Bridging Heterogeneous Domains With Parallel Transport For Vision and Multimedia Applications

Raghuraman Gopalan
Dept. of Video and Multimedia Technologies Research
AT&T Labs-Research
San Francisco, CA 94108

Abstract

Accounting for different feature types across datasets is a relatively under-studied problem in domain adaptation. We address this heterogeneous adaptation setting using principles from parallel transport and hierarchical sparse coding. By learning generative subspaces from each domain, we first perform label-independent cross-domain feature mapping using parallel transport, and obtain a collection of paths (bridges) that could compensate domain shifts. We encode the information contained in these bridges into an expanded prior, and then integrate the prior into a hierarchical sparse coding framework to learn a selective subset of codes representing holistic data properties that are robust to domain change and feature type variations. We then utilize label information on the sparse codes to perform classification, or in the absence of labels perform clustering, and obtain improved results on several previously studied heterogeneous adaptation datasets. We highlight the flexibility of our approach by accounting for multiple heterogeneous domains in training as well as in testing, and by considering the zero-shot domain transfer scenario where there are data categories in testing which are not seen during training. In that process we also empirically show how existing heterogeneous adaptation solutions can benefit from the findings of our study.

1 INTRODUCTION

Domain adaptation, which addresses the problem of change in data characteristics across training (source domain) and testing (target domain) datasets, has received substantial attention over the last few years [Daume III and Marcu, 2006, Gopalan et al., 2015]. Its utility for visual recognition problems has been adequately demonstrated by several efforts that have utilized principles from distance transforms [Saenko et al., 2010], max-margin methods [Duan et al., 2009], manifolds [Gong et al., 2012], dictionary learning [Qiu et al., 2012] among others. Many of these approaches have addressed settings where the source domain is labeled and the target domain is unlabeled (unsupervised adaptation), when the target domain also has partial labels (semi-supervised adaptation) and when there is more than one domain in the source and/or target (multiple domain adaptation). However they assume that data, across domains, is represented by same type of features with same dimensions. This is not always possible in practice, given the prevalence of multi-modal sensors, and there have been relatively few efforts in the literature addressing the heterogeneous domain adaptation (HDA) setting which allows different feature types across domains.

One of the earliest efforts is by [Dai et al., 2008] that proposed a translated learning framework using risk minimization principles to bridge data across different feature types. [Prettenhofer and Stein, 2010] utilized structural correspondence learning by identifying pivot features that share similarities across domains. While such models have restrictions on the type of applications where they can be deployed, [Shi et al., 2010] proposed a more generic heterogeneous spectral mapping framework that learns an embedding to obtain a common domain-invariant feature subspace that optimally represents data from each domain. Along similar lines, [Wang and Mahadevan, 2011] proposed a manifold alignment approach that bridges source and target domain manifolds through a latent space that in addition to preserving topology of each domain, maximizes the intra-class similarities and inter-class dis-similarities across domains. [Kulis et al., 2011] approached this problem by learning asymmetric kernel transformations that perform cross-domain data mapping using semantic similarity. More recently, [Li et al., 2014] proposed a feature augmentation strategy coupled with max-margin classifiers, whose formulation resembles multiple kernel learning that in turn guarantees global optimal solution, and [Yeh et al., 2014] studied the utility of canonical correlation subspaces to this problem. Most of these methods either address the HDA problem by learning projections for each domain onto a common latent space where certain properties are satisfied, or by learning feature mapping from one domain onto another directly.

We take a rather different approach which, instead of learning a *few complex transformations* to map domains, learns *several simpler transformations* that explain how data from different domains can be bridged. In learning such transformations we bring the domains to a common dimension using simpler generative information, instead of more complex objectives tied to the domain structure and data similarities. What we gain by doing so is a richer 'expanded' set of prior information which could be harnessed by hierarchical learning methodologies to perform inference. More specifically, given N domains, with each domain representing the data with different feature type (thus having different dimension), we group the data from each domain into k clusters based on their feature similarity and obtain generative subspaces corresponding to each cluster by doing principal component analysis (PCA). We fix all subspaces to have the same dimension p using a heuristic tied to the amount of energy preserved in doing the dimensionality reduction. We then perform parallel transport [Edelman et al., 1998] between subspaces in every domain pair and obtain several intermediate representations that describe how data across domains can be bridged. We subsequently project the data from each domain onto all these intermediate representations to obtain an expanded prior, and integrate it with hierarchical sparse coding [Jenatton et al., 2011] to obtain compact codes on which we use label information, if any, to perform cross-domain inference. More details are provided in Section 2. The construction of our approach has the following benefits.

