SEMI-SUPERVISED MULTI-DOMAIN REGRESSION WITH DISTINCT TRAINING SETS

Tomer Michaeli, Yonina C. Eldar

Technion–Israel Institute of Technology

Guillermo Sapiro

University of Minnesota

ABSTRACT

We address the problems of multi-domain and single-domain regression based on distinct labeled training sets for each of the domains and a large unlabeled training set from all domains. We formulate these problems as ones of Bayesian estimation with partial knowledge of statistical relations. We propose a worst-case design strategy and study the resulting estimators. Our analysis explicitly accounts for the cardinality of the labeled sets and includes the special cases in which one of the labeled sets is very large or, in the other extreme, completely missing. We demonstrate our estimators in the context of audio-visual word recognition and provide comparisons to several recently proposed multi-modal learning algorithms.

Index Terms— Bayesian estimation, multi-modal learning.

1. INTRODUCTION

In many application areas one can access data from multiple domains to perform a task. For example, word recognition can greatly benefit from the availability of joint audio-visual measurements [1]. Identity recognition and verification can be performed much more accurately by fusing information from several modalities such as facial images, iris scans, voice recordings, handwritings, and more.

A difficulty in fusing multiple sources, though, is that one can often access only distinct labeled training sets for the different domains and does not have paired labeled examples from all domains. Consider, for instance, audio-visual gender recognition. There are numerous existing data-sets of labeled voice recordings as well as labeled data-sets of facial images. However, there are only a few audio-visual data-sets (where the audio and video are paired and labeled), with limited number of subjects each. Thus, it is easy to train a classifier based only on audio or only on image data, but it is not clear how the two modalities should be best fused.

While paired multi-domain labeled examples are typically scarce, paired unlabeled examples are often abundant. For instance, enormous amounts of speaker video sequences (together with audio) can be easily collected. These videos, though, often do not come with labels. However, they can be used to unveil the statistical relations between audio and video. An important question is how to best fuse audio- and image-based predictors, given these relations.

An even more interesting and practical question is whether the availability of multiple data sources can aid a machine learning algorithm during training, when not all are available during testing. For example, suppose we want to predict the age of a speaker. Assume we have a labeled audio training set, a labeled image training set, and a large amount of unlabeled audio-visual examples. Can the visual examples help construct a predictor, which is solely based on audio?

In this paper we address the problems of multi-domain and single-domain regression based on distinct labeled single-domain training sets and unlabeled multi-domain data. Specifically, focusing on two domains for simplicity, we consider the situation in which available to us is a very large unlabeled training set $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i\}$ and two (mutually unpaired) labeled sets $\{\boldsymbol{x}_1^i, \boldsymbol{y}^i\}$ and $\{\boldsymbol{x}_2^i, \boldsymbol{y}^i\}$. Using this training data, we treat the problems of designing a predictor of $\boldsymbol{y}$ based on $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (multi-domain regression) and a predictor of $\boldsymbol{y}$ based on $\boldsymbol{x}_1$ alone (single-domain regression). Our analysis explicitly accounts for the cardinality of the labeled sets. In particular, it includes the case in which one or both labeled sets are very large as well as the case in which one labeled set is completely missing.

A problem related to ours is *multi-view learning* [2] in general and multi-view regression [3] in particular. These techniques make use of a large training set of data from multiple domains (views), containing only a few labeled examples. If the views tend to agree, then the unlabeled examples are useful [2, 3]. In our setting, however, we do not observe even a single multi-domain labeled example $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \boldsymbol{y}^i\}$ and also make no assumptions on the underlying distribution. Situations in which labeled samples from a source domain are used to construct a predictor for a target domain fall under the category of *transfer learning* [4]. Nevertheless, in these settings, paired unlabeled examples from the two domains are not accessible.

