias et. al [21]. We also compare against single view performance using GP regression-based classifiers in each view and multi-view GP kernel combination with and without self-training. We evaluate each approach under the Correct Classification Rate (CCR) evaluation metric defined as

$$\text{CCR} = \frac{\#\text{ samples correctly classified}}{\#\text{ of samples}} \tag{5.7}$$

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.

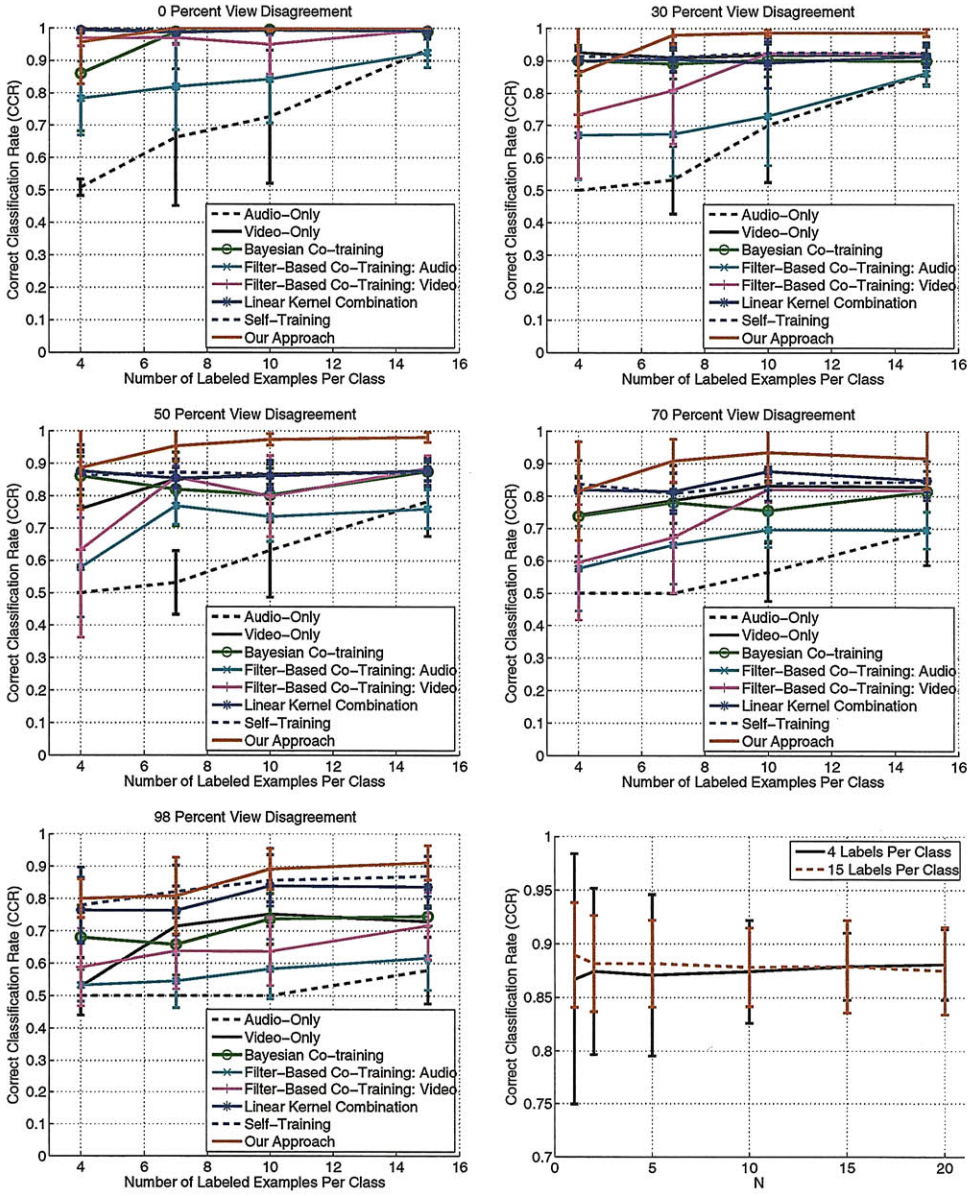


Figure 5-4: **Audio-visual recognition with varying training set sizes**: Classification accuracy as a function of the number of training samples across different amounts of view disagreement. Performance is shown averaged across 10 splits, the error bars indicate $\pm$ 1 std. deviation. Comparison with single-view and multi-view baseline approaches, including Bayesian co-training, the audio and video classifiers from filter-based co-training [21] and the results of multi-view GP kernel combination with and without self-training (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). The performance of self-training as a function of N is also shown. Self-training is fairly insensitive to the setting of this parameter.

5.3.1 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 5-2 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 different amounts of 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 5-3 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 5-2. 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 5-4 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 for different amounts of view disagreement. As before, the performance of single view GP regression-based classification and Bayesian co-training are also shown. Figure 5-5

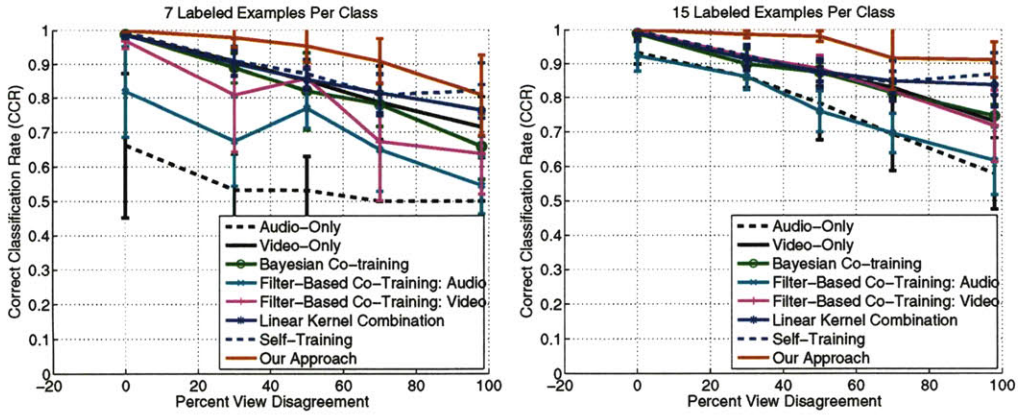


Figure 5-5: **Audio-visual recognition with varying levels of view disagreement**: Classification accuracy as a function of the level of view disagreement. Performance is shown averaged over 10 splits, error bars indicate ± 1 std. deviation. Comparison with single-view and multi-view baseline approaches, including Bayesian co-training, the audio and video classifiers from filter-based co-training [21] and the results of multi-view GP kernel combination with and without self-training (see text for details). 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.

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. [21] on the audio-visual user agreement dataset. Figure 5-4 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 [21] averaged

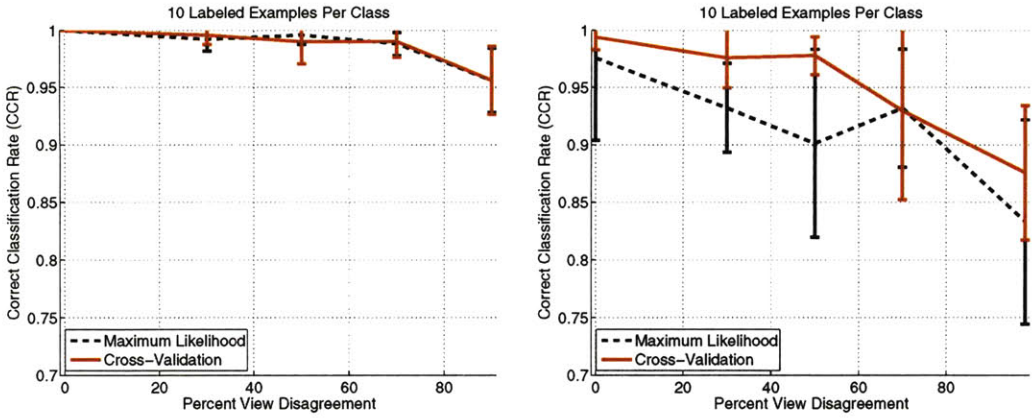


Figure 5-6: **Cross-Validation vs. Maximum Likelihood**: Average performance is shown over 10 splits with 10 labeled examples per class for (top) the multi-view image database and (bottom) the audio-visual gesture database. Cross-validation either matches or outperforms maximum likelihood across both datasets.

over 10 splits of the data with training set sizes M = 4,7,10,15 and for different amounts of view disagreement. Similarly, Figure 5-5 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.

Multiple kernel combination approaches have recently received much attention in the machine learning and computer vision literature [7]. 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 5-4 and 5-5 also display the performance of a kernel combination GP baseline whose covariance function is modeled as the weighted sum of the covariance functions from each

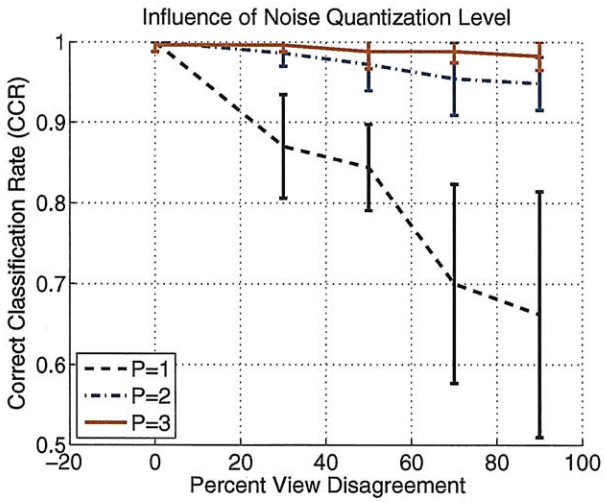


Figure 5-7: **Simultaneously coping with partial occlusion and view disagreement**: Influence of the number of noise levels P on classification accuracy when the multi-view image data is corrupted by view disagreement and partial occlusion. Performance is shown averaged over 10 splits with $M = 7$, error bars indicate $\pm$ 1 std. deviation. As expected performance improves with increasing model components. With $P = 1$ our model is equivalent to [128].

view, with and without self-training. The performance of the multi-view kernel combination baseline degrades in the presence of view disagreement on this audio-visual gesture recognition task.

For self-training, the kernel combination baseline was bootstrapped on the unlabeled test data, in which N examples per class were added at each iteration. Results are displayed in Figure 5-4 using a conservative setting of N ($N = 2$), although performance of the self-training baselines as a function of N is also displayed and self-training exhibits similar performance across the different settings of this parameter. In contrast to our co-training approach, self-training displays similar performance to supervised kernel combination and is unable to effectively leverage both labeled and unlabeled data.

Finally, we evaluated the performance of our approach using both maximum likelihood and n-fold cross-validation parameter learning. Figure 5-6 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.

5.3.2 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 5-7 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.

5.4 Chapter Summary

In this chapter 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 correlated 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.

Our heteroscedastic Bayesian co-training approach makes the limiting assumption of knowing the noise component indicator matrices over the labeled data. In general, one would expect that these matrices can be difficult to known a priori and a mechanism for learning them is desirable. In the following chapter we investigate the related problem of multiple kernel learning and propose an algorithm within this framework for learning the indicator matrices and computing a local weighting of the input space useful for performing multiple kernel combination over insufficient, noisy views.

Chapter 6

Localized Multiple Kernel Learning with Gaussian Processes

The previous chapters have focused on semi-supervised learning from multiple sources with a focus on coping with insufficient, noisy views. This chapter looks at the related problem of supervised multiple kernel learning and investigates the use of local view combination models within this domain.