Accommodating Unlabeled Data: In many practical situations, with the widespread availability of multi-modal data on the web, we have very few (or at times no) labeled data and lots of unlabeled data. Our approach can readily utilize such big unlabeled data as we rely on generative modeling in addressing the heterogeneous domain shift. By doing so, when the source and target domains contain the same categories/classes, our final inference can range from the classification scenario where we have labels for all categories in source domain and the target domain may or may not have partial labels, to the clustering scenario where both the source and target domains are unlabeled. The label information is utilized while training a discriminative classifier such as support vector machines (SVM) on the learnt sparse codes, and if no labels are available we perform clustering on the sparse codes using methods such as k-means.

Zero-Shot Domain Transfer: We can also address the zero-shot scenario in which there are categories in the target domain that are not present in the source domain. This is somewhat different from the scenarios discussed above where we at least had unlabeled data in source domain for all target categories to support inference models. We could handle such a zero-shot scenario as our model is generative and therefore the learned domain shift would have pertinent information for reasoning out new categories.

Multiple Heterogeneous Domains: Finally we can easily accommodate multiple heterogeneous domains in the source as well as in target since we obtain the expanded prior by doing parallel transport between each domain pair. This does not pose a computational bottleneck as we are eventually learning sparse codes in a hierarchical learning setting that could handle big data.

To the best of our knowledge, our proposed approach is the first to handle these varied aspects of the HDA problem, and while some existing methods could handle a subset of these in principle, we make explicit discussions on these practically relevant requirements. We tested our approach on existing heterogeneous adaptation datasets and obtained good performance improvement over previous results for diverse tasks such as object recognition, event classification, text categorization and sentiment analysis. Detailed experimental analysis is provided in Section 3, and concluding remarks are given in Section 4.

2 APPROACH

Problem Setting: We assume there are N heterogeneous domains $D = \{D_i\}_{i=1}^N$, where each domain $D_i = \{x_i^j, y_i^j\}_{j=1}^{n_i}$ contains n_i data samples with $x_i^j \in \mathbb{R}^{d_i}$ denoting the feature vector of dimension d_i and y_i^j denoting the corresponding label information (if any) belonging to one of M different categories. These domains could be partitioned into source and target domains depending on the problem situation. With this information, the goal of this work is to account for heterogeneous domain shift in inferring the labels of the unlabeled target domain data.

2.1 PRELIMINARIES

Before we proceed, we will first review relevant details about the tools we use in our approach.

Parallel Transport: We will be working on subspaces derived from the data, and we will generally have multiple subspaces extracted from each domain. In domain adaptation literature, the notion of geodesic on the Grassmann manifold has been used as a bridge to connect a pair of subspaces [Gopalan et al., 2011]. When we need to bridge two 'sets' of subspaces instead, parallel transport [Edelman et al., 1998] provides a way by learning multiple paths by which subspace sets can be bridged. More specifically, let $S_1 = \{S_1^i\}_i$ and $S_2 = \{S_2^i\}_i$ denote two sets of p-dimensional subspaces in $\mathbb{R}^d$ corresponding to domains D_1 and D_2 respectively, where each subspace say S_1^1 is a point on the Grassmannian $\mathcal{G}_{d,p}$. Let $g_A(t)$ denote the geodesic with the initial direction $A \in \mathbb{R}^{(d-p)\times p}$ connecting the means of S_1 and S_2, and $\bar{S}_1^1$ denote the tangent space representation of S_1^1 obtained using inverse exponential mapping computed at the mean of S_1. The parallel transport of $\bar{S}_1^1$ is then given as

$$\gamma\bar{S}_1^1(t) = Q_{S_1^1}\exp\left(t\begin{bmatrix}0 & A^T\\ -A & 0\end{bmatrix}\right)\begin{bmatrix}0\\ B\end{bmatrix} \qquad (1)$$

where exp is the matrix exponential, $Q_{S_1^1} \in SO(d)$ is the orthogonal completion of S_1^1, and $B \in \mathbb{R}^{(d-p)\times p}$ is the initial direction to reach from S_1^1 to the exponential map of $\bar{S}_1^1$. Similar directions can be obtained for all subspaces in the sets S_1 and S_2 using the above tangent space approximation. Please refer to [Edelman et al., 1998] for more details. We will discuss how to utilize these directions (bridges) for HDA in Section 2.2.

Hierarchical Sparse Coding: In sparse coding [Yang et al., 2009] the goal is to represent each input vector $x \in \mathbb{R}^p$ as a sparse linear combination of basis vectors. Given a stacked input data matrix $X \in \mathbb{R}^{p\times n}$, where n is the number of data, it seeks to minimize:

$$\underset{Z\in\mathcal{Z},C}{\arg\min}||X - ZC||_2^2 + \lambda\Omega(C) \qquad (2)$$

where $Z \in \mathbb{R}^{p\times r}$ is the dictionary of basis vectors, $\mathcal{Z}$ is the set of matrices whose columns have small ℓ_2 norm and $C \in \mathbb{R}^{r\times n}$ is the code matrix, λ is a regularization hyperparameter, and Ω is the regularizer. In hierarchical sparse coding, such a scheme is extended in a layered fashion using a combination of coding and pooling steps and we pursue the schema presented in [Jenatton et al., 2011]. Our modification comes in the way in which the dictionary is initialized and we present the details in Section 2.2.