More related to our setting are the *cross-modality* and *shared-representation* learning scenarios recently studied in [1] in the context of multi-modal learning. In both settings, unlabeled training data $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i\}$ from multiple modalities, such as audio and video, are used to perform a *feature learning* stage. In cross-modality learning, then, one constructs a predictor based on $\boldsymbol{x}_1$ using a labeled training set $\{\boldsymbol{x}_1^i, \boldsymbol{y}\}$. For example, we may want to build a classifier operating on audio features by observing labeled audio examples in addition to unlabeled audio-visual instances. In shared-representation learning, one constructs a predictor based on $\boldsymbol{x}_1$ using a labeled training set $\{\boldsymbol{x}_2^i, \boldsymbol{y}\}$. For instance, we may want to train an audio classifier by observing labeled visual examples in addition to unlabeled audio-visual instances. Shared-representation regression was studied from a Bayesian estimation perspective in [5], in which a link to instrumental variable regression was also discussed. Both cross-modality and shared-representation regression are special cases of the general setting we address here, corresponding to the situation in which there are zero examples in one of the labeled sets.

Finally, in statistics, regression involving two types of covariates is often performed via partially linear models. These methods can be applied in multi-domain settings in which both the unlabeled set and one of the labeled sets are very large whereas the other labeled set is small. Such situations fall within our problem formulation as well.

Due to space limitations, we state here the main results without proofs, which will appear in [6].

2. PROBLEM FORMULATION

We assume we are given access to three data-sets as follows:

1. labeled examples $\{(\boldsymbol{x}_1^\ell, \boldsymbol{y}^\ell)\}_{\ell=1}^{L_1}$ from domain 1;

This work was supported in part by a Google Research Award, NSF, ONR, NGA, DARPA, ARO, and NSSEFF.

2. labeled examples $\{(\boldsymbol{x}_2^\ell, \boldsymbol{y}^\ell)\}_{\ell=L_1+1}^{L_1+L_2}$ from domain 2;
3. paired unlabeled examples $\{(\boldsymbol{x}_1^u, \boldsymbol{x}_2^u)\}_{u=L_1+L_2+1}^{L_1+L_2+U}$.

These training sets correspond to independent draws from the distributions F_{X_1Y}, F_{X_2Y}, and $F_{X_1X_2}$, respectively. Our focus is on situations in which the number U of unlabeled examples is very large, so that the joint distribution $F_{X_1X_2}$ can be assumed known (or very well approximated). The cardinalities L_1 and L_2 of the labeled sets are arbitrary. In particular, one of them can be zero. In this case no knowledge whatsoever is available regarding the statistical relation between Y and the associated domain. On the other extreme, one (or both) of the labeled sets may be very large, in which case the associated single-domain MMSE estimator, say $\mathbb{E}[Y|X_1]$, can be assumed known.

In terms of testing, we address two tasks. The first is *multi-domain regression*, in which the algorithm is asked to predict the value $\boldsymbol{y}$ based on an observation of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The second is *single-domain regression*, where prediction should be based solely on $\boldsymbol{x}_1$.

We adopt and generalize the frameworks of [5, 7] by posing our problem as one of estimation with partial knowledge of statistical relations. Before formalizing our problem in estimation theoretic terms, we first recall the common practice for regression from one domain with a limited number of examples.

2.1. Single-Domain Regression with Single-Domain Training

Suppose we are given a sample $\{\boldsymbol{x}^\ell, \boldsymbol{y}^\ell\}_{\ell=1}^L$, independently drawn from the joint distribution of the random variables (RVs) X and Y, which take values in $\mathbb{R}^M$ and $\mathbb{R}^N$, respectively. If L is very large, then nonparametric methods can be used to approximate the conditional expectation curve $\varphi(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$ with great accuracy at any $\boldsymbol{x}$. Such estimates, however, are often far from accurate when L is small. Common practice in such situations is to use parametric or semi-parametric methods that impose some structure on the sought predictor. In other words, rather than trying to approximate the regression function $\varphi(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$, which minimizes the MSE among all functions of X, we settle for approximating the optimal predictor among some family $\mathcal{A}$ of functions:

$$\varphi^* = \arg\min_{\varphi\in\mathcal{A}} \mathbb{E}\left[\|Y - \varphi(X)\|^2\right]. \tag{1}$$

The less rich the class $\mathcal{A}$ is, the more accurate we can typically approximate $\varphi^*(X)$ from the training data. This comes at the cost that the (theoretical) MSE that $\varphi^*(X)$ achieves is higher. In the sequel, we term φ^* of (1) the $\mathcal{A}$-optimal estimator of Y from X.