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 chapter, 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.

6.1 Introduction

Multiple kernel learning approaches to multi-view learning [7, 105, 128] 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. [38] 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. [66] learn an ensemble of SVM classifiers defined on a per-example basis for coping with local variability. Similarly, Gonen and Alpaydin [43] 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 chapter 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 [38], 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 [115, 133], 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.

6.2 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 multi-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

$$p(\mathbf{f}|\bar{\mathbf{X}}) = \mathcal{N}(0, \bar{\mathbf{K}}) \tag{6.1}$$

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. The Bayesian co-training kernel of [128] defines a transductive kernel from multiple views and is useful for classification in the context of both labeled and unlabeled data. In the previous chapter we saw how a per-sample weighting model is important for performing co-training from insufficient views corrupted by complex noise. Here, we pursue a similar model for supervised linear kernel combination and show the importance of a local weighting in this context.

We construct our covariance as a linear combination of covariance matrices

$$\bar{\mathbf{K}} = \sum_{v}^{V} \mathbf{K}^{(v)} + \sigma^2 \mathbf{I} \tag{6.2}$$

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

$$K_{ij}^{(v)} = k_{np}^{(v)}(i,j) \cdot k_p^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)}) \tag{6.3}$$

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

$$\mathbf{K}_{np}^{(v)} = (\mathbf{g}^{(v)})^T \mathbf{g}^{(v)} \tag{6.4}$$

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

$$g_j^{(v)} = \boldsymbol{\alpha}^{(v)} \cdot \mathbf{e}_j \tag{6.5}$$

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

$$\mathcal{L} = \frac{1}{2} \ln |\bar{\mathbf{K}}| + \frac{1}{2} tr(\bar{\mathbf{K}}^{-1} \mathbf{Y}\mathbf{Y}^T) + \lambda \sum_i \sum_j \frac{1}{(\alpha_j^{(i)})^2} \tag{6.6}$$

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. (6.6) 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.

6.2.1 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

$$\mathcal{L}_{multi} = \frac{C}{2}\ln|\bar{\mathbf{K}}| + \sum_{c=1}^{C}\frac{1}{2}tr(\bar{\mathbf{K}}^{-1}\mathbf{Y}^{(c)}\mathbf{Y}^{(c),T}) + \lambda C\sum_i\sum_j\frac{1}{(\alpha_j^{(i)})^2} \quad (6.7)$$

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

6.2.2 Inference

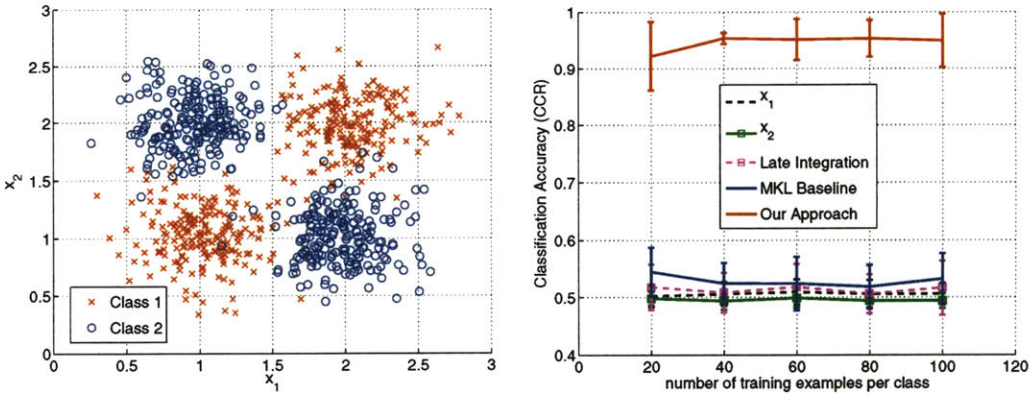
The mean prediction is an estimator of the distance to the margin, and thus one can choose the label for each test data point as the one with the largest mean prediction among all the 1-vs-all classifiers

$$\bar{\mathbf{y}}_* = \max_c\{\bar{\mathbf{k}}_*^T\bar{\mathbf{K}}^{-1}\mathbf{Y}^{(c)}\} \quad (6.8)$$

where $\bar{\mathbf{k}}_*$ is the kernel computed between the training and test data using Eq. (6.5). 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.

6.3 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.,



(a) (b)

Figure 6-1: **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 6-5).

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., 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 [130]. 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 $\sigma^2 = 0.01$.

6.3.1 Synthetic example

We first consider the two-view, two-class synthetic example depicted in Figure 6-1(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 6-1(a), each distribution having a std. deviation of 0.25 and means $(1,1), (1,2), (2,1), (2,2)$ respectively. Figure 6-1(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 (see Figure 6-5 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.

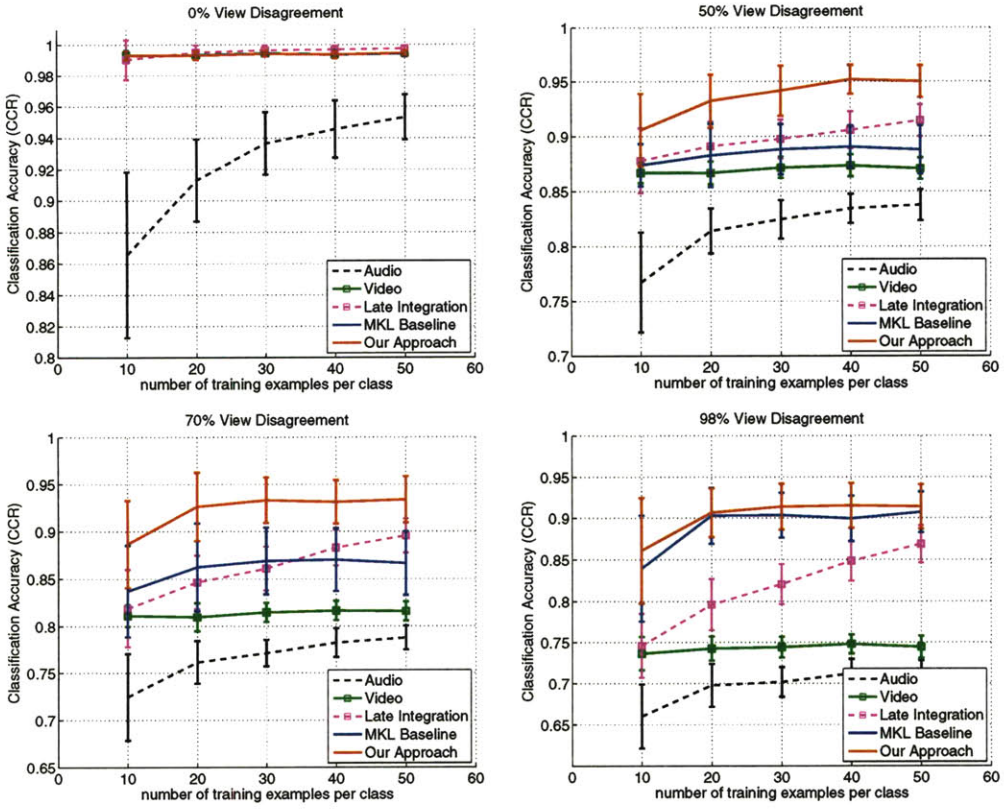


Figure 6-2: **Audio-visual user agreement experiments.** The performance of our approach is shown along with the baseline approaches averaged over 50 splits as a function of the number of training samples per class, error bars indicate ± 1 std. deviation. Unlike the baseline methods, our approach is able to achieve accurate classification performance despite the per-sample view corruption.

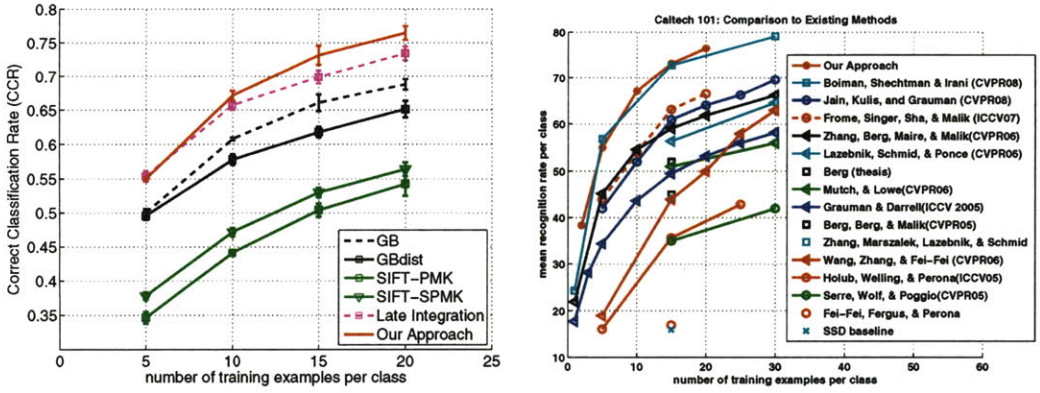
6.3.2 Audio-visual user agreement in the present of view disagreement

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 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 [21]. 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.

[21], 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 [46]. 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 6-2 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 [21], 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,



(a) (b)

Figure 6-3: **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.

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 6-5.

6.3.3 Object recognition

Finally we evaluate our approach on the Caltech-101 benchmark that is comprised of images from 101 object categories [35]. We use four different image features. For the first two feature types we used the geometric blur features described in [131]. 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 [44] and spatial PMK [60] 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 6-3(a) displays the performance of our approach with $P = 6$ averaged over 5 splits for varying number of training examples per-class, Figure 6-5(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 6-3(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 [12], with a Gaussian Process used in place of the naive Bayes nearest-neighbor classifier in each view. Moreover, the method in [12] 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 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 [104]. In Figure 6-4(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 6-4(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 advan-

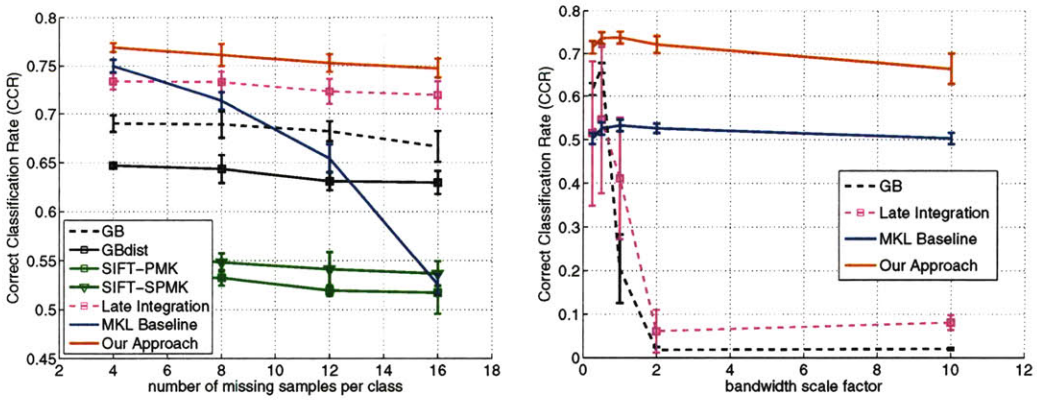


Figure 6-4: **Caltech-101 with missing data.** (a) Unlike conventional kernel combination our approach can take advantage of partially observed multi-view samples. (b) Late integration is sensitive to bandwidth selection. The performance of our approach is relatively un-affected by the corrupted view 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.