While there has been an attempt in using parallel transport for unsupervised homogeneous domain adaptation [Shrivastava et al., 2014], our construction vastly differs from that work as we handle multiple heterogeneous domains without using label information to bridge the domain shift. Moreover, ours is the first approach to integrate parallel transport information with hierarchical sparse coding for adaptation problems.

2.2 PROPOSED HETEROGENEOUS ADAPTATION ALGORITHM

Step 1: We first bring data from all N domains, $D = \{D_i\}_{i=1}^N$, onto a common dimension d by performing PCA on each D_i and choosing the resultant subspace dimension as the largest dimension required among all N subspaces such that 90% of the signal energy is preserved for that decomposition. Then we project data from each domain onto its corresponding subspace. We now have d-dimensional data across all domains, say $\bar{X} = \{\bar{x}_i^j\}_{i,j}$, where i ranges from 1 to N and j ranges from 1 to n_i.

Step 2: From each domain D_i, we then derive k generative subspaces by partitioning $\{\bar{x}_i^j\}_{j=1}^{n_i}$ into k clusters using the k-means algorithm based on the similarity of the d dimensional features, and performing PCA on each cluster. We ensure all the subspaces are of dimension p, by choosing p as the largest dimension required for a subspace, amongst all subspaces obtained by doing k-means in each of the N domains, such that 90% of the signal energy is preserved by that decomposition[1]. Thus from every domain D_i we have a set of p-dimensional subspaces in $\mathbb{R}^d$ denoted by $S_i = \{S_i^j\}_{j=1}^k$. Each subspace in this set is a point on the Grassmann manifold, $\mathcal{G}_{d,p}$. Let $X \in \mathbb{R}^{p\times n}, n = \sum_{i=1}^N n_i$ denote the matrix containing the projections of each data in $\bar{X}$ onto its appropriate subspace in $S = \{S_i\}_{i=1}^N$. This is our input data matrix for sparse coding.

Step 3: We then perform parallel transport between S_i's using the method described in Section 2.1, and obtain a collection of directions between each pair of (S_i, S_j), i=1,..N-1, $j = i$+1,..,N. We uniformly sample points along these directions using exponential mapping, which results in new subspaces that have information on how domain shift information flows between domains. We project each data in $\bar{X}$ onto these subspaces to get the expanded prior $\mathcal{P} \in \mathbb{R}^{p\times r}$, which we in turn use to initialize the dictionary Z.

Step 4: Finally we perform hierarchical sparse coding [Jenatton et al., 2011] with the input data matrix X from Step 2 and the initial dictionary Z obtained from Step 3. At the output of each layer of hierarchical sparse coding, we apply Steps 2 and 3 obtain another set of expanded prior which is then used to complement the dictionary of the following layer. Let the final output (from the last layer) of hierarchical sparse coding corresponding to the original data X be denoted by $\hat{X} = \{\hat{x}_i^j\}_{i,j}$, and their corresponding label information (if any) is denoted as before by $Y = \{y_i^j\}_{i,j}$. Note that we have not used any label information thus far.

2.3 INFERENCE

We now perform cross-domain inference using the information contained in $W = (\hat{X}, Y)$. Note that W contains data from both source and target domains, and depending on the dataset we may have one or more domains in the source and target.

Classification: For the classification scenarios widely studied in HDA, source domain contains labeled data for all M categories, and the target domain may or may not have partial labels, and both the source and target domains

[1] While this is a simple heuristic in addressing variations in feature dimensions, we show that it works well empirically. As stated in the introduction, our main proposal for HDA is by generating an expanded prior on how these domains interact, rather than addressing domain shift 'during' the process of bringing domains to a common dimension as done by most existing methods.

have the same M categories. So we consider labeled data present in W to train a multi-class SVM [Crammer and Singer, 2002] with default parameters for linear kernel, and then use the SVM similarity score to classify the unlabeled target domain data into one of M categories. Classification accuracy is computed as the percentage of unlabeled target data that were correctly assigned their category label (using ground truth). Note that while we could have used the label information in any of the previous stages, be it during parallel transport or in sparse coding, we did not because we would like the learned cross-domain representations $\hat{X}$ to be generic to support other inference scenarios discussed next. Nevertheless, we make some observations regarding this later in Section 3.6.

