RECOGNIZING HETEROGENEOUS CROSS-DOMAIN DATA VIA GENERALIZED JOINT DISTRIBUTION ADAPTATION

Yuan-Ting Hsieh[*1], *Shi-Yen Tao*[*1], *Yao-Hung Hubert Tsai*[2], *Yi-Ren Yeh*[3], *Yu-Chiang Frank Wang*[2]

[1]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2]Research Center for IT Innovation,Academia Sinica, Taipei, Taiwan
[3]Department of Mathematics, National Kaohsiung Normal University, Kaohsiung, Taiwan

{b01901046,b01901055}@ntu.edu.tw, y.h.huberttsai@gmail.com, yryeh@nknu.edu.tw, ycwang@citi.sinica.edu.tw

ABSTRACT

In this paper, we propose a novel algorithm of *Generalized Joint Distribution Adaptation* (G-JDA) for heterogeneous domain adaptation (HDA), which associates and recognizes cross-domain data observed in different feature spaces (and thus with different dimensionality). With the objective to derive a domain-invariant feature subspace for relating source and target-domain data, our G-JDA learns a pair of feature projection matrices (one for each domain), which allows us to eliminate the difference between projected cross-domain heterogeneous data by matching their marginal and class-conditional distributions. We conduct experiments on cross-domain classification tasks using data across different features, datasets, and modalities. We confirm that our G-JDA would perform favorably against state-of-the-art HDA approaches.

Index Terms— heterogeneous domain adaptation, object recognition, text categorization

1. INTRODUCTION

In the areas of machine learning and pattern recognition, domain adaptation (DA) deals with the learning of data collected across domains. For example, one needs to recognize the object in an image captured by a smartphone, while the labeled training data is collected from the Internet. As a result, training and test data would exhibit distinct feature distributions. DA thus aims to exploit rich label information observed in a source domain, so that the data of interest in the target domain can be processed or recognized accordingly [1, 2].

When relating cross-domain data, two different settings are generally considered: semi-supervised or unsupervised DA. The former allows a small number of labeled data to be collected in the target domain, while the latter only observes unlabeled target-domain data during the adaptation process. Nevertheless, existing DA approaches either focus on deriving a common feature space for associating cross-domain data [1, 2, 3], or choose to identify/weight representative data instances (i.e., landmarks) [4, 5, 6] for performing adaptation.

It is worth noting that, most existing DA approaches require the source and target-domain data to be represented by the same type of features (and thus with the same feature dimensionality). This is referred to as homogeneous domain adaptation. When cross-domain data are described by distinct features (e.g., color vs. texture), or if such data are collected from different domains (e.g., image vs. text), one would encounter the challenging problem of *heterogeneous domain adaptation* (HDA). In order to relate heterogenous cross-domain data, HDA typically requires a small number of labeled data in the target domain. Moreover, depending on the presence of unlabeled target-domain data during the adaptation process, *supervised* or *semi-supervised* settings can be considered for HDA [7, 8]. Nevertheless, existing homogeneous DA approaches cannot be directly applied for solving HDA problems.

With recent research attention gradually shifts to HDA, a number of solutions have been proposed. For example, some researchers choose to learn a feature transformation, which either project data from one domain to the other, or project cross-domain data to a common subspace for adaptation [9, 10, 11, 12, 7]. On the other hand, aiming at adapting the classification model, others propose to match the classifiers or prediction models trained on different heterogeneous domains [13, 8].

In this paper, we propose *Generalized Joint Distribution Adaptation* (G-JDA), which is a feature-transformation based learning algorithm for *semi-supervised* HDA. Inspired by a recent work of [3], we aim to learn a pair of feature transformation (one for each domain), which allows us to derive a domain-invariant feature space for associating heterogeneous cross-domain data. When deriving such domain-specific feature transformation, our objective is to simultaneously match cross-domain marginal and conditional feature distributions, which would relate cross-domain data with adaptation and classification performance guarantees.

The contributions of this paper are highlighted below:

- We propose Generalized Joint Distribution Adaptation (G-JDA) for heterogeneous domain adaptation, in which source and target-domain data are collected from different domains and feature spaces.