One of the simplest structural restrictions corresponds to linear estimation, so that $\mathcal{A}$ is the set of all linear functions from $\mathbb{R}^M$ to $\mathbb{R}^N$. In this case,

$$\varphi^*(X) = \boldsymbol{\Gamma}_{YX}\boldsymbol{\Gamma}_{XX}^\dagger X. \tag{2}$$

The second-order moments $\boldsymbol{\Gamma}_{YX} = \mathbb{E}[YX^T]$ and $\boldsymbol{\Gamma}_{XX} = \mathbb{E}[XX^T]$ can be estimated from the training set, for example, by using sample moments. A more general model corresponds to the collection $\mathcal{A}$ of all functions of the form $\varphi(X) = \sum_{k=1}^K a_k\varphi_k(X)$, where $\{\varphi_k\}_{k=1}^K$ is a predefined set of functions from $\mathbb{R}^M$ to $\mathbb{R}^N$. The optimal coefficients $\boldsymbol{a} = (a_1, \cdots, a_K)^T$ can be obtained similar to the linear setting.

In both examples, $\mathcal{A}$ forms a linear subspace of functions, as for every $\varphi^1, \varphi^2 \in \mathcal{A}$ and $\alpha, \beta \in \mathbb{R}$, the function $\alpha\varphi^1 + \beta\varphi^2$ also belongs to $\mathcal{A}$. We note that this claim is also trivially true when $\mathcal{A}$ is taken to be the set of all (Borel-measurable) functions, in which case $\varphi^*(X) = \mathbb{E}[Y|X]$, and when $\mathcal{A}$ contains only the zero function, in which case $\varphi^*(X) = 0$.

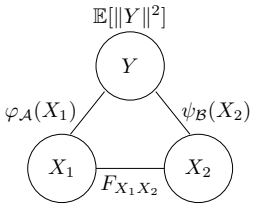

Fig. 1: Known statistical relationships.

2.2. Statistical Relations Deduced from Separate Training Sets

In our setting we have access to two sperate sets of labeled examples, one for each domain. We can therefore determine the $\mathcal{A}$-optimal predictor of Y given X_1 as well as the $\mathcal{B}$-optimal predictor of Y from X_2, were $\mathcal{A}$ and $\mathcal{B}$ are classes of functions chosen in accordance with the cardinality of the the two sets. We model the existence of numerous unlabeled examples (X_1, X_2) by the assumption that the joint distribution of X_1 and X_2 is known. We also assume that the second-order moment of Y is known (or accurately estimated from the labeled sets). The statistical relationships assumed known are depicted in Fig. 1.

In a more mathematical language, assume we are given two functions $\varphi^* : \mathbb{R}^{M_1} \to \mathbb{R}^N$ and $\psi^* : \mathbb{R}^{M_2} \to \mathbb{R}^N$, a cumulative probability function $F_{X_1X_2} = \mathbb{P}(X_1 \le \boldsymbol{x}_1, X_2 \le \boldsymbol{x}_2)$ over $\mathbb{R}^{M_1\times M_2}$, and a scalar $c > 0$. Then what we know regarding the RVs X_1, X_2 and Y is that their distribution $F_{X_1X_2Y}$ belongs to the set $\mathcal{F}$ of distributions satisfying

$$\varphi^* = \arg\min_{\varphi\in\mathcal{A}} \mathbb{E}[\|Y - \varphi(X_1)\|^2],\ \psi^* = \arg\min_{\psi\in\mathcal{B}} \mathbb{E}[\|Y - \psi(X_2)\|^2],$$
$$F_{X_1X_2Y}(\boldsymbol{x}_1, \boldsymbol{x}_2, \infty) = F_{X_1X_2}(\boldsymbol{x}_1, \boldsymbol{x}_2), \quad \mathbb{E}[\|Y\|^2] = c, \tag{3}$$

where $\mathcal{A}$ and $\mathcal{B}$ are linear subspaces of functions.