Clustering: We also address the clustering scenario where both source and target domain data are unlabeled, and they have the same M categories. In this case we cluster all the data in $\hat{X}$ into M groups using k-means, and compute the clustering accuracy using a standard method of labeling each of the resulting clusters with the majority class label according to the ground truth, and measuring the number of mis-classifications made by each cluster grouping.

Zero-shot Learning: We finally account for the zero-shot learning scenario where the target domain has some categories that are not a part of the M source domain categories. For this case, we use labels for M categories in the source domain and (if available) in the target domain to train the SVM as discussed for the classification scenario. We then threshold the SVM similiarity score for the unlabeled target data, with the hope that if such data comes outside of the M source categories, the similarity score will be less. We then cluster such data using k-means to obtain groupings, with the number of clusters set to the number of new target categories known apriori, and evaluate the accuracy as discussed above for the clustering scenario. If we do not even know the number of new target categories, then it becomes difficult to quantify clustering accuracy.

3 EXPERIMENTS

We first discuss the classification scenario and experiment with the setup used by [Li et al., 2014] for the problems of heterogeneous object recognition, text categorization and sentiment analysis. These experiments have only one domain each in the source and target. We then consider the event classification experiment designed by [Chen et al., 2013] which consists of multiple source domains and a single target domain. We then provide a detailed analysis of the findings from our study.

3.1 OBJECT RECOGNITION

We work with the Office dataset [Saenko et al., 2010] that contains a total of 4106 images from 31 categories collected from three sources: amazon (object images downloaded from Amazon), dslr (high-resolution images taken from a digital SLR camera) and webcam (low-resolution images taken from a web camera). SURF features are extracted for all the images. The images from amazon and webcam are clustered into 800 visual words by using k-means. After vector quantization, each image is represented as a 800 dimensional histogram feature. Similarly, we represent each image from dslr as a 600-dimensional histogram feature. In the experiments, dslr is used as the target domain, while amazon and webcam are considered as two individual source domains. For training the SVM, we randomly select 20 labeled images per category for the source domain amazon, and 8 labeled images per category for webcam as source domain. For the target domain dslr, 3 labeled images are randomly selected from each category. The remaining target domain data is used for testing. We present results of our HDA approach in Table 1 along with other methods studied in [Li et al., 2014].

Methods	Source Domain	
	amazon	webcam
[Shi et al., 2010]	42.8±2.4	42.2±2.6
[Wang and Mahadevan, 2011]	53.3±2.3	53.2±3.2
[Kulis et al., 2011]	53.1±2.4	53.0±3.2
[Li et al., 2014]	55.4±2.9	54.3±3.6
Ours	62.1±1.7	61.5±2.1

Table 1: Mean and std. deviation of classification accuracy (%) on Object Recognition Dataset with target domain dslr.

3.2 TEXT CATEGORIZATION

We use the Reuters multilingual dataset [Amini et al., 2009] which contains about 11K newswire articles from 6 categories in 5 languages namely, English, French, German, Italian and Spanish. All documents are represented by using the TF-IDF feature. We perform PCA based on the TF-IDF features from each domain with 60% energy preserved and thus the features for each language have the following dimensions respectively, 1131, 1230, 1417, 1041 and 807. We consider Spanish as the target domain in the experiments and use each of the other four languages as individual source domains. For each category, we randomly sample 100 labeled documents from the source domain and either 10 or 20 labeled documents from the target domain to train the SVM. The remaining documents in the target domain are used as the test data. We report classification results of our HDA approach in Table 2 along with the results of the other approaches discussed in [Li et al., 2014].

3.3 SENTIMENT ANALYSIS

We use the Cross-Lingual Sentiment (CLS) dataset [Prettenhofer and Stein, 2010], which is an extended version of the Multi-Domain Sentiment Dataset [Blitzer et al., 2007] widely used for domain adaptation. It is collected from

Methods	Source Domain							
	10 labels per target class				20 labels per target class			
	English	French	German	Italian	English	French	German	Italian
[Shi et al., 2010]	54.7±7.4	55.0±9.4	58.0±7.9	59.4±3.7	65.7±3.1	64.2±4.2	64.6±3.6	65.8±2.3
[Wang and Mahadevan, 2011]	65.0±2.9	66.9±2.1	67.5±2.1	68.5±2.8	72.4±2.4	72.8±2.0	72.9±2.3	73.3±2.1
[Kulis et al., 2011]	65.7±2.7	66.9±1.7	68.7±2.9	67.9±2.8	72.9±2.0	73.5±1.8	74.7±1.6	74.0±2.0
[Li et al., 2014]	68.6±2.3	69.5±1.9	69.8±2.7	69.8±2.5	75.3±1.7	75.7±1.6	76.1±1.5	75.8±1.8
Ours	73.2±1.6	73.8±1.9	74.1±2.3	74.0±1.7	83.5±2.4	84.1±1.9	83.8±2.1	84.2±1.5