tage 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 6-4(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 6-4, both the late integration baseline and our approach assume an equal weighting over the

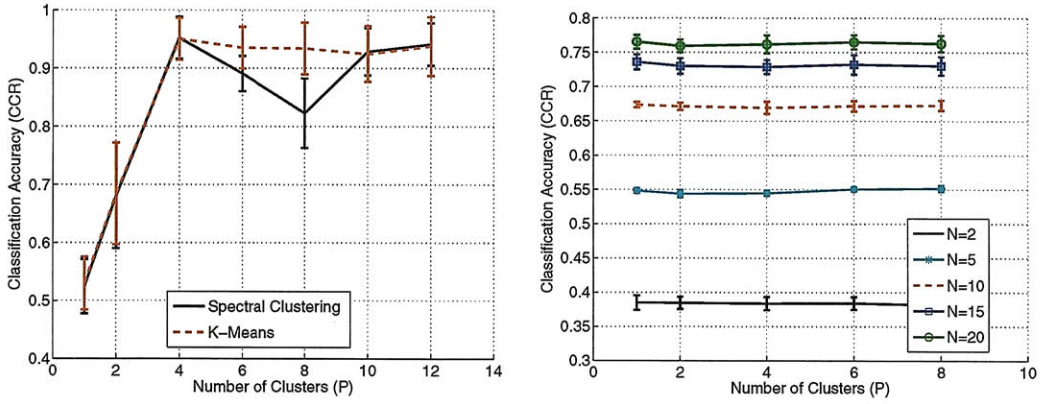


Figure 6-5: **Influence of the number of clusters**. Average performance is shown for each dataset, error bars indicate ± 1 std. deviation. (top) Influence on synthetic dataset. Performance is averaged over 50 splits with $N = 60$ and the rest test. Performance with spectral and k-means clustering is 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. (bottom) 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.

views. The performance of the 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.

6.3.4 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 6-5 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.

6.4 Chapter Summary

In this chapter 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. Avenues of future work include the use of soft clustering as well as the application of our approach to other domains, e.g., pose estimation.

This chapter concludes our work on supervised and semi-supervised multi-view learning approaches. In the following chapter we consider the separate, but related problem of unsupervised multi-view feature selection for classification from multiple sensors. In particular, we consider a distributed feature selection algorithm with GPs useful for multi-sensor classification in distributed networks, where bandwidth is limited and communication between sensors prohibitively expensive. We demonstrate our approach on the task of distributed compression of visual features for multi-view object recognition and provide an evaluation of our method on a publically available multi-view object database.

Chapter 7

Unsupervised Visual Feature Selection via Distributed Coding

As the previous chapters focused on multi-view semi-supervised learning, this chapter focuses on a separate, yet related problem in multi-sensor classification[1]. Feature selection is an important problem in machine learning that has close connections with data compression techniques in information theory. These techniques seek to find compact, informative feature representations for performing classification. In the case of multi-sensor data, feature selection is of particular importance since the data from multiple sources is often high dimensional and highly redundant, and feature selection plays a central in the development of efficient and accurate classification and learning algorithms in multiple sensor systems.

In this chapter, we consider the specific problem of object recognition from multiple cameras belonging to a distributed network, where the computation at each sensor is limited and communication between sensors is prohibitively expensive due to bandwidth constraints. For this setting, we develop a distributed feature selection algorithm with Gaussian Processes (GPs) borrowing concepts from distributed source coding in the information theory literature. We evaluate our approach both on synthetic and real-world datasets, and achieve high distributed compression rates

[1]The work described in this chapter is published in the Conference on Computer Vision and Pattern Recognition, 2008, Christoudias, Urtasun and Darrell [22].

while maintaining accurate multi-view recognition performance.

7.1 Introduction

Object recognition often benefits from integration of observations at multiple views. Contemporary methods for object recognition use local feature representations and perform recognition over sets of local features corresponding to each image [62, 80, 45, 109]. Several techniques have been proposed that generalize these methods to include object view-point in addition to appearance [88, 108, 109, 92]. Rothganger et. al. [88] present an approach that builds an explicit 3D model from local affine-invariant image features and uses that model to perform view-point invariant object recognition. Thomas et. al. [108] extend the Implicit Shape Model (ISM) of Leibe and Schiele [62] for single-view object recognition to multiple views by combining the ISM model with the recognition approach of Ferrari et. al. [36]. Similarly, Savarese and Li [92] present a part-based approach for multi-view recognition that jointly models object view-point and appearance.

Traditionally, approaches to multi-view object recognition use only a single input image at test time [88, 108, 109, 92]. Recently, there has been a growing interest in application areas where multiple input views of the object or scene are available. The presence of multiple views can lead to increased recognition performance; however, the transmission of data from multiple cameras places an additional burden on the network. When multiple camera sensors exist in a bandwidth limited environment it may be impossible to transmit all the visual features in each image. Also, when the task or target class is not known *a priori* there may be no obvious way to decide which features to send from each view. If redundant features are chosen at the expense of informative features, performance can be worse with multiple views than with a single view, given fixed bandwidth.

We consider the problem of how to select which features to send in each view to achieve optimal results at a centralized recognition or indexing module (see Figure 7-1). An efficient encoding of the streams might be possible in theory if a class label

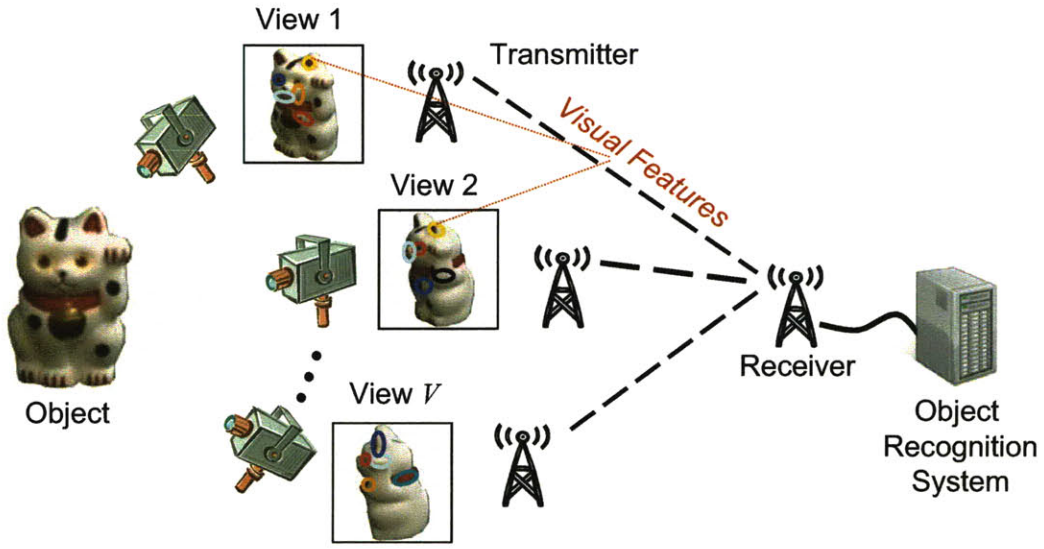


Figure 7-1: Distributed object recognition. Messages are only sent between each camera (transmitter) and the recognition module (receiver). An efficient joint feature selection is achieved *without* directly sharing information between cameras.

could be inferred at each camera, enabling the use of supervised feature selection techniques to encode and send only those features that are relevant of that class. Partial occlusions, unknown camera viewpoint, and limited computational power, however, limit the ability to reliably estimate the image class label at each camera. Instead we propose an unsupervised feature selection algorithm to obtain an efficient encoding of the feature streams.

If each camera sensor had access to the information from all views this could trivially be accomplished by a joint compression algorithm that could, e.g., encode the features of the v-th view based on the information in the previous $v-1$ views. We are interested, however, in the case where there is *no* communication between cameras themselves, and messages are only sent from the cameras to the recognition module with a limited back-channel back to the cameras. In practice, many visual category recognition and indexing applications are bandwidth constrained (e.g., wireless surveillance camera networks, mobile robot swarms, mobile phone cameras), and it is infeasible to broadcast images across all cameras or to send the raw signal from each camera to the recognition module.

It is possible to achieve very efficient encoding without any information exchange between the cameras, by adopting a distributed encoding scheme that takes advantage of known statistics of the environment [84, 103, 93, 42]. We develop a new method for distributed encoding based on a Gaussian Process (GP) formulation, and demonstrate its applicability to encoding visual-word feature histograms; such representations are used in many contemporary object indexing and category recognition methods [102, 80, 45]. Our approach bears similarity to that of Kapoor et. al. [52] that use GP prediction uncertainty as a criteria for example selection in active learning.

With our algorithm a statistical model of the dependency between feature streams is learned during an off-line training phase at the receiver. This model is then used along with previously decoded streams to aid feature selection at each camera and if the streams are redundant, then only a few features need to be sent. In this chapter, we consider bag-of-words representations [80, 45] and model the dependency between visual feature histograms. As shown in our experiments, our algorithm is able to achieve an efficient joint encoding of the feature histograms without explicitly sharing features across views. This results in an efficient unsupervised feature selection algorithm that improves recognition performance in the presence of limited network bandwidth.

We evaluate our approach using the COIL-100 multi-view image database [77] on the tasks of instance-level retrieval and recognition from multiple views; we compare unsupervised distributed feature selection to independent stream encoding. For a two-view problem, our algorithm achieves a compression factor of over 100:1 in the second view while preserving multi-view recognition and retrieval accuracy. In contrast, independent encoding at the same rate does not improve over single-view performance.

7.2 Distributed Object Recognition

We consider the distributed recognition problem of V cameras transmitting information to a central common receiver with no direct communication between cameras (see Figure 7-1). In our problem, each camera is equipped with a simple encoder used

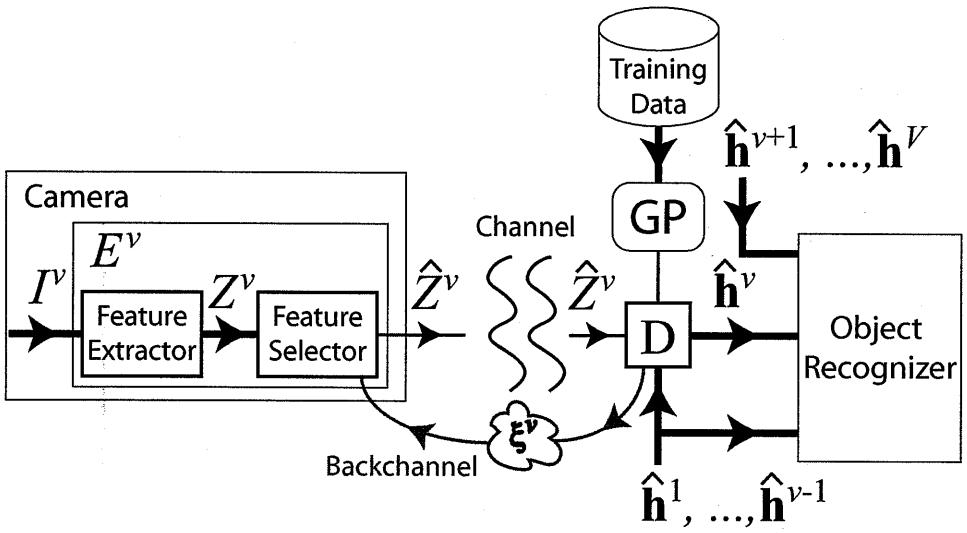


Figure 7-2: System diagram. Image I^v is coded by encoder E^v and decoder D. $\hat{Z}^v$ are the encoded image features, $\hat{\mathbf{h}}^v$ the reconstructed histograms, and ξ^v the non-redundant bin indices for views $v = 1, ..., V$ (see Section 7.2 for details).

to compress each signal before transmission. A common decoder receives the encoded signals and performs a joint decoding of the signal streams using a model of the joint statistics. Note that this coding scheme off-loads the computational burden onto the decoder and allows for computationally in-expensive encoders. In what follows, we assume a noiseless channel, but our approach is also applicable to the more general case.