Any predictor of Y, whether a function of X_1 and X_2 or of X_1 alone, may perform well under certain distributions $F_{X_1X_2Y} \in \mathcal{F}$ and worse under others. Our goal is therefore to uniformly optimize the performance over $\mathcal{F}$.

3. MULTI-DOMAIN REGRESSION

For any distribution $F_{X_1X_2Y}$, the MSE attained by an estimator $\hat{Y} = \rho(X_1, X_2)$ is defined as

$$\mathrm{MSE}(F_{X_1X_2Y}, \rho) = \mathbb{E}\left[\|Y - \rho(X_1, X_2)\|^2\right], \tag{4}$$

where the expectation is with respect to $F_{X_1X_2Y}$. Since the MSE depends on $F_{X_1X_2Y}$, which is unknown, our approach is to seek the estimator whose worst-case MSE over $F_{X_1X_2Y} \in \mathcal{F}$ is minimal. Namely, we are interested in

$$\rho^* = \arg\min_{\rho} \sup_{F_{X_1X_2Y}\in\mathcal{F}} \mathrm{MSE}(F_{X_1X_2Y}, \rho). \tag{5}$$

The next theorem provides a means for solving this problem.

Theorem 1 *Choose any $F_{X_1X_2Y} \in \mathcal{F}$ and consider the estimator*

$$\rho_{\mathcal{C}}^* = \arg\min_{\rho\in\mathcal{C}} \mathrm{MSE}(F_{X_1X_2Y}, \rho), \tag{6}$$

where $\mathcal{C} = \{\rho : \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \phi(\boldsymbol{x}_1) + \psi(\boldsymbol{x}_2),\ \phi \in \mathcal{A},\ \psi \in \mathcal{B}\}$. Then

1. *the function $\rho_{\mathcal{C}}^*$ does not depend on $F_{X_1X_2Y} \in \mathcal{F}$;*

2. $\mathrm{MSE}(F_{X_1X_2Y}, \rho_{\mathcal{C}}^*)$ *does not depend on* $F_{X_1X_2Y} \in \mathcal{F}$*;*

3. *the estimator* $\rho_{\mathcal{C}}^*$ *of* (6) *is also the solution* ρ^* *to* (5)*.*

Theorem 1 shows that instead of tempting to solve the minimax problem (6), we can equivalently solve the minimization problem (5). Namely, all we need to do is determine the MMSE estimator of Y among all functions of the form $\phi(X_1)+\psi(X_2)$ with $\phi \in \mathcal{A}$ and $\psi \in \mathcal{B}$. As we now show, in many practical cases the latter possesses a closed form solution.

3.1. Single-Domain Training

Suppose that we have at our disposal only labeled examples from one domain, say X_1. In this case $\mathcal{B} = \{0\}$ so that $\mathcal{C} = \mathcal{A}$. Consequently, the solution to (6) is simply

$$\rho^*(X_1, X_2) = \varphi^*(X_1). \tag{7}$$

This shows that, at least from a worst-case perspective, there is no gain in basing the prediction on the domain X_2 for which we have no labeled training examples. Namely, for any estimator that differs from $\varphi^*(X_1)$, and, in particular, a function of X_2, there exist distributions $F_{X_1X_2Y} \in \mathcal{F}$ (one maybe being the true underlying distribution) under which $\varphi^*(X_1)$ performs better.

This result does not stand in contrast to the basic observation in multi-view learning that unlabeled data helps [2]. This is because in our setting, we do not assume that the two views are "coherent" or tend to agree in any sense, as done *e.g.,* in [3].