Table 2: Mean and std. deviation of classification accuracy (%) on Reuters Multilingual Dataset with target domain Spanish

Methods	Target Domain		
	German	French	Japanese
[Shi et al., 2010]	50.4±0.6	49.8±0.6	51.3±1.0
[Wang and Mahadevan, 2011]	64.6±1.9	65.7±1.8	64.4±1.8
[Kulis et al., 2011]	58.3±3.0	59.4±4.3	57.5±1.9
[Li et al., 2014]	66.5±2.2	66.9±2.1	64.2±2.5
Ours	71.2±2.1	71.5±1.8	70.9±1.5

Table 3: Mean and std. deviation of classification accuracy (%) on Cross-lingual Sentiment Dataset with source domain English.

Amazon and contains about 800,000 reviews of three product categories: Books, DVDs and Music, and written in four languages: English, German, French, and Japanese. The English reviews were sampled from the Multi-Domain Sentiment Dataset and reviews in other languages are crawled from Amazon. For each category and each language, the dataset is partitioned into a training set, a test set consisting of 2,000 reviews each. We take English as the source domain and each of the other three languages as an individual target domain in the experiment. We randomly sample 500 reviews from the training set of the source domain and 100 reviews from the training set of the target domain as the labeled data to train the SVM. The test set is the official test set for each category and each language. As with text categorization experiment, we extracted the TF-IDF features and performed PCA with 60% energy preserved to result in dimensions 715, 929, 964 and 874 respectively for the four languages discussed above. Results on this dataset are presented in Table 3.

3.4 EVENT CLASSIFICATION

We then worked on the event classification dataset of [Chen et al., 2013] that accounts for multiple heterogeneous source domains and a single target domain. The events pertain to six classes namely, birthday, picnic, parade, show, sports and wedding. Three source domains correspond to Google and Bing image search, and Youtube video search corresponding to these events. Kodak and CCV dataset serve as the individual target domain. The Google and Bing domains are represented by a 4000 dimensional bag-of-words codebooks learnt on SIFT features, while YouTube, Kodak and CCV datasets are represented by 6000 dimensional spatio-temporal features corresponding to histogram of oriented gradient, histogram of optical flow and motion boundary histogram. By considering the source domains to be labeled and target domain completely unlabeled we trained the SVM on the source domain samples to perform separate inference on CCV and Kodak datasets as the target domain. The cross-domain event classification results are presented in Table 4.

Methods	Target Domain	
	Kodak	CCV
[Bruzzone and Marconcini, 2010]	43.49	41.55
[Duan et al., 2012b]	44.21	38.56
[Duan et al., 2012a]	46.21	43.44
[Chen et al., 2013]	49.61	44.52
Ours	54.56	51.24

Table 4: Mean average precision (%) on the Event classification dataset with multiple heterogeneous source domains from Google image search, Bing image search and YouTube video search.

3.5 DISCUSSION

Clustering and Zero-shot Learning: We see that our HDA approach outperforms existing methods on diverse heterogeneous classification tasks. We also performed these experiments for the clustering scenario by not considering any labeled data from the source and target domains using the approach discussed in Section 2.3. The performance decreased by around 15% on average from the classification results. While this is reasonable since label information always helps in performing class-specific inference, it also shows that our generative heterogeneous model outputs $\hat{X}$ contain information agnostic to domain and feature variations and thus is relevant in grouping data categories. We verified this by performing a baseline using k-means after Step 1 (i.e. without the adaptation procedure) and the results dropped further by around 18% on average.

We then considered the zero-shot learning scenario, by considering the classification experiments and holding out 10%, 20% and 30% of the categories as being exclusive to the target which the source domain has not seen. As per the discussion in Section 2.3, we first thresholded the SVM similarity scores for data from these new categories. With the similarity score ranging from 0 (low) to 1(high), we tried three thresholds namely 0.3, 0.2 and 0.1. Ideally these new categories should have lower similarities since they are not part of the trained model. On average we obtained a 95% filtering accuracy with these thresholds, across all datasets discussed before, and then we grouped the filtered data to get an accuracy of 85% on average. Note that these results are only for the new data categories, which are much less in number than those considered for classification scenario and hence can not be compared with those results. As a sanity check we tested the unlabeled data from the target categories that are present in the source, and observed that their SVM similarity scores were always greater than 0.3 for all datasets. These results convey that our model outputs are quite useful in reasoning out never seen before categories, which is very important in practice.

Parameter Tuning: For all the results discussed thus far, we used $k = 10$ clusters within each domain, and uniformly sampled 10 points on the parallel transport directions. We tried other values 8 and 12 for both clusters and sampling on directions, and PCA energy of 80% and 85% in learning the generative subspaces from the domains, and found the results decreased at the most by 2%. We used default parameters for other tools we have employed in the approach. This sheds light on the robustness of our approach. It takes about 5 to 10 seconds on a single 2GHz machine to perform inference over the range of scenarios discussed here.