Figure 7-2 illustrates our proposed distributed coding algorithm at a single camera. With our algorithm, the decoder iteratively queries each of the V cameras and specifies the desired encoding rate the camera should use. At the v-th view, the decoder uses its model of joint statistics along with *side information*, i.e., the previously decoded streams, to decode the signal. The use of side information allows the encoder to work at a lower encoding rate than if the stream were encoded independently. As discussed below, the decoder selects the camera encoding rate based on the joint stream statistics and transmits this information back to the encoder. If the v-th view is highly redundant with respect to the side information, then little-to-no information needs to be encoded and sent to the decoder.

In this work, we consider bag-of-words models for object recognition [80, 45]. With these models, an image is represented using a set of local image descriptors extracted from the image either at a set of interest point locations (e.g., those computed using a Harris point detector [73]) or on a regular grid. In our experiments, we employ the latter feature detection strategy in favor of simplicity at the encoder. To perform feature coding, the local image features are quantized using a global vocabulary that is shared by the encoder and decoder and computed from training images.

Let I^v, $v = 1, ..., V$ be a collection of V views of the object or scene of interest, imaged by each camera and Z^v be the set of quantized local image features corresponding to image I^v computed by the v-th encoder, E^v. In this context (see Figure 7-2), the encoders transmit quantized features to the central receiver and the encoding rate is the number of features sent.

In theory, distributed coding with individual image features (e.g., visual words) might be possible, but preliminary experiments have shown that distributed coding of local features does not improve over independent encoding at each camera. Using a joint model over quantized features on COIL-100 with a 991 word vocabulary gave an entropy of 9.4 bits, which indicates that the joint feature distribution is close to uniform (for a 991 word feature vocabulary, the uniform distribution has an entropy of 10 bits). This is expected since a local image feature is a fairly weak predictor of other features in the image.

We have found, however, that distributed coding of histograms of local features is effective. As seen in our experiments, the distribution over features in one view is predictive of the distribution of features in other views and, therefore, feature histograms are a useful image representation for distributed coding.

7.2.1 Joint Feature Histogram Model

Let $\hat{Z}^v$ be the set of encoded features of each view, $v = 1, ..., V$. To facilitate a joint decoding of the feature streams, the decoder first computes a feature histogram, $\hat{\mathbf{h}}^v = h(\hat{Z}^v)$, using the global feature vocabulary. Note, in our approach, the form of $h(\cdot)$ can either be a flat [102] or hierarchical histogram [80, 45]; we present a

general distributed coding approach applicable to any bag-of-words technique. At the decoder, the joint stream statistics are expressed over feature histograms,

$$p(\mathbf{h}^1, \mathbf{h}^2, ..., \mathbf{h}^V) = p(\mathbf{h}^1) \prod_{v=2}^{V} p(\mathbf{h}^v | \mathbf{h}^{v-1}, ..., \mathbf{h}^1), \tag{7.1}$$

where the conditional probabilities are learned from training data as described in Section 2.2. In what follows, without loss of generality, we assume that the histograms $\mathbf{h}^v$ are normalized to sum to 1, such that $\|\mathbf{h}^v\|_1 = 1$, and regress to a continuous-valued output space, although a discrete-valued output space could also be modeled within our framework and we leave this as a topic for future work.

Assuming independence between the histogram bins and pair-wise dependence between histograms we write

$$p(\mathbf{h}^v | \mathbf{h}^{v-1}, ..., \mathbf{h}^1) = \prod_{k=1}^{v-1} \prod_{b=1}^{B} p(h^{v,b} | \mathbf{h}^k) \tag{7.2}$$

where $h^{v,b}$ is the b-th bin of histogram $\mathbf{h}^v$, and B is the number of bins.

The joint model of Equation 7.1 is used to determine which features at a given camera are redundant with the side information. In particular, redundant features are those that are associated with the redundant bins of the histogram of the current view. Since we are ultimately interested in the feature histograms for performing recognition, the encoders can send either histogram bin counts or the quantized visual features themselves.

We obtain a reconstruction of the feature histogram of each view from the view's encoded features and its side information. Let $\mathbf{h}^v$ be the histogram of interest and $\mathbf{h}^k$, $k = 1, ..., v-1$, its side information, where v is the current view considered by the decoder. From Equation 7.2 the probability of a histogram $\mathbf{h}^v$ given its side information is found as,

$$p(\mathbf{h}^v | \mathbf{h}^{v-1}, ..., \mathbf{h}^1) = \prod_{k=1}^{v-1} p(\mathbf{h}^v | \mathbf{h}^k) \tag{7.3}$$

Algorithm 5 GP Distributed Feature Selection

Let E^v be an encoder, ξ^v be defined over sets of feature histogram bin indices, $\hat{Z}^v$ be defined over sets of encoded features, $\mathbf{H}^v$ be a $N \times B$ matrix of N training examples, $v = 1, ..., V$, and $R_{\max}$ be the desired encoding rate.

$\hat{\mathbf{h}} = \emptyset$
$\xi^1 = \{1, ..., B\}$
for $v = 1, ..., V$ **do**
  $\hat{Z}^v = \texttt{request}(E^v, \xi^v)$
  for $b = 1, ..., B$ **do**
    if $b \in \xi^v$ **then**
      $\hat{h}^{v,b} = \psi(\hat{Z}^{v,b})$
    else
      $\hat{h}^{v,b} = (\mathbf{k}_*^{v-1,b})^T (\mathbf{K}^{v-1,b})^{-1} \mathbf{H}^{v,b}$
    end if
  end for
  $\hat{\mathbf{h}} = (\hat{\mathbf{h}}, \hat{\mathbf{h}}^v)$
  if $v < V$ **then**
    for $b = 1, ..., B$ **do**
      $\sigma^{v+1,b} = (k^{v,b}(\hat{\mathbf{h}}, \hat{\mathbf{h}}) - (\mathbf{k}_*^{v,b})^{\mathrm{T}} (\mathbf{K}^{v,b})^{-1} \mathbf{k}_*^{v,b})^{\frac{1}{2}}$
    end for
    $\xi^{v+1} = \texttt{select}(\boldsymbol{\sigma}^{v+1}, R_{\max})$
  end if
end for

We model the above conditional probability using a GP prior over feature histograms. To make learning more tractable we assume independence between histogram bins

$$p(\mathbf{h}^v | \mathbf{h}^{v-1}, ..., \mathbf{h}^1) = \prod_{b=1}^{B} \mathcal{N}(0, \mathbf{K}^{v-1,b}) \tag{7.4}$$

where a GP is defined over each bin with kernel matrix $\mathbf{K}^{v-1,b}$. We compute $\mathbf{K}^{v-1,b}$ with a covariance function defined using an exponential kernel over the side information,

$$k^{v,b}(\mathbf{h}_i, \mathbf{h}_j) = \prod_{r=1}^{v} \gamma_b^{-v} \exp\left(\frac{d(\mathbf{h}_i^r, \mathbf{h}_j^r)^2}{\alpha_b^2}\right) + \eta_b \delta_{ij} \tag{7.5}$$

where $\hat{\mathbf{h}}_i = (\hat{\mathbf{h}}_i^1, ..., \hat{\mathbf{h}}_i^v)$ and $\hat{\mathbf{h}}_j = (\hat{\mathbf{h}}_j^1, ..., \hat{\mathbf{h}}_j^v)$ are multi-view histogram instances, γ_b, α_b are the kernel hyper-parameters of bin b, which we assume to be the same across views, and η_b is a per-bin additive noise term. Given training data $\mathbf{H}^v$, where

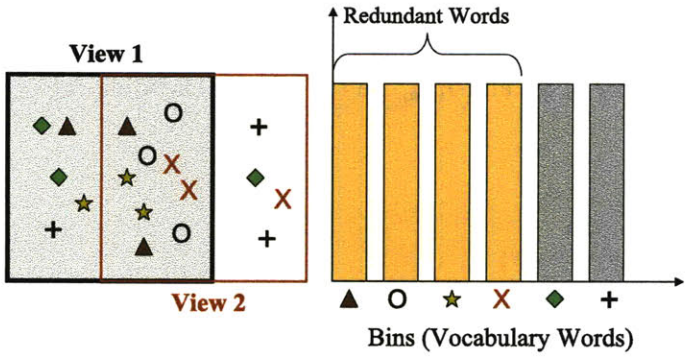


Figure 7-3: Synthetic example considered below. This scenario consists of two overlapping views of an object, which is presumed to fill the scene. Image features are represented using a 6 word vocabulary.

$\mathbf{H}^v$ is a $N \times B$ matrix of N training examples for the $v = 1, ..., V$ views, the kernel hyper-parameters are learned as described in Section 2.2. We define a different set of kernel hyper-parameters per bin since each bin can exhibit drastically different behavior with respect to the side information.

The variance of each GP can be used to determine whether a bin is redundant: a small bin variance indicates that the GP model is confident in its prediction, and therefore the features corresponding to that bin are likely to be redundant with respect to the side information. In our experiments, we found that redundant bins generally exhibit variances that are small and similar in value and that these variances are much smaller than those of non-redundant bins.

7.2.2 GP Distributed Feature Selection

Distributed feature selection is performed by the decoder using an iterative process. The decoder begins by querying the first encoder to send all of its features, since in the absence of any side information no feature is redundant. At the v-th view, the decoder requests only those features corresponding to the non-redundant histogram bins of that view, whose indices are found using the bin variances output by each GP. At each iteration, the GPs are evaluated using the reconstructed histograms of

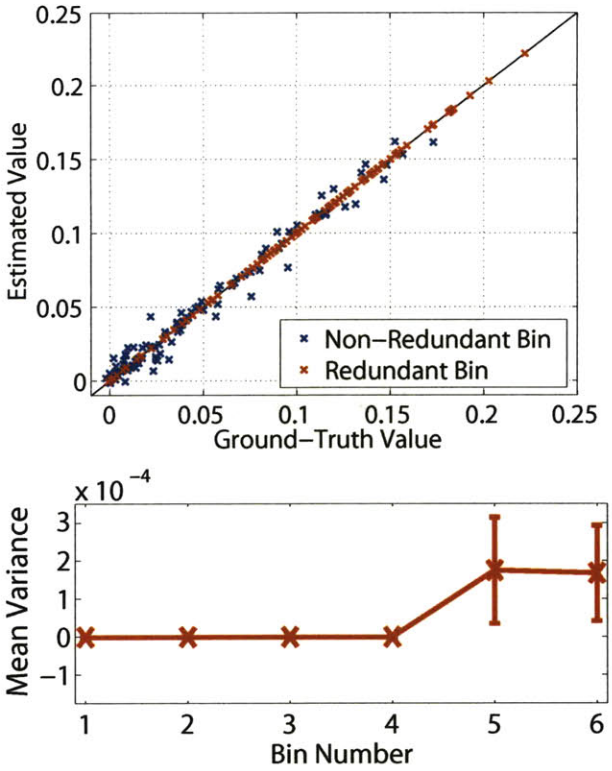


Figure 7-4: GP variance is correlated with bin redundancy. The GP mean prediction for the second view is plotted vs. ground-truth values for both a redundant and non-redundant bin. The GP variance for each of the 6 histogram bins, averaged across examples is also shown; error bars indicate ±1 std. deviation. The variance of non-redundant bins is noticeably higher than that of redundant bins.

previous iterations as illustrated in Algorithm 1.