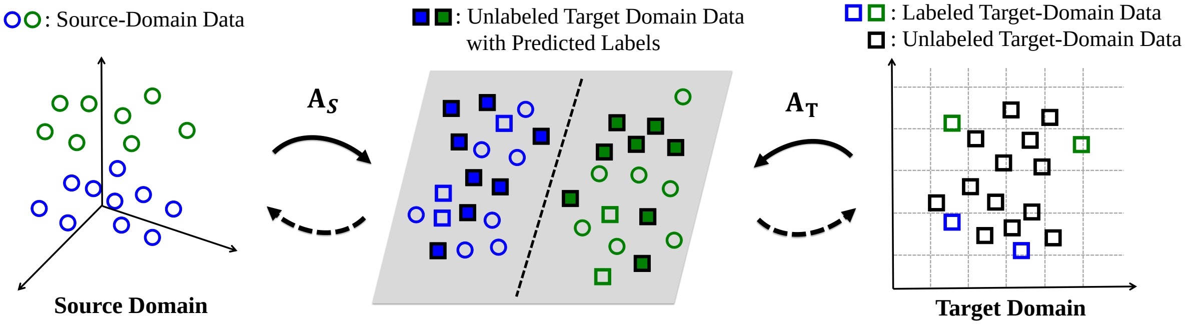

○○: Source-Domain Data
■■: Unlabeled Target Domain Data with Predicted Labels
□□: Labeled Target-Domain Data
□: Unlabeled Target-Domain Data
$\mathbf{A}_S$
$\mathbf{A}_T$
Source Domain
Target Domain

Fig. 1. Illustration of Generalized Joint Distribution Adaptation (G-JDA) for heterogeneous domain adaptation (HDA).

- Our G-JDA learns a pair of feature transformation, which allows one to match cross-domain marginal and conditional data distributions in a domain-invariant feature space for adaptation and classification.

- Our experiments confirm that our G-JDA is able to perform favorably against state-of-the-art HDA approaches on the task of classification across across features, datasets, and modalities.

2. RELATED WORK

Most existing DA approaches consider a homogeneous setting, in which both source and target-domain data are described by the same type of features. With the goal to derive a common feature space for associating cross-domain data, Pan *et al.* [1] proposed transfer component analysis (TCA), which learns a projection matrix by matching marginal probability distributions of cross-domain data based on the *Maximum Mean Discrepancy* [14] (MMD) criterion. Later, Long *et al.* [3] extended TCA to Joint Distribution Adaptation (JDA), which jointly minimizes the difference between the marginal and conditional probability distributions of cross-domain data. On the other hand, Gong *et al.* [6] proposed a landmark selection method, which first selects the source-domain instances with similar distributions as those of the target-domain instances, followed by the learning of feature projection via Geodesic Flow Kernel (GFK) [15].

Heterogeneous domain adaptation (HDA) deals with cross-domain data collected from different domains and described by distinct types of features. Existing HDA methods can be divided into two categories. One aims to transform heterogeneous data from one domain to the other for adaptation purposes [9, 12, 13]. For example, Kulis *et al.* [9] proposed asymmetric regularized cross-domain transformation (ARC-t), which can be viewed as nonlinear metric learning with particular data similarity constraints. Hoffman *et al.* [12] presented a Max-Margin Domain Transform (MMDT) method, which transforms the target-domain data to the source domain while deriving a max-margin classifiers. Instead of mapping cross-domain data, Zhou *et al.* [13] proposed sparse heterogeneous feature representation (SHFR) for learning a sparse transformation matrix, which is able to transform the classifiers across domains.

The other category of HDA approaches choose to learn separate projection matrices, which transform cross-domain data into a common feature subspace [11, 10, 16, 7, 8]. For example, Shi *et al.* [11] proposed heterogeneous spectral mapping (HeMap) to derive mapping matrices based on spectral embedding. Wang and Mahadevan [10] chose to solve domain adaptation by manifold alignment (DAMA), and derived projection matrices by manifold alignment with data locality constraints. Duan *et al.* [16] proposed heterogeneous feature augmentation (HFA), which first produces two transformation matrices to project the data into a common feature space, followed by the learning of classifiers in that space.

As noted in Section 1, in addition to the presence of labeled data in source and target domains, semi-supervised HDA also observes the unlabeled target-domain data during the adaptation process. Recent works on semi-supervised HDA show that improved adaptation and classification performance can be further achieved. For example, Li *et al.* [7] extended HFA to a semi-supervised version (i.e., SHFA) by incorporating unlabeled target-domain data in the training stage. Xiao *et al.* [8] proposed a semi-supervised subspace co-projection method (SCP) for HDA, which projects data across domains into a common latent subspace. By observing cross-domain data with MMD criterion, their method is able to derive the learning model while preserving data consistency in the resulting feature space.