3.2. Multi-Domain Linear Regression

Suppose that the labeled training sets we have suffice to identify (with very high precision) the optimal linear estimator from each view. In this case $\mathcal{A}$ and $\mathcal{B}$ correspond to the collection of all linear functions from $\mathbb{R}^{M_1}$ to $\mathbb{R}^N$ and from $\mathbb{R}^{M_2}$ to $\mathbb{R}^N$, respectively. Consequently, $\mathcal{C}$ is the set of all linear functions from $\mathbb{R}^{M_1} \times \mathbb{R}^{M_2}$ to $\mathbb{R}^N$. This implies that the solution to (6) is simply the best linear predictor of Y based on X_1 and X_2, namely

$$\rho^*(X_1, X_2) = \begin{pmatrix} \mathbf{\Gamma}_{YX_1} & \mathbf{\Gamma}_{YX_2} \end{pmatrix} \begin{pmatrix} \mathbf{\Gamma}_{X_1X_1} & \mathbf{\Gamma}_{X_1X_2} \\ \mathbf{\Gamma}_{X_2X_1} & \mathbf{\Gamma}_{X_2X_2} \end{pmatrix}^{\dagger} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \tag{8}$$

The second-order moments $\mathbf{\Gamma}_{X_iX_j}$, $i,j \in \{1,2\}$, can be accurately estimated from the unlabeled training set. Similarly, the matrices $\mathbf{\Gamma}_{YX_j}$, $i,j \in \{1,2\}$, can be determined from the labeled sets.

This analysis trivially extends to the case in which the training sets suffice to identify the optimal parametric estimators of the forms $\varphi(X_1) = \sum_{k=1}^{K_1} a_k^1 \varphi_k(Y)$ and $\psi(X_2) = \sum_{k=1}^{K_2} a_k^2 \psi_k(X_2)$, where $\{\varphi_k\}_{k=1}^{K_1}$ and $\{\psi_k\}_{k=1}^{K_2}$ are given functions.

3.3. Multi-Domain Partially Linear Regression

Suppose that we have numerous labeled examples from the first domain, allowing us to determine $\mathbb{E}[Y|X_1]$, and only a limited amount of labeled examples from the second domain, so that we can only determine the best linear predictor of Y from X_2. In this setting, Theorem 1 implies that the minimax-optimal predictor based on X_1 and X_2 is the estimator minimizing the MSE among all functions of the form

$$\rho(X_1, X_2) = \boldsymbol{a}(X_1) + \boldsymbol{B}X_2, \tag{9}$$

where $\boldsymbol{a} : \mathbb{R}^{M_1} \to \mathbb{R}^N$ and $\boldsymbol{B} \in \mathbb{R}^{N \times M_2}$. In [7] it was shown that the solution to this problem is given by

$$\rho(X_1, X_2) = \mathbb{E}[Y|X_1] + \mathbf{\Gamma}_{YW}\mathbf{\Gamma}_{WW}^{\dagger} W, \tag{10}$$

where $W = X_2 - \mathbb{E}[X_2|X_1]$.

The intuition here is that we need to make sure we do not account for variations in Y twice when fusing information from X_1 and X_2. Thus, we start with the estimate $\varphi^*(X_1) = \mathbb{E}[Y|X_1]$, and then update it with the optimal linear estimate of Y based on the *innovation* $X_2 - \varphi^*(X_1)$ of X_2 with respect to $\varphi^*(X_1)$.

In practice, $\mathbb{E}[Y|X_1]$ and $\mathbb{E}[X_2|X_1]$ can be approximated from the labeled and unlabeled training sets, respectively, using nonparametric methods. The second term in (10) can then be obtained by linearly regressing Y against $X_2 - \mathbb{E}[X_2|X_1]$.

4. SINGLE-DOMAIN REGRESSION

Next, we address the setting in which at the testing stage our predictor is only supplied with one type of features, say X_1. The interesting question in this context is how to take into account the training sets of both domains in order to design an improved estimator of Y based on X_1 alone.