Design Choice Analysis: We now analyze the rationale behind some choices we made in the approach. Firstly, we tested whether hierarchical sparse coding is necessary, or will a single layer sparse coding be sufficient. We also tested how many layers are necessary by experimenting up to five layers. The results reported in the paper are with three layers, and when we used four and five layers, the results reduced by 2.5% and 3% on average, using two layers saw average performance reduction of around 8% and using one layer saw a reduction of 17%. This suggests that hierarchical feature learning is useful, and the results reach a plateau around three layers for our experiments.

Then we inspect whether learning multiple bridges across domains using parallel transport is necessary, or just a single bridge using the geodesic will suffice. Results using only the geodesic was inferior by around 21% which highlights the utility of parallel transport to the HDA problem. We also test whether we need to do parallel transport to obtain expanded prior on the outputs of each layer of hierarchical sparse coding, by just initializing the first layer with the prior and doing hierarchical feature learning on it. That resulted in a performance drop of about 12% on average.

Multiple Source and Multiple Target Domains: Our experiments so far contained single source and single target domain, or multiple source domains and single target domain. We now pursue multiple source domains and multiple target domains on these datasets, where possible. Given N domains, we try all possible combinations across source and target domains. For example, if we have four domains, we consider 1 to 3 source domains that are accompanied by 3 to 1 target domains respectively. Our results for such a setting improves the results discussed in Sec 3.1 to 3.4 by at least 3% and up to 12%. This explains the utility of our method for HDA with increasing availability of domains.

3.6 EXTENSIONS

In this section we discuss some extensions of our approach, by relaxing certain assumptions that were made to facilitate its generalizability to different adaptation settings.

Using Label Information In Modeling Stage: Till now the labels were used only in the classification stage (Sec 2.3) and not in the modeling stage (Sec 2.2) as we wanted the model to handle classification, clustering and zero-shot learning. But in cases where say, the goal is just classification and there are labels available for training the model, it makes sense to use them in the model building stage itself. To support such a scenario, we modify our approach (in Step 4) by performing discriminative hierarchical sparse coding. We use the method of [Ji et al., 2011] and feed it with data labels contained in Y. Thus, we obtain the sparse codes output $\hat{X}$ which in addition to minimizing the reconstruction error of the data samples, also separates samples belonging to one class from other classes. Let us call this Case A. With this modification, the performance for classification experiments reported in Section 3.1 to 3.4 improved by at least 3% and up to 15% on average.

Another way to utilize label information is in forming the clusters within each domain (Step 2). Instead of using the similarity of the d dimensional features to group the data into k clusters, we group the data using their labels into M clusters and then perform the remaining steps as outlined in Section 2.2. So in this case the parallel transport information will have a notion of class discrimination in traversing the domain shift. Let us call this Case B. This resulted in an average performance improvement of at least 1.8% and up to 9% for the classification experiments reported in Section 3.1 to 3.4. We then used Case A and Case B together, and this improved the results by at least 5% and up to 20% on average.

Integrating With Other Heterogeneous Adaptation Strategies: As mentioned in the introduction, one of the goals of our study was to see how a large number of simple transformations would fare against few complex transformations to handle heterogeneous domain shift, which was the reason to map all domains to a common dimension using simpler generative information (Step 1). This strategy was shown to be successful through detailed experiments. Now we study how to get the best of both approaches. For this, instead of Step 1, we use outputs from existing heterogeneous adaptation techniques which map different dimensions onto a common one using more involved objectives related to domain structure. We tried two such techniques, one based on spectral mapping [Shi et al., 2010] and the

Domain		Classification accuracy (in %), mean±std. deviation					
Source	Target	No labeled target data			Few labeled target data		
		[Mahsa et al., 2014]	[Long et al., 2014]	Ours	[Ni et al., 2013]	[Mahsa et al., 2013]	Ours
Caltech	Amazon	52.3±1.1	46.76	56.3±1.1	49.5±2.6	61.8±2.5	65.3±1.2
Caltech	Dslr	53.0±2.3	44.59	55.2±1.8	76.7±3.9	65.8±3.5	79.8±1.2
Amazon	Caltech	44.4±1.4	39.45	49.2±1.3	27.4±2.4	47.8±1.5	51.1±0.3
Amazon	Webcam	48.5±2.6	42.03	51.6±2.1	72.0±4.8	72.5±3.1	75.5±1.1
Webcam	Caltech	39.3±0.5	30.19	42.8±1.8	29.7±1.9	43.6±1.2	45.5±2.8
Webcam	Amazon	44.3±0.9	29.96	45.2±2.5	49.4±2.1	53.4±1.9	55.5±1.7
Dslr	Amazon	39.4±1.1	32.78	44.3±1.2	48.9±3.8	56.9±1.6	57.5±1.5
Dslr	Webcam	88.8±1.0	85.42	91.2±1.7	72.6±2.1	89.1±1.6	92.3±2.6