Given the encoded features $\hat{Z}^v$, the decoder reconstructs histograms $\hat{\mathbf{h}}^v$, $v = 1, ..., V$, such that bins that are non-redundant are those received and the redundant bins are estimated from the GP mean prediction

$$\hat{h}^{v,b} = \begin{cases} h(\hat{Z}^{v,b}), & b \in \xi^v \\ (k_*^{v-1,b})^{\mathrm{T}}(\mathbf{K}^{v-1,b})^{-1}\mathbf{H}^{v,b}, & \text{otherwise.} \end{cases} \tag{7.6}$$

where $\mathbf{H}^{v,b} = (h_1^{v,b}, ..., h_N^{v,b})^T$ are the bin values for view v and bin b in the training data, and ξ^v are the bin indices of the non-redundant bins of the histogram of view v.

The GP distributed feature selection algorithm achieves a compression rate proportional to the number of bin indices requested for each view. For view v the compression rate of our algorithm in percent bins transmitted is

$$R = \frac{r}{B} = \frac{2|\xi^v|}{B}, \tag{7.7}$$

where B is the total number of histogram bins and r is the number of bins received, which is proportional to twice the number of non-redundant bins as a result of the decoder `request` operation. Note, however, that in the case of large amounts of redundancy there are few non-redundant bins encoded at each view and therefore a small encoding rate is achieved.

As mentioned above the bin indices ξ^v are chosen using the GP prediction uncertainty. If a desired encoding rate $R_{\max}$ is provided, the decoder requests the $r_{\max}/2$ histogram bins associated with the highest GP prediction uncertainty (see Equation 7.7). If $R_{\max}$ is not known, the encoding rate can be automatically determined by grouping the histogram bins at each view into two groups corresponding to regions of high and low uncertainty; ξ^v is then defined using the bins associated with the high uncertainty group. Both strategies exploit the property that prediction uncertainty is correlated with bin redundancy to request the non-redundant bins at each view. Many grouping algorithms are applicable for the latter approach, e.g., conventional clustering. In practice, we use a simple step detection technique to form each group by sorting the bin variances and finding the maximum local difference.

7.3 Experiments

We evaluate our distributed coding approach on the tasks of object recognition and indexing from multiple views. Given $\mathbf{h}^v$, $v = 1, .., V$, multi-view recognition is performed using a nearest-neighbor classifier over a fused distance measure, computed

as the average distance across views

$$D_{i,j}(\mathbf{h}_i, \mathbf{h}_j) = \frac{1}{V} \sum_{v=1}^{V} d(\mathbf{h}_i^v, \mathbf{h}_j^v) \tag{7.8}$$

where for flat histograms we define $d(\cdot)$ using the L_2 norm, and with pyramid match similarity [45] for multi-resolution histograms[2].

We use a majority vote performance metric for nearest-neighbor recognition. Under this metric a query example is correctly recognized if a majority ($\geq k/2$) of its k nearest-neighbors are of the same category or instance. We also experiment with an at-least-one criterion to evaluate performance in an interactive retrieval setting: with this scheme an example is correctly retrieved if one of the first k examples has the true label. We compare distributed coding to independent encoding at each view with a random feature selector that randomly selects histogram bins according to a uniform distribution, and report feature selection performance in terms of percent bins encoded, R (see Equation 7.7).

In what follows, we first present experiments on a synthetic example with our approach and then discuss our results on COIL-100.

7.3.1 Synthetic Example

To demonstrate our distributed feature selection approach we consider the scenario illustrated in Figure 7-3. An object is imaged with two overlapping views, and the histograms of each view are represented using a 6 word vocabulary. As shown by the figure, the images are redundant in 4 of the 6 words, as 2 of the words (i.e., diamond and plus) do not appear in the overlapping portion of each view. Although real-world problems are much more complex than described above, we use this simple scenario to give intuition and motivate our approach.

We first consider the case where there is no noise between the redundant features in each view and the redundant features appear only in the overlapping region. We

[2]Note our distributed coding algorithm is independent of the choice of classification method.

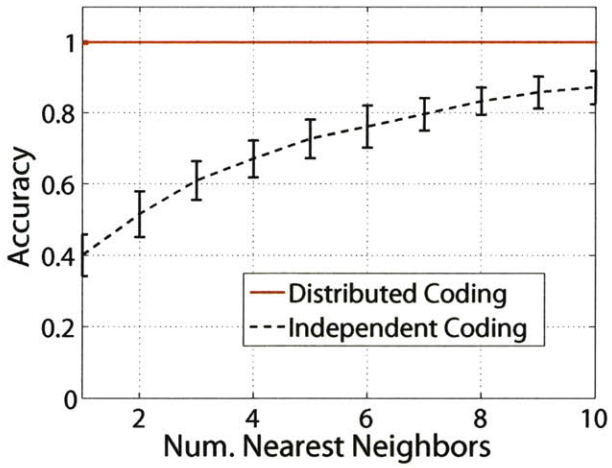


Figure 7-5: Nearest-neighbor instance-level retrieval for the two-view synthetic dataset; average retrieval accuracy is plotted over varying neighborhood sizes. For a fixed rate, our algorithm far outperforms the independent encoding baseline (see text for details).

randomly generated $N = 100$, 6-D histograms, where each histogram was generated by sampling its bins from a uniform distribution between 0 and 1, and the histograms were normalized to sum to one. Each histogram was split into two views by replicating the first 4 bins in each view and randomly splitting the other two bins. The above data was used to form a training set of examples, where each pair of histograms corresponds to a single object instance. To form the test set, zero mean Gaussian noise was added to the training set with $\sigma = 0.01$ and the test set histograms were split into two views using the same split ratios as the training set.

For distributed coding we trained 6 GPs, one per dimension, using each view. Figure 7-4 displays the predicted bin value vs. ground truth for 2 of the bins (one redundant and the other non-redundant) evaluated on the second view of the test set. The GPs are able to learn the deterministic mapping that relates the redundant bins. For the non-redundant bin, the variance of the GP's predictions is quite large compared to that of the redundant bin. Also shown in Figure 7-4, are the mean GP variances plotted for each histogram bin. The error bars in the plot indicate the standard deviation. The GP variance is much larger for the non-redundant bins than those of the redundant ones whose variances are small and centered about 0. This is

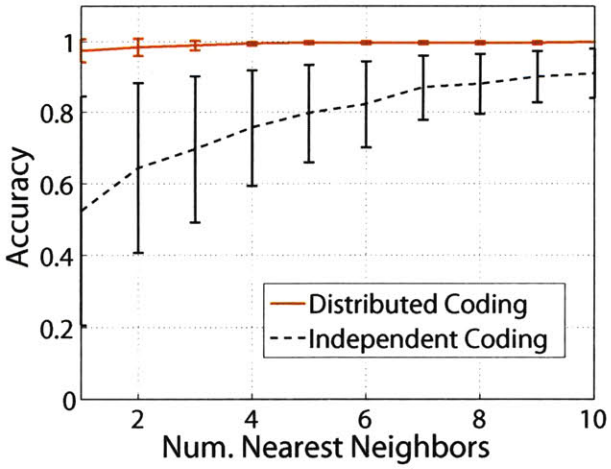


Figure 7-6: Nearest-neighbor instance-level retrieval on the two-view synthetic dataset with partial redundancy, plotted over varying neighborhood sizes. Our distributed coding algorithm performs favorably to independent encoding even when the bins are only partially redundant.

expected since non-redundant bins are less correlated and therefore the GPs are less certain in their prediction of the value of these bins from side information.

Evaluating our distributed coding algorithm on the above problem gave a bin rate of $R = 0.66$ in the second view. Figure 7-5 displays the result of nearest-neighbor instance-level retrieval over each of the 100 instances in the training set for varying neighborhood sizes. The average retrieval accuracy, averaged over 10 independent trials, is shown for both distributed and independent coding of the second view, where for independent coding features were selected at the same rate as distributed coding. Distributed coding far outperforms independent encoding in the above scenario.

We also considered the case of partial redundancy, where the redundant bins are only partially correlated as a result of noise. To simulate partial redundancy we added zero mean Gaussian noise to the split ratios of the first 4 bins with $\sigma = \{0, 0.01, 0.05, 0.1\}$. Figure 7-6 displays the result of nearest-neighbor recognition with distributed and independent coding of the second view. In the plot, recognition performance is reported, averaged across the different σ values, along with error bars indicating the standard deviation. For this experiment, an average bin rate of $R = 0.78 \pm 0.23$ was achieved with our distributed feature selection algorithm. Our

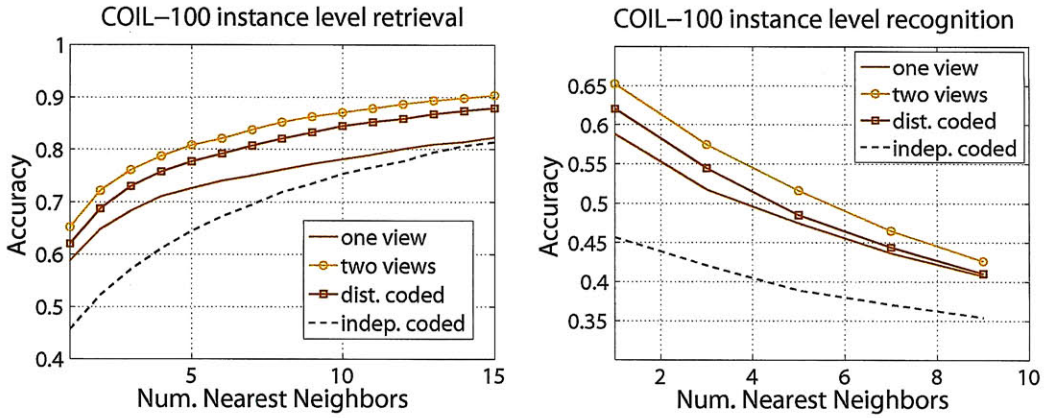


Figure 7-7: Nearest-neighbor (top) retrieval and (bottom) recognition with two-views on COIL-100. Our algorithm performs significantly better over single view performance under each task while achieving a very low encoding rate. For the retrieval task, our approach performs near multi-view performance. The independent encoding baseline is also shown, where independent feature selection was performed at the same rate as our algorithm. Note that independent encoding with two views does worse than a single view when operating at such a low encoding rate.

distributed coding algorithm can perform favorably to independent encoding even when the bins are only partially redundant.