Since our estimator operates on X_1 and is judged by the proximity of its output to Y, its performance is only affected by the joint distribution of Y and X_1. It may thus seem at first that the second set of features X_2 cannot be of help in improving estimation accuracy. However, note that F_{X_1Y} is not fully known in our setting. Thus, being told the statistical relations between Y and X_2 and between X_1 and X_2, helps us narrow down the set of candidate distributions F_{X_1Y} for which we need to design an estimator.

In the single-domain setting we know that, whatever we do, our estimator will not achieve lower MSE than the conditional expectation $\mathbb{E}[Y|X_1]$. We are thus interested in minimizing the *regret* of our estimator $\xi(X_1)$, which is defined as the difference between the MSE it achieves and the MSE of the MMSE solution:

$$R(F_{X_1X_2Y}, \xi) = \mathbb{E}\left[\|Y - \xi(X_1)\|^2\right] - \mathbb{E}\left[\|Y - \mathbb{E}[Y|X_1]\|^2\right]. \tag{11}$$

Similar to the MSE, the regret depends on $F_{X_1X_2Y}$, which is unknown. We would therefore like to design an estimator whose worst-case regret over $F_{X_1X_2Y} \in \mathcal{F}$ is minimal, namely

$$\xi^* = \arg\min_{\xi} \sup_{F_{X_1X_2Y} \in \mathcal{F}} R(F_{X_1X_2Y}, \xi), \tag{12}$$

where now $\xi(X_1)$ is only a function of X_1.

The next theorem describes the single-domain minimax regret estimator in terms of the multi-domain minimax MSE solution.

Theorem 2 *The solution to problem* (12) *is given by*

$$\xi^*(X_1) = \mathbb{E}[\rho^*(X_1, X_2)|X_1] \tag{13}$$

where $\rho^*(X_1, X_2)$ *is the multi-domain minimax estimator* (5)*.*

This result has a very simple and intuitive explanation. We know that $F_{X_1X_2Y}$ belongs to the set $\mathcal{F}$, and therefore $\rho^*(X_1, X_2)$ is the optimal estimate of Y in a minimax-MSE sense. However, we cannot use this estimate as it is a function of X_2, which is not measured in our setting. What Theorem 2 shows is that the optimal strategy is to estimate $\rho^*(X_1, X_2)$ based on the available measurements, which are X_1 alone. Computation of the conditional expectation $\mathbb{E}[\rho^*(X_1, X_2)|X_1]$ only requires knowledge of the marginal distribution $F_{X_1X_2}$, which is available in our setting from the unlabeled data.

We now apply this result to several interesting scenarios.

4.1. Cross-Modality Regression

In the cross-modality learning setting, introduced in [1], we only have labeled examples of the domain X_1 and not of X_2. The intuition, as presented in [1], is that the unlabeled data should somehow help boost the performance of the best single-domain estimator $\varphi^*(X_1)$ that can be designed based on the available labeled set.

This scenario can be treated within our framework by setting $\psi^*(X_2) = 0$. As we have seen in Section 3.1, in this situation $\rho^*(X_1, X_2) = \varphi^*(X_1)$. Therefore, the single-domain minimax-regret predictor of Y from X_1 is given by

$$\xi(X_1) = \mathbb{E}[\varphi^*(X_1)|X_1] = \varphi^*(X_1). \tag{14}$$

We see that despite the fact that we know $F_{X_1X_2}$, there is no better strategy than using the estimator $\varphi^*(X_1)$ here. This implies that cross-modality learning is not useful unless additional knowledge on the underlying distributions is available.

The application of cross-modality learning to classifying isolated words from either audio or video (lipreading) was studied in [1]. It was reported that unlabeled audio-visual examples helped improve visual recognition but failed to boost the performance of an audio classifier. This aligns with our analysis, which states that, in the worst-case scenario, there is nothing better to do than disregarding the modality for which no labeled examples are available.