3. OUR PROPOSED METHOD

3.1. Problem Setting and Notations

We first define the semi-supervised HDA problem, and introduce the notations which will be used in the remaining of this paper. Let $\mathcal{D}_S = \{\mathbf{X}_S, Y_S\} = \{\mathbf{x}_S^i, y_S^i\}_{i=1}^{n_S}$ denote the source-domain data, where $\mathbf{x}_S^i \in \mathbb{R}^{d_S}$ indicates the d_S-dimensional source-domain instance, and y_S^i is its corresponding label from the label set $\mathcal{L} = \{1, 2, .., C\}$. Similarly, we have $\mathcal{D}_T = \{\mathbf{X}_T, Y_T\} = \{\mathbf{x}_T^i, y_T^i\}_{i=1}^{n_T}$ as the target-domain data where $\mathbf{x}_T^i \in \mathbb{R}^{d_T}$ and $y_T^i \in \mathcal{L}$. For semi-supervised HDA, we further define $\mathcal{D}_T$ as an integration of labeled and unlabeled data subsets $\mathcal{D}_L = \{\mathbf{X}_L, Y_L\} =$

$\{\mathbf{x}_L^i, y_L^i\}_{i=1}^{n_L}$ and $\mathcal{D}_U = \{\mathbf{X}_U, Y_U\} = \{\mathbf{x}_U^i, y_U^i\}_{i=1}^{n_U}$, respectively. Target-domain labels $\{y_L^i\}_{i=1}^{n_L}$ are known in advance, but $\{y_U^i\}_{i=1}^{n_U}$ are to be predicted. Recall that, HDA deals with source and target-domain data in different feature spaces and exhibiting distinct distributions. Thus, we have $\mathbb{R}^{d_S} \neq \mathbb{R}^{d_T}$ and marginal distributions $\mathcal{P}_S(\mathbf{X}_S) \neq \mathcal{P}_T(\mathbf{X}_T)$.

It is worth noting that, semi-supervised HDA allows a sufficient number of labeled data to be collected in the source domain, but only few labeled target-domain data are available (i.e., $n_L \ll \{n_S, n_U\}$). In addition, we assume that at least one labeled instance is available for each class c from $\mathcal{L}$ in both source and target domains. In a nutshell, the main goal of HDA is to predict the labels of unlabeled target-domain data $\{y_U^i\}_{i=1}^{n_U}$ by leveraging the information across heterogeneous domains.

3.2. Generalized Joint Distribution Adaptation (G-JDA)

As noted in Section 2, *Joint Distribution Adaptation* [3] has been successfully applied to domain adaptation tasks. However, JDA can only deal with cross-domain data lying in the same (homogeneous) feature space, and thus it cannot be directly applied to HDA.

Inspired by [10, 7], we aim at finding distinct feature transformation for source and target-domain data, so that such cross-domain data can be projected into a domain-invariant feature space for adaptation and recognition. When deriving the above feature transformation, our goal is to match cross-domain data distributions, so that the domain difference (or mismatch) in the resulting feature space can be eliminated.

We now detail our proposed algorithm. As noted above, we aim to learn transformation matrices $\mathbf{A}_S \in \mathbb{R}^{d_S \times d_K}$ and $\mathbf{A}_T \in \mathbb{R}^{d_T \times d_K}$ for $\mathbf{X}_S$ and $\mathbf{X}_T$, respectively, so that *both* marginal and conditional distributions across domains can be matched (i.e., $\mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S) \approx \mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T)$ and $\mathcal{P}_S(Y_S|\mathbf{A}_S^\top \mathbf{X}_S) \approx \mathcal{P}_T(Y_T|\mathbf{A}_T^\top \mathbf{X}_T)$). Note that d_K denotes the dimensionality of the resulting feature space.

Since $\mathcal{P}_S(Y_S|\mathbf{A}_S^\top \mathbf{X}_S)$ and $\mathcal{P}_T(Y_T|\mathbf{A}_T^\top \mathbf{X}_T)$ cannot be estimated directly, we turn to calculate $\mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S|Y_S)$ and $\mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T|Y_T)$ instead [3]. As a result, our proposed *Generalized Joint Distribution Adaptation* (G-JDA) can be formulated as

$$\begin{aligned} \min_{\mathbf{A}_S, \mathbf{A}_T} \; & E_{mar}(\mathbf{A}_S, \mathbf{A}_T) + \sum_{c=1}^{C} E_{cond}^{(c)}(\mathbf{A}_S, \mathbf{A}_T) \\ & + \lambda \left(\|\mathbf{A}_S\|^2 + \|\mathbf{A}_T\|^2 \right) \\ \text{s.t.} \quad & \hat{\mathbf{X}}\mathbf{H}\hat{\mathbf{X}}^\top = \mathbf{I}, \end{aligned} \tag{1}$$