7.3.2 COIL-100 Experiments

We evaluated our distributed feature selection algorithm using the COIL-100 multi-view object database [77] that consists of 72 views of 100 objects viewed from 0 to 360 degrees in 5 degree increments. A local feature representation is computed for each image using 10 dimensional PCA-SIFT features [54] extracted on a regular grid using a 4 pixel spacing. We evaluate our distributed coding algorithm and perform recognition with the COIL-100 dataset using multi-resolution vocabulary-guided histograms [45] computed with LIBPMK [61]. We split the COIL-100 dataset into train and test sets by taking alternating views of each object. We then paired images 50 degrees apart to form the two views of our problem.

Using the training image features we perform hierarchical k-means clustering to compute the vocabulary used to form the multi-resolution pyramid representation. Using 4 levels and a tree branch factor of 10 gave a 991 word vocabulary at the finest

level of the hierarchy. GP distributed feature selection is performed over the finest level of the histogram, such that the encoders and decoder only communicate bins at this level. The upper levels of the tree are then recomputed from the bottom level when performing recognition. To perform GP distributed coding we used a kernel defined using L2 distance over a coarse, flat histogram representation.

Figure 7-7 displays nearest-neighbor retrieval and recognition accuracy using one and two views. A significant performance increase is achieved by using the second view when there are no bandwidth constraints. Applying GP distributed feature selection on the above dataset resulted in a bin rate of $R < 0.01$ in the second view; this is a compression rate of over 100:1. Figure 7-7 displays the performance of our GP distributed feature selection algorithm. By exploiting feature redundancy across views, our algorithm is able to perform significantly better than single view performance while achieving a very low encoding rate. The result of independent encoding is also shown in the Figure, where independent feature selection was performed at the same rate as our algorithm. In contrast to our approach, independent encoding is not able to improve over single-view performance and in fact does worse at such low encoding rates.

We also tested our approach over different encoding rates, where the desired rate is provided as input to the algorithm. Figure 7-8 displays the nearest-neighbor performance of our approach over different encoding rates. As expected, nearest-neighbor performance increases for larger encoding rates. Performance saturates at about $r = 50$ bins and remains fairly constant for larger rates. Of coarse, for $r = B$ one would expect to recover ground-truth performance. The slow convergence rate of our approach to ground-truth performance with increasing encoding rate suggests the need for better bin selection criteria, which we plan to investigate as part of future work. The independent encoding baseline is also shown. Recall that at rate R the baseline approach transmits twice the number of bins as our approach as a result of the request operation. Independent encoding needs to transmit nearly the entire histogram ($|\xi^2| = 400$) before reaching a recognition performance close to our approach. Our approach achieves similar performance with only $|\xi^2| = 10$.

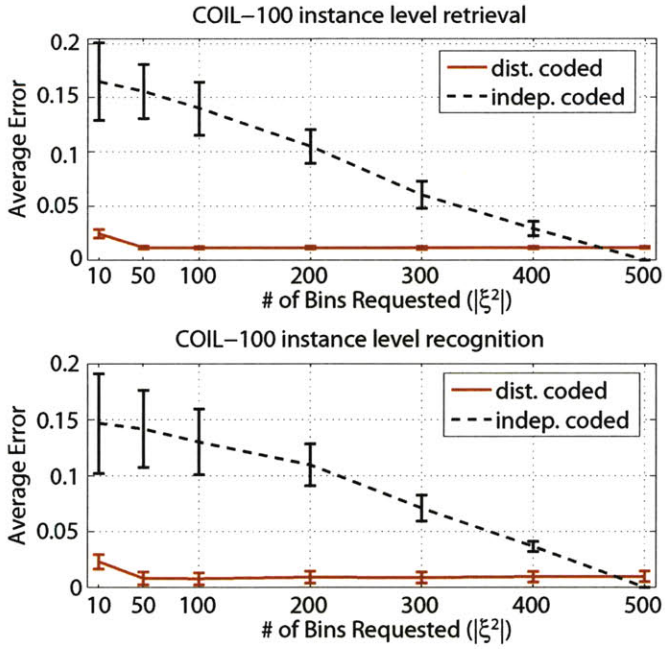


Figure 7-8: Nearest-neighbor performance increases with encoding rate. Nearest-neighbor performance is shown for the tasks of (top) retrieval and (bottom) recognition. The accuracy difference between our approach and ground-truth two-view performance is shown averaged over neighborhood size; error bars indicate ± 1 std. deviation. The independent encoding baseline is also shown.

7.4 Chapter Summary

In this chapter we presented a distributed coding method for unsupervised distributed feature selection and showed its application to multi-view object recognition. We developed a new algorithm for distributed coding with Gaussian Processes and demonstrated its effectiveness for encoding visual word feature histograms on both synthetic and real-world datasets. For a two-view problem with COIL-100, our algorithm was able to achieve a compression rate of over 100:1 in the second view, while significantly increasing accuracy over single-view performance. At the same coding rate, independent encoding was unable to improve over recognition with a single-view. For future work, we plan to investigate techniques for modeling more complex dependencies as well as one-to-many mappings between views and evaluate our approach under different bin selection criteria.

The method pursued in this chapter is unsupervised in the class label, however, still requires a training stage of a joint feature model that can be difficult to learn and can perform poorly in the face of limited training data. Approaches in compressed sensing provide an alternative solution to the distributed coding problem, in which no model of joint feature statistics is required. Instead, features are compressed using random projections and are reconstructed after transmission by exploiting joint sparsity between the feature patterns in each view.

In recent work [124], we develop a compressed sensing approach that exploits joint sparsity to perform distributed coding for multi-view object recognition and achieves accurate recognition performance at relatively low rates, without learning a joint model. Still, when training data is available these methods can also benefit from added information that can be used to help regularize the solution, e.g., by learning a latent space used in place of random projection or by adding a regularization term to the optimization process that favors similarity between the obtained solution and the training samples. The development of such approaches is an interesting avenue for future work.

Chapter 8

Conclusion

8.1 Summary

In this thesis, we considered the problem of classification and learning from multiple noisy sensors in perceptual learning tasks, including speech and gesture interfaces and multi-view object recognition. In particular, this thesis focused on the problems of learning from labeled and unlabeled data by exploiting multiple, potentially redundant input sources, performing multi-source model adaptation to automatically adapt a system to a new user or working condition, and for performing distributed feature selection for classification from multi-source, bandwidth limited distributed networks. Challenges in each of these areas include coping with sensor noise or view insufficiencies, and forming appropriate models of joint sensor statistics for performing multi-view learning and feature selection. This thesis proposed novel probabilistic modeling approaches built upon multi-view Gaussian Processes for coping with each of these challenges.

Multi-modal classification is well suited for multi-view learning because each modality provides a potentially redundant view to the learning algorithm. While the concept of multi-modal co-training was mentioned as promising future work in the seminal Blum and Mitchell paper, it appears that there has been relatively little subsequent work on cross-modal co-training. The first part of this thesis investigated the use of co-training for learning audio-visual speech and gesture classifiers in human-

computer interfaces and developed a novel multi-view model adaptation algorithm, co-adaptation, that adapts a generic model to a new user or working condition. Both co-training and co-adaptation were evaluated on the tasks of audio-visual user agreement classification from speech and gesture and audio-visual speech recognition. On both tasks improved performance was demonstrated with co-training by exploiting both labeled and unlabeled data and/or by adapting to a new user or environment.

The second part of this thesis focused on the problem of view disagreement—i.e., when a sample belongs to a different class than the samples in the other views as a result of view corruption or noise—and other forms of view insufficiency caused by view corruption common to multi-sensor perceptual learning tasks, such as occlusion, uni-modal expression, and other forms of missing data and complex per-sample noise processes. We proposed a filter-based co-training algorithm that utilizes an information theoretic criterion to detect and filter view disagreement samples during co-training. Our experiments on audio-visual user agreement classification demonstrate that unlike existing techniques, our filter-based co-training algorithm is able to learn accurate multi-view classifiers despite view disagreement.

The filter-based co-training approach was tailored to modeling view disagreement, a particular kind of binary per-sample view corruption. In the more general setting, however, the per-sample noise is continuous with certain samples being more corrupted than others; also a sample can still be informative even if it is corrupted by noise, and a filter-based approach would in effect ignore such samples. This thesis develops a probabilistic co-training framework that extends Bayesian co-training to model per-sample noise and other complex noise processes. Our heteroscedastic Bayesian co-training approach simultaneously discovers the per-sample noise while solving the classification task and can handle arbitrary view corruption processes including binary view disagreement. We demonstrate our approach on the tasks of audio-visual user agreement and multi-view object classification and show improved performance compared to our filter-based co-training approach other state-of-the-art techniques. For the related problem of supervised multi-view learning, we also proposed a Bayesian multiple kernel learning method capable of learning a local weighting

over the input space, and showed how under this approach the noise indicator matrices assumed by our heteroscedastic Bayesian co-training algorithm can be computed in an unsupervised fashion using data clustering techniques.

Finally, this thesis investigated the problem of unsupervised feature selection for classification from multiple sensors on a bandwidth-limited distributed network, where communication between sensors is prohibitively expensive. We developed a distributed feature selection algorithm with Gaussian Processes borrowing concepts from distributed source coding in information theory. Our GP-based distributed feature selection algorithm was demonstrated on the task of visual feature selection for multi-view object classification from a distributed multi-camera system. Our approach was evaluated on both synthetic and real-world datasets, and achieves high distributed compression rates while maintaining accurate multi-view recognition performance without explicitly sharing information between the different camera sensors.

8.2 Future Work

There are many interesting extensions and avenues for future work concerning the challenges in multi-view learning and feature selection addressed in this thesis.

Co-training seems naturally suited for the task of semi-supervised learning in multi-sensor classification systems. Yet, it has received limited application in these domains. This thesis demonstrated the successful application of co-training for the tasks of audio-visual speech and gesture interfaces and multi-view object recognition, however, the potential of co-training for generally solving semi-supervised learning problems in multi-sensory domains has yet to be realized. In this thesis, we have argued that this is partially due to its restrictive assumptions, such as view insufficiency, and have developed multi-view learning techniques that overcome some of these challenges. The application of multi-view learning methods, including those explored in this thesis, to other multi-sensory problem domains is an interesting avenue for future work that can provide further incites into the multi-view learning problem.

Model adaptation is an interesting application of multi-view semi-supervised learn-

ing. The co-adaptation approach explored in this thesis exploits multiple views to perform model adaptation with co-training. Model adaptation is a well studied problem in the speech recognition literature, where adaptation techniques have been explored for the single-view learning scenario. The combination of co-adaptation with existing single-view adaptation techniques such as maximum likelihood linear regression (MLLR) is a promising direction for future research.

Multi-view learning approaches have difficult dealing with noise. This thesis proposed a heteroscedastic Bayesian co-training algorithm for learning from noisy views corrupted by complex noise processes, such as occlusion, that are common to multi-sensor problems. To make learning and inference tractable in the per-sample noise model, a quantized noise model was assumed, whose noise components matrices were assumed known on the labeled data samples. The investigation of alternative noise models for the representation and learning of complex noise processes is an important avenue of future work. The correlated noise model is of particular interest, since its a general, yet constrained representation in that it can model a variety of different noise processes, and depending on the employed basis has a manageable number of model parameters.

Bayesian co-training bears close similarity to co-regularization based multi-view learning. Co-regularization methods form an interesting class of related multi-view learning techniques that provide alternative optimization and regularization strategies to Bayesian co-training. The further investigation of the connection between these two multi-view learning algorithms, along with the use of sample-dependent noise models within co-regularization based frameworks is an exciting area for future work that can lead to a better understanding of existing multi-view learning methods and result in a new class of multi-view learning techniques.