Table 5: Comparison with homogeneous adaptation methods on the Office-Caltech dataset with 10 object categories.

other based on manifold alignment [Wang and Mahadevan, 2011]. The mapped output of these techniques, where the data from all domains $\{D_i\}_{i=1}^N$ will have the transformed to the same dimension d, signifies $\bar{X}$ which is then fed into Step 2, and the remaining steps from Sec 2.2 and 2.3 are followed. This resulted in a performance improvement of at least 5% and up to 18%, and at least 4.2% and up to 16.5% while using [Shi et al., 2010] and [Wang and Mahadevan, 2011] respectively, for classification, clustering, and zero-shot learning experiments reported in Section 3.1 to 3.5. These results hold promise on the utility of our approach to existing adaptation solutions.

Heterogeneous View Of Homogeneous Adaptation: Encouraged by these observations, we considered homogeneous adaptation problems that have been extensively studied in the literature [Saenko et al., 2010], which assumes the data is represented by same features (same dimensions) in both source and target domains. The goal of our study here is to represent such data with many different features, with each feature forming a separate domain, and empirically investigate whether multiple features can make adaptation problems easier to handle. We experiment with different combinations of heterogeneous features in the source and target and at the same time making sure that the same feature type is not used for both source and target. The results presented below are averaged over such combinations.

We first consider the Office-Caltech dataset [Gong et al., 2012] for adaptive object recognition that contains 10 objects with four domains namely, amazon, dslr, webcam, and Caltech. The data is represented by bag-of-words codebooks learnt from SURF features. We additionally extracted histogram of oriented gradients, local binary patterns, local phase quantization, and GIST descriptors. Thus we have five domains each in the source and target domain. We then followed the adaptation protocol of [Gong et al., 2012] that first considered labeled data from source domain and no labels from target domain, and then allows few labeled samples from target as well. Our results are provided in Table 5.

We then worked on adaptive face recognition and considered the following facial features, image intensities in RGB and HSV color spaces, edge magnitudes and gradient direction, multi-scale block LBP, self-quotient image and Gabor wavelets. We first followed the protocol of [Shekhar et al., 2013] that used the Multi-PIE dataset [Gross et al., 2010] with images of 129 subjects in frontal pose as the source domain, and five other off-frontal poses as the target domain. Images under five illumination conditions across source and target domains were used for training with which images from remaining 15 illumination conditions in the target domain were recognized. Face recognition accuracy for this experiment is given in Table 6. We also performed an experiment on the PIE dataset [Sim et al., 2002] using the protocol of [Ni et al., 2013], where the domain change is caused by illumination and blur. Frontal pose faces of 34 subjects under 11 different illumination conditions formed the source domain, while the unlabeled target domain consisted of frontal images of the same subjects under 10 other lighting conditions that were also blurred using four kernels, a Gaussian with standard deviation of 3 and 4 and motion blur with length 9, angle 135° and length 11, angle 45°. The results are provided in Table 7.

Method	Mean classification accuracy (%) Target domain pose				
	15°	30°	45°	60°	75°
[Sharma and Jacobs, 2011]	92.1	89.7	88.0	86.1	83.0
[Sharma et al., 2012]	99.7	99.2	98.6	94.9	95.4
[Yang et al., 2011]	96.8	90.6	94.4	91.4	90.5
[Shekhar et al., 2013]	98.4	98.2	98.9	99.1	98.8
Ours	99.5	99.3	99.1	99.4	98.9

Table 6: Comparison with homogeneous adaptation methods on CMU Multi-PIE dataset for face recognition across pose and lighting variations with few labeled target data.

Finally we performed adaptation experiments for action recognition on the IXMAS multi-view action dataset [Weinland et al., 2007] that contains eleven action categories including walk, kick and throw. Each action was performed three times by twelve actors taken from five different views, which include four side views and one top view. From each action video we extracted the following features, histograms of oriented gradients and optical flow (HOG/HOF), 3DHOG that is based on 3D gradi-

Method	Mean classification accuracy (%)			
	$\sigma = 3$	$\sigma = 4$	$len = 9$	$len = 11$
[Ahonen et al., 2008]	66.5	32.9	73.8	62.1
[Gopalan et al., 2011]	70.9	60.3	72.4	67.9
[Gong et al., 2012]	78.5	77.7	82.4	77.7
[Ni et al., 2013]	80.3	77.9	85.9	81.2
Ours	86.3	85.2	87.9	87.2

Table 7: Comparing homogeneous adaptation methods on CMU PIE dataset for face recognition across blur and lighting variations with no labeled target data. σ: std. deviation of the Gaussian blur, and len: length of linear motion blur.

ent orientations, and extended SURF descriptor for videos. The experiment was to recognize actions with the source and target domain pertaining to different views, thereby making 20 source-target combinations. The average cross-view recognition accuracy results for all these view combinations are given in Table 8. The correspondence mode contains matched instances across source and target domains, partial labels mode having fewer target labels and the non-discriminative virtual views (NDVV) comprising of partially-labeled and unlabeled target data, as studied in [Li and Zickler, 2012].