4.2. Shared-Representation Regression

Consider next the shared-representation learning setting [1] in which we have no labeled examples from the domain X_1 but rather only from X_2. As we have seen in Section 3.1, in this setting $\rho^*(X_1, X_2) = \psi^*(X_2)$. Therefore, the single-domain minimax-regret predictor of Y from X_1 is given in this case by

$$\xi(X_1) = \mathbb{E}[\psi^*(X_2)|X_1]. \tag{15}$$

This expression can be approximated using nonparametric methods by using the unlabeled training examples. This result indicates that when performing prediction based on a modality for which we do not have labels, having such labels for the other modality can help.

5. AUDIO-VISUAL WORD RECOGNITION

We now illustrate the approach derived from our theoretical study in the tasks of spoken digit classification from audio-only and video-only measurements. We used the Grid Corpus [8], which consists of speakers saying simple-structured sentences. Every sentence contains one digit, which we isolated using the supplied transcriptions. We constructed three distinct training sets, one of labeled audio examples (4 males, 4 females), one of visual examples (4 males, 4 females) and one of unlabeled audio-visual examples (6 males, 4 females). Six speakers were used for testing (3 males, 3 females).

We used face detection followed by mean-shift on the gradient image map to extract the lip region. Segments of duration 320msec were used for recognition. This corresponded to 8 consecutive video frames and 1600 audio samples. The image frames were reduced to 10 dimensions using PCA, resulting in an 80-dimensional video feature-vector. The dimension of the spectogram of the audio was reduced to 180, to constitute the audio features. In all experiments Y was a 10-dimensional vector with 1 at the location corresponding to the spoken digit and 0 elsewhere.

Our approach is designed for regression, so that the predicted $\hat{Y}$ is a continuous variable. To perform classification, we chose the maximal element in $\hat{Y}$. For simplicity, $\mathcal{A}$ and $\mathcal{B}$ were taken as the sets of all linear functions (linear regression). This choice yields rather poor classification results based solely on audio or solely on video. Our goal, though, is to demonstrate that, even with such naive single-domain predictors, we can attain good recognition accuracy by using our approach, which cleverly fuses the two domains.

Features		Accuracy	
Training	Testing	Minimax (Grid corpus)	Deep RBM (CUAVE)
Audio	Audio	69.3%	95.8%
Video	Video	52.0%	69.7%
Video	Audio	50.1%	27.5%
Audio	Video	44.6%	29.4%

Table 1: Digit classification performance.

Table 1 compares the accuracy of the proposed approach with that attained by the deep restricted Boltzmann machine (RBM) [1] on the CUAVE dataset [9]. The Grid corpus, used here, is more challenging in that the digits appear within sentences, rather than individually. As can be seen, the single-domain predictors we start with perform relatively poorly (rows 1 and 2). Nevertheless, in the shared-representation settings (rows 3 and 4), our predictors perform much better than the RBM method. Their accuracy is only between 7% and 20% worse than the corresponding single domain estimators (rows 1 and 2, respectively). By contrast, the difference in success rates for the RBM predictor is between 30% and 70%.

6. REFERENCES

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on computational learning theory*, 1998, pp. 92–100.

[3] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Proceedings of the 20th annual conference on learning theory*, 2007, pp. 82–96.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] T. Michaeli and Y. C. Eldar, "Hidden relationships: Bayesian estimation with partial knowledge," *IEEE Trans. Signal Process.*, vol. 59, no. 5, 2011.

[6] T. Michaeli, Y. C. Eldar, and G. Sapiro, "Semi-supervised single- and multi-domain regression with multi-domain training," in preparation.

[7] T. Michaeli, D. Sigalov, and Y. C. Eldar, "Partially linear estimation with application to sparse signal recovery from measurement pairs," *IEEE Trans. Signal Process.*, 2012, submitted. [Online]. Available: http://arxiv.org/abs/1103.5639

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[9] E. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 2017–2020.