The multi-view learning algorithms explored in this thesis comprise an alternative set of optimization strategies for learning from partially labeled data compared to manifold-based semi-supervised learning methods. The development of learning algorithms that simultaneously exploit manifold structure when available in addition to agreement-based priors is a promising future direction. This includes an evaluation

of the tradeoff between each of these approaches and the potential benefit of their combination.

Finally, the last portion of this thesis focused on unsupervised feature selection algorithms borrowing ideas from distributed source coding. Our GP-based distributed feature selection algorithm made the simplifying assumption of feature bin independence, however, more compact representations can be achieved if inter-bin dependencies are accounted for. The use of latent variable representations for model feature dependencies within our GP-based distributed coding approach is an compelling area of future research that can lead to more compact feature encodings and increased applicability of our approach. This includes the combination of our approach with compressed sensing strategies as discussed previously.

The experiments in this thesis have been limited to the task of instance-level object recognition. The application of our distributed coding approach for performing category-level recognition is a interesting research endeavor. As the joint feature distribution is arguably more complex for category level problems, a purely unsupervised feature selection approach will likely have limited success in this domain, and supervised approaches are necessary to reduce problem complexity and derive efficient feature encodings. In the context of limited computational power at each sensor node, however, the incorporation of supervision for performing distributed feature selection is a challenging problem and forms an interesting area for future research whose study can further bridge connections between the areas of machine learning and information theory leading to new incites and approaches.

Appendix A

Heteroscedastic Bayesian Co-training Derivations

Heteroscedastic Bayesian co-training proposes a Gaussian Processes graphical model for multi-view learning. This model is depicted in Figure A-1. Here, $\mathbf{x}_j$, $j = 1, ..., m$ are the m views of the learning problem and f_j are the m latent functions defined in each view. f_c is a latent variable referred to as the consensus function that depends on f_j and represents the fused decision of the m functions. y are either labels or target values depending on whether the task is classification or regression.

Under this model we can write the following joint probability,

$$p(\mathbf{y}, \mathbf{f}_c, \mathbf{f}_1, ..., \mathbf{f}_m) = p(\mathbf{y}|\mathbf{f}_c) \prod_j p(\mathbf{f}_c|\mathbf{f}_j) \prod_j p(\mathbf{f}_j), \tag{A.1}$$

where the vectors are defined over examples in the training set (e.g., $\mathbf{f}_j = (f_j^1, ..., f_j^N)$ where N is the number of training examples), and

$$\begin{aligned} p(\mathbf{f}_c|\mathbf{f}_j) &= \mathcal{N}(\mathbf{f}_j, \mathbf{A}_j), \\ p(\mathbf{f}_j) &= \mathcal{N}(0, \mathbf{K_j}). \end{aligned} \tag{A.2}$$

In the derivation that follows, we use the following property involving the product

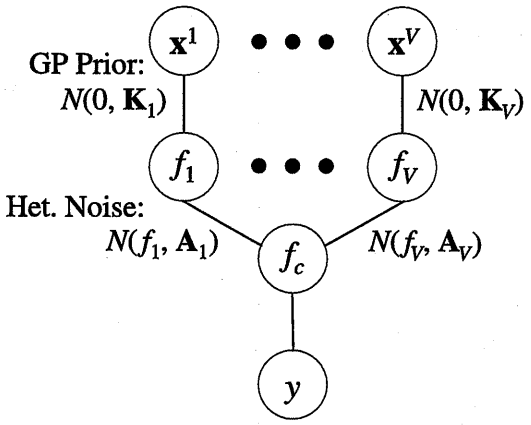


Figure A-1: Graphical model for heteroscedastic Bayesian co-training.

of two Gaussians with means $\mathbf{a}$ and $\mathbf{b}$ and covariances $\mathbf{A}$ and $\mathbf{B}$,

$$\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{B}) = Z\mathcal{N}(\mathbf{x}|\mathbf{c}, \mathbf{C}) \tag{A.3}$$

where

$$\begin{aligned} \mathbf{c} &= \mathbf{C}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}), \\ \mathbf{C} &= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}, \\ Z &= \mathcal{N}(\mathbf{b}|\mathbf{a}, \mathbf{A} + \mathbf{B}). \end{aligned} \tag{A.4}$$

Note that Z could alternatively be expressed as a normal distribution over $\mathbf{a}$ with mean $\mathbf{b}$ depending on which is the variable of interest.

By an iterative application of Eq. A.3 it is straightforward to show that a product of n Gaussians with means $\boldsymbol{\mu}_j$ and covariances $\mathbf{C}_j$, $j = 1, ..., n$ is a Gaussian with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ given by

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{C}\textstyle\sum_{j=1}^{m} \mathbf{C}_j^{-1}\boldsymbol{\mu}_j, \\ \mathbf{C} &= \left(\textstyle\sum_{j=1}^{m} \mathbf{C}_j^{-1}\right)^{-1}. \end{aligned} \tag{A.5}$$

Before deriving the Bayesian co-training kernel, we first derive an expression for $p(\mathbf{f}_c|\mathbf{f}_1, ..., \mathbf{f}_j)$. Note that from Eq. A.1 and observing Figure A-1 we can write,

$$p(\mathbf{f}_c|\mathbf{f}_1, ..., \mathbf{f}_m) = \prod_j p(\mathbf{f}_c|\mathbf{f}_j) = \prod_j \mathcal{N}(\mathbf{f}_j, \mathbf{A}_j). \tag{A.6}$$

The product of m Gaussian distributions over the same input variable is itself a Gaussian. Using Eq. A.5 we can re-express the above Equation as,

$$p(\mathbf{f}_c|\mathbf{f}_1, ..., \mathbf{f}_m) = Z\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{A}_c), \tag{A.7}$$

where,

$$\boldsymbol{\mu}_c = \mathbf{A}_c \left(\sum_j \mathbf{A}_j^{-1}\mathbf{f}_j\right), \quad \mathbf{A}_c = \left(\sum_j \mathbf{A}_j^{-1}\right)^{-1}, \tag{A.8}$$

and Z is a normalization constant.

We now derive the Bayesian co-training kernel. Here we are interested in obtaining an expression for $p(\mathbf{f}_c)$ obtained by marginalizing over $\mathbf{f}_j$ in Eq. A.1. In particular, we are interested in finding,

$$p(\mathbf{f}_c) = \int_{\mathbf{f}_1,...,\mathbf{f}_m} p(\mathbf{f}_c, \mathbf{f}_1, ..., \mathbf{f}_m)\ \mathrm{d}\mathbf{f}_1 ... \mathrm{d}\mathbf{f}_m \tag{A.9}$$

Using Eq. A.1 we can re-express the above Equation as,

$$p(\mathbf{f}_c) = \int_{\mathbf{f}_1,...,\mathbf{f}_m} \prod_j \mathcal{N}(\mathbf{f}_c, \mathbf{A}_j) \prod_j \mathcal{N}(0, \mathbf{K}_j)\ \mathrm{d}\mathbf{f}_1 ... \mathrm{d}\mathbf{f}_m \tag{A.10}$$

where $\mathcal{N}(\mathbf{f}_j, \mathbf{A}_j)$ is re-expressed as $\mathcal{N}(\mathbf{f}_c, \mathbf{A}_j)$ since the integral is over $\mathbf{f}_j$ and $\mathbf{f}_c$ is held constant.

By group terms according by $\mathbf{f}_j$ the above Equation can be simplified to get,

$$p(\mathbf{f}_c) = \prod_j \int_{\mathbf{f}_j} \mathcal{N}(\mathbf{f}_c, \mathbf{A}_j)\mathcal{N}(0, \mathbf{K}_j)\ \mathrm{d}\mathbf{f}_j. \tag{A.11}$$

Using Equation A.3 we have,

$$p(\mathbf{f}_c) = \prod_j \int_{\mathbf{f}_j} Z_j \mathcal{N}(\mathbf{c}, \mathbf{C})\ \mathrm{d}\mathbf{f}_j, \tag{A.12}$$

where,

$$Z_j = \mathcal{N}(0, \mathbf{K}_j + \mathbf{A}_j) \tag{A.13}$$

Evaluating the integral gives,

$$p(\mathbf{f}_c) = \prod_j Z_j = \prod_j \mathcal{N}(0, \mathbf{K}_j + \mathbf{A}_j) \tag{A.14}$$

Using Eq. A.5 the above expression is further simplified to give,

$$p(\mathbf{f}_c) = Z\mathcal{N}(0, \mathbf{K}_c), \tag{A.15}$$

where,

$$\mathbf{K}_c = \left(\sum_j (\mathbf{K}_j + \mathbf{A}_j)^{-1} \right)^{-1} \tag{A.16}$$

Equation A.16 is the heteroscedastic co-training kernel, $\mathbf{K}_c$, of Chapter 5.

Bibliography

[1] A. Aaron and B. Girod. Compression with side information using turbo codes. In *Data Compression Conference*, April 2002.

[2] S. Abney. Understanding the Yarwosky algorithm. *Computational Linguistics*, 30(3):365–395, September 2004.

[3] I.F. Akyildiz, S. Weilian, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications*, 40(8):102–114, August 2002.

[4] M. Alvarez and N. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems (NIPS)*, December 2008.

[5] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research (JMLR)*, 6:1817–1853, November 2005.

[6] R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *International Conference on Machine Learning (ICML)*, June 2007.

[7] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *International Conference on Machine Learning (ICML)*, 2004.

[8] N. Balcan, A. Blum, and K. Yang. Co-training and expansion: towards bridging theory and practice. In *Advances in Neural Information Processing (NIPS)*, December 2004.

[9] S. Bickel and T. Scheffer. Estimation of mixture models using co-em. In *Proceedings of the ICML Workshop on Learning with Multiple Views*, August 2005.

[10] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, August 1999.

[11] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Workshop on Computational Learning Theory (COLT)*, July 1998.

[12] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[13] E.V. Bonilla, K.M.A. Chai, and C.K.I. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems (NIPS)*, December 2008.

[14] C. Buehler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen. Unstructured lumigraph rendering. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, August 2001.

[15] L.-P. Morency C. M. Christoudias and T. Darrell. Light field appearance manifolds. In *European Conference on Computer Vision (ECCV)*, May 2004.

[16] J. Quinonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research (JMLR)*, 6:1939–1959, December 2005.

[17] G. A. Di Caro, F. Ducatelle, and L.M. Gambardella. *Applications of Evolutionary Computing*, chapter Wireless Communications for Distributed Navigation in Robot Swarms, pages 21–30. Springer, 2009.

[18] 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.D. Tygar, and S. S. Sastry. CITRIC: A low-bandwidth wireless camera network platform. In *International Conference on Distributed Smart Cameras (ICDSC)*, December 2008.

[19] J. Chou, S. S. Pradhan, and K. Ramchandran. Turbo and trellis-based constructions for source coding with side information. In *Data Compression Conference*, March 2003.

[20] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *International Conference on Multimodal Interfaces (ICMI)*, November 2006.

[21] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2008.

[22] C. M. Christoudias, R. Urtasun, and T. Darrell. Unsupervised feature selection via distributed coding for multi-view object recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[23] C. M. Christoudias, R. Urtasun, A. Kapoor, and T. Darrell. Co-training with noisy perceptual observations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[24] T. P. Coleman, A. H. Lee, M. Medard, and M. Effros. On some new approaches to practical Slepian-Wolf compression inspired by channel coding. In *Data Compression Conference*, March 2004.