All these results show our method outperforming other homogeneous adaptation approaches, which suggests that such approaches could, in some form, benefit by expanding the type of features used to represent the data. The features we have used are only somewhat representative of the vast literature and more of them can be used with our approach.

4 CONCLUSION

We have approached the problem of bridging heterogeneous domains by learning a large set of intermediate representations, born out of simpler generative information, and integrated them with hierarchical feature learning mechanisms to perform inference pertaining to classification, clustering and zero-shot learning. We demonstrated superior empirical performance over existing methods on a wide range of problems involving different data modalities such as images, videos and natural language. We also discussed the utility of our design choices, and the robustness of the approach in dealing with multiple domains, feature types, unlabeled data and new unseen data categories. While our approach offers an alternative to many existing heterogeneous solutions that address domain shift through a latent space modeling, it also opens up opportunities for harnessing the benefits of the two strategies, which we highlighted through some initial studies. Such a mechanism could eventually pave way for establishing error bounds on the nature of heterogeneous shifts an approach can handle, within a reasonable set of domain shift assumptions reflecting practical data acquisition constraints.

References

Timo Ahonen, Esa Rahtu, Ville Ojansivu, and J Heikkila. Recognition of blurred faces using local phase quantization. In *ICPR*, 2008.

Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, 2009.

John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE TPAMI*, 32:770–787, 2010.

Chao-Yeh Chen and Kristen Grauman. Inferring unseen views of people. In *CVPR*, 2014.

Lin Chen, Lixin Duan, Dong Xu, and Dong Xu. Event recognition in videos by learning from heterogeneous web sources. In *CVPR*, 2013.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.

Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, 2008.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

Lixin Duan, Ivor W Tsang, Dong Xu, and Stephen J Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009.

Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, 2012a.

Lixin Duan, Dong Xu, and Ivor W Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Trans Neural Networks and Learning Systems*, 23:504–518, 2012b.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20: 303–353, 1998.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

Method	Mean classification accuracy (%)		
	correspondence	partial labels	NDVV
[Liu et al., 2011]	75.3	-	-
[Chen and Grauman, 2014]	78.8	-	-
[Li and Zickler, 2012]	81.2	61.2	61.1, 26.0
[Zhang et al., 2013]	85.8	69.0	69.2, 35.3
Ours	89.5	74.3	75.3, 45.4

Table 8: Comparison with homogeneous adaptation methods on the IXMAS dataset for cross-view action recognition.

Raghuraman Gopalan, Ruonan Li, Vishal M Patel, and Rama Chellappa. Domain adaptation for visual recognition. *Foundations and Trends® in Computer Graphics and Vision*, 8(4):285–378, 2015.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28:807–813, 2010.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *JMLR*, 12:2297–2334, 2011.

Zhengping Ji, Wentao Huang, Garrett Kenyon, and Luis Bettencourt. Hierarchical discriminative sparse coding via bidirectional connections. In *IJCNN*, 2011.

Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

Ruonan Li and Todd Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, 2012.

Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE TPAMI*, 36:1134–1148, 2014.

Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *CVPR*, 2014.

Baktashmotlagh Mahsa, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.

Baktashmotlagh Mahsa, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Domain adaptation on the statistical manifold. In *CVPR*, 2014.

Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, 2013.

Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *ACL*, 2010.

Qiang Qiu, Vishal M Patel, Pavan Turaga, and Rama Chellappa. Domain adaptive dictionary learning. In *ECCV*. 2012.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*. 2010.

Abhishek Sharma and David W Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*, 2011.

Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.

Sumit Shekhar, Vishal M Patel, Hien V Nguyen, and Rama Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, 2013.

Xiaoxiao Shi, Qi Liu, Wei Fan, Philip S Yu, and Ruixin Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *ICDM*, 2010.

Ashish Shrivastava, Sumit Shekhar, and Vishal M Patel. Unsupervised domain adaptation using parallel transport on grassmann manifold. In *WACV*, 2014.

Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *FG*, 2002.

Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 2011.

Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007.

Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

Meng Yang, David Zhang, and Xiangchu Feng. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011.

Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23, 2014.

Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, Shuang Liu, and Cunzhao Shi. Cross-view action recognition via a continuous virtual path. In *CVPR*, 2013.