[25] T. P. Coleman, A. H. Lee, M. Medard, and M. Effros. Low-complexity approaches to Slepian-Wolf near-lossless distributed data compression. *IEEE Transactions on Information Theory*, 52(8):3546–3561, August 2006.

[26] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, June 1999.

[27] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley and Sons, 1991.

[28] S.C. Dass, K. Nandakumar, and A.K. Jain. A principled approach to score level fusion in multimodal biometric systems. In *Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, June 2005.

[29] V. R. de Sa. Spectral clustering with two views. In *European Conference on Machine Learning (ECML)*, October 2005.

[30] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, August 2001.

[31] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, August 2003.

[32] C. Ding, X. He, and H. D. Simon. On the equivalence of non-negative matrix factorization and spectral clustering. In *SIAM Data Mining Conference*, 2005.

[33] J. Dy and C. Brodley. Feature subset selection and order identification for unsupervised learning. In *International Conference on Machine Learning (ICML)*, July 2000.

[34] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

[35] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(4):594–611, 2006.

[36] V. Ferrari, T. Tuytelaars, and L. Van Gool. Integrating multiple model views for object recognition. June 2004.

[37] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[38] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *International Conference on Computer Vision (ICCV)*, October 2007.

[39] R. G. Gallager. *Low density parity check codes.* PhD thesis, MIT, 1963.

[40] A. Geiger, R. Urtasun, and T. Darrell. Rank priors for continuous non-linear dimensionality reduction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[41] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems (NIPS)*, November 1994.

[42] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero. Distributed video coding. *Proceedings of the IEEE*, 93(1):71–83, January 2005.

[43] M. Gonen and E. Alpaydin. Localized multiple kernel learning. In *International Conference on Machine Learning (ICML)*, July 2008.

[44] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *International Conference on Computer Vision (ICCV)*, October 2005.

[45] K. Grauman and T. Darrell. The pyramid match kernel: efficient learning with sets of features. *Journal of Machine Learning Research (JMLR)*, 8:725–760, May 2007.

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

[47] T. J. Hazen, K. Saenko, C. H. La, and J. Glass. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. In *International Conference on Multimodal Interfaces (ICMI)*, October 2005.

[48] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Conference on Human Language Technology (HLT)*, May 2003.

[49] J. Huang, E. Marcheret, and K. Visweswariah. Rapid feature space speaker adaptation for multi-stream HMM-based audio-visual speech recognition. In *International Conference on Multimedia and Expo (ICME)*, July 2005.

[50] O. Javed, Z. Rasheed, K Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *International Conference on Computer Vision (ICCV)*, October 2003.

[51] S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *Conference on Learning Theory (COLT)*, June 2007.

[52] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with Gaussian processes for object categorization. In *International Conference in Computer Vision (ICCV)*, October 2007.

[53] A. Kapoor, A. Qi, H. Ahn, and R. W. Picard. Hyperparameter and kernel learning fro graph based semi-supervised classification. In *Advances in Neural Information Processing Systems (NIPS)*, December 2005.

[54] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2004.

[55] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 1999.

[56] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(3):226–239, 1998.

[57] N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research (JMLR)*, 6:1783–1816, December 2005.

[58] N. D. Lawrence and M. I. Jordan. Semi-supervised learning via Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, December 2004.

[59] N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with Gaussian processes. In *International Conference in Machine Learning (ICML)*, June 2009.

[60] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[61] J. J. Lee. LIBPMK: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-017, MIT CSAIL, 2008.

[62] B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *DAGM Pattern Recognition Symposium*, August 2004.

[63] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *International Conference on Computer Vision (ICCV)*, October 2003.

[64] M. Levoy and P. Hanrahan. Light field rendering. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, October 1996.

[65] T. Li and M. Ogihara. Semi-supervised learning from different information sources. *Knowledge Information Systems Journal*, 7(3):289–309, March 2005.

[66] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[67] H. Liu and L. Yu. Feature selection for data mining. Technical report, Arizona State University, 2002.

[68] A. D. Liveris, C. Lan, K. Narayanan, Z. Xiong, and C. N. Georghiades. Slepian-Wolf coding of three binary sources using ldpc codes. In *International Symposium on Turbo Codes and Related Topics*, March 2003.

[69] A. D. Liveris, Z. Xiong, and C. N. Georghiades. A distributed source coding technique for highly correlated images using turbo-codes. In *International Conference on Acoustic Speech and Signal Processing (ICASSP)*, May 2002.

[70] A. D. Liverisa, Z. Xiong, and C. N. Georghiades. Compression of binary sources with side information at the decoder using ldpc codes. In *IEEE Global Communications Symposium*, October 2002.

[71] D. J. C. MacKay and R. M. Neal. Near shannon limit performance of low density parity check codes. *Electronics Letters*, 33(6):457–458, March 1997.

[72] B. Maeireizo, D. Litman, and R. Hwa. Co-training for predicting emotions with spoken dialogue data. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, July 2004.

[73] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, 65(1-2):43–72, November 2005.

[74] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[75] I. Muslea, S. Minton, and C. A. Knoblock. Adaptive view validation: a first step towards automatic view detection. In *International Conference on Machine Learning (ICML)*, July 2002.

[76] K. Nandakumar. *Multibiometric systems: fusion strategies and template security.* PhD thesis, Michigan State University, 2008.

[77] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (COIL-100). Technical report, Colmubia University, February 1996.

[78] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Technical Report WS00AVSR, Johns Hopkins University, CLSP, 2000.

[79] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Workshop on Information and Knowledge Management*, 2000.

[80] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[81] S. Pan, S. Shen, M. X. Zhou, and K. Houck. Two-way adaptation for robust input interpretation in practical multimodal conversation systems. In *Conference on Intelligent User Interfaces (IUI)*, January 2005.

[82] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(8):1226–1238, August 2005.

[83] G. Potamianos, C. Neti, G. Gravier, and A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. In *Proceedings of the IEEE*, volume 91, pages 1306–1326, September 2003.

[84] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): design and construction. *IEEE Transactions on Information Theory*, 49(3):626–643, March 2003.

[85] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

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

[87] B. Rimoldi and R. Urbanke. Asynchronous Slepian-Wolf coding via source-splitting. In *International Symposium on Information Theory*, June 1997.

[88] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision (IJCV)*, pages 231–259.

[89] K. Saenko and T. Darrell. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing (NIPS)*, December 2008.

[90] M. Salzmann, R. Urtasun, and P. Fua. Local deformation models for monocular 3D shape recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[91] M. Sartipi and F. Fekri. Distributed source coding in wireless sensor networks using ldpc codes: A non-uniform framework. In *Data Compression Conference*, March 2005.

[92] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. In *International Conference on Computer Vision (ICCV)*, October 2007.

[93] D. Schonberg. *Practical distributed source coding and its application to the compression of encrypted data.* PhD thesis, University of California, Berkeley, 2007.

[94] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.

[95] V. Sindhwani, W. Chu, and S. S. Keerthi. Semi-supervised Gaussian processes. In *International Joint Conference on Artificial Intelligence (IJCAI)*, January 2007.

[96] V. Sindhwani, J. Hu, and A. Mojsilovic. Regularized co-clustering with dual supervision. In *Advances in Neural Information Processing Systems (NIPS)*, December 2008.

[97] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*, August 2005.

[98] V. Sindhwani and D. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *International Conference on Machine Learning (ICML)*, July 2008.

[99] J. Sinkkonen, J. Nikkila, L. Lahti, and S. Kaski. Associative clustering. In *European Conference on Machine Learning (ECML)*, September 2004.

[100] M. R. Siracusa and J. W. Fisher III. Dynamic dependency tests: analysis and applications to multi-modal data association. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[101] M. R. Siracusa, K. Tieu, A. Ihler, J. Fisher III, and A. S. Willsky. Estimating dependency and significance for high-dimensional data. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2005.

[102] J. Sivic and A. Zisserman. *Toward category-level object recognition*, chapter Video Google: efficient visual search of videos, pages 127–144.

[103] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, July 1973.

[104] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, January 2005.

[105] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research (JMLR)*, 7:1531–1565, July 2006.

[106] V. Stankovic, A. D. Liveris, Z. Xiong, and C. N. Georghiades. Design of Slepian-Wolf codes by channel code partitioning. In *Data Compression Conference*, March 2004.

[107] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 22(8):747–757, August 2000.

[108] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiel, and L. Van Gool. Towards multi-view object class detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[109] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(5):854–869, May 2007.

[110] R. Urtasun and T. Darrell. Discriminative Gaussian process latent variable models for classification. In *International Conference in Machine Learning (ICML)*, June 2007.

[111] R. Urtasun and T. Darrell. Local probabilistic regression for activity-independent human pose inference. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[112] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. Darrell, and N. D. Lawrence. Topologically-constrained latent variable models. In *International Conference in Machine Learning (ICML)*, July 2008.

[113] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *International Conference in Computer Vision (ICCV)*, October 2005.

[114] N. Vandapel, R. R. Donamukkala, and M. Hebert. Unmanned ground vehicle navigation using aerial ladar data. *International Journal of Robotics Research*, 25(1):31–51, January 2006.

[115] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *International Conference on Computer Vision (ICCV)*, October 2007.

[116] J. Wang, D.J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):283–298, February 2008.

[117] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Advances in Neural Information Processing (NIPS)*, December 2000.

[118] F. M. J. Willems. Totally asynchronous Slepian-Wolf data compression. *IEEE Transactions on Information Theory*, 34(1):35–44, January 1988.

[119] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, January 1976.

[120] A. D. Wyner. Recent results in the shannon theory. *IEEE Transactions on Information Theory*, 20(1):2–10, January 1974.

[121] B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. L. Oviatt. Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In *International Conference on Multimodal Interfaces (ICMI)*, November 2003.

[122] Z. Xiong, A. D. Liveris, and S. Cheng. Distributed source coding for sensor networks. *IEEE Signal Processing Magazine*, 21(5):80–94, September 2004.

[123] R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005.

[124] A. Y. Yang, S. Maji, C. M. Christoudias, T. Darrell, J. Malik, and S. S. Sastry. Multiple-view object recognition in band-limited distributed camera networks. In *International Conference on Distributed Smart Cameras (ICDSC)*, June 2009.

[125] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1995.

[126] C. Yeo, P. Ahammad, and K. Ramchandran. A rate-efficient approach for establishing visual correspondences via distributed source coding. In *Conference on Visual Communications and Image Processing (VCIP)*, January 2008.

[127] C. Yeo, P. Ahammad, and K. Ramchandran. Rate-efficient visual correspondences using random projections. In *International Conference on Image Processing (ICIP)*, October 2008.

[128] S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and R. B. Rao. Bayesian co-training. In *Advances in Neural Information Processing (NIPS)*, December 2007.

[129] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Transactions on Information Theory*, 48(6):1250–1276, June 2002.

[130] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, December 2004.

[131] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: discriminative nearest-neighbor classification for visual category recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006.

[132] Y. Zhao and J. Garcia-Frias. Data compression of correlated non-binary sources using punctured turbo codes. In *Data Compression Conference*, April 2002.

[133] A. Zhen and C. S. Ong. Multiclass multiple kernel learning. In *International Conference on Machine Learning (ICML)*, June 2007.

[134] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, October 2003.

[135] X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning from Gaussian fields to Gaussian processes. Technical report, Carnegie Mellon University, 2003.