where $\hat{\mathbf{X}} = [\mathbf{A}_S^\top \mathbf{X}_S, \mathbf{A}_T^\top \mathbf{X}_T]$ are the projected cross-domain data and E_{mar} and $E_{cond}^{(c)}$ indicate the matching of cross-domain marginal and conditional distributions (of class c), respectively. Parameters λ regularizes the derivation of transformation, and $\mathbf{I}$ is the identity matrix. It is worth noting that, our G-JDA enforces the constraint of $\hat{\mathbf{X}}\mathbf{H}\hat{\mathbf{X}}^\top = \mathbf{I}$, where $\mathbf{H} = \mathbf{I}_{n_S+n_T} - \frac{1}{n_S+n_T}\mathbf{1}_{n_S+n_T}$, and $\mathbf{1}_{n_S+n_T}$ is the matrix with all elements equal to one. This constraint is to preserve the variance of the projected cross-domain data, which implies additional data discriminating ability [1, 3].

By applying *Maximum Mean Discrepancy* [14] (MMD) as the criterion of measuring distribution similarity, E_{mar} can be calculated by the distance between the two empirical sample means [3]. In other words, the difference between $\mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S)$ and $\mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T)$ can be estimated as follows:

$$E_{mar}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbf{A}_S^\top \mathbf{x}_S^i - \frac{1}{n_T} \sum_{j=1}^{n_T} \mathbf{A}_T^\top \mathbf{x}_T^j \|^2. \tag{2}$$

Similarly, the E_{cond} term is approximated by measuring the distance between $\mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S|Y_S = c)$ and $\mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T|Y_T = c)$. Thus, we define E_{cond} as follows:

$$E_{cond}^{(c)}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S^c} \sum_{i=1}^{n_S^c} \mathbf{A}_S^\top \mathbf{x}_S^{i,c} - \frac{1}{n_T^c} \sum_{j=1}^{n_T^c} \mathbf{A}_T^\top \mathbf{x}_T^{j,c} \|^2, \tag{3}$$

where $\mathbf{x}_S^{i,c}$ denotes the i-th source-domain instance of class c, and n_S^c is the total number of such instances in that class. Note that, when matching cross-domain data during the adaptation process, we apply the predicted labels (i.e., *pseudo labels*) for the unlabeled target-domain data. Thus, the original semi-supervised setting in the target domain can be converted into a supervised one. Thus, we have $\mathbf{x}_T^{j,c}$ and n_T^c denote the j-th target-domain instance of class c and the total number of such instances, respectively. In the following subsection, we will detail how we solve (1) for HDA.

3.3. Optimization of G-JDA

To solve the optimization problem of (1), we first define the data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_S & \mathbf{0}_{d_S \times n_T} \\ \mathbf{0}_{d_T \times n_S} & \mathbf{X}_T \end{pmatrix},$$

and an augmented transformation matrix $\mathbf{A} = [\mathbf{A}_S; \mathbf{A}_T]$. We have a label set defined as $Z = \{z_i\}_{i=1}^{n_S+n_T} = \{\{y_S^i\}_{i=1}^{n_S}, \{y_T^i\}_{i=1}^{n_T}\}$, in which target-domain labels $\{y_T^i\}_{i=1}^{n_T} = \{\{y_L^i\}_{i=1}^{n_L}, \{\widehat{y}_U^i\}_{i=1}^{n_U}\}$. We note that, y_L^i and $\widehat{y}_U^i$ indicate the true and pseudo labels for the target-domain instances, respectively.

As suggested in [1, 3] , we can rewrite (2) as

$$E_{mar} = tr(\mathbf{A}^\top \mathbf{X} \mathbf{M}_0 \mathbf{X}^\top \mathbf{A}), \tag{4}$$

where $tr(\cdot)$ denotes the trace sum, and the each entry in the matrix $\mathbf{M}_0 \in \mathbb{R}^{(n_S+n_T)\times(n_S+n_T)}$ is calculated as

$$(\mathbf{M}_0)_{ij} = \begin{cases} \frac{1}{n_S n_S} & \text{if } i, j \le n_S \\ \frac{1}{n_T n_T} & \text{if } i, j > n_S \\ \frac{-1}{n_S n_T} & \text{otherwise.} \end{cases}$$

Similarly, we simplify (3) into the following formulation:

$$E_{cond}^{(c)} = tr(\mathbf{A}^\top \mathbf{X} \mathbf{M}_c \mathbf{X}^\top \mathbf{A}), \tag{5}$$

where $\mathbf{M}_c \in \mathbb{R}^{(n_S+n_T)\times(n_S+n_T)}$ with each entry

$$(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{n_S^c n_S^c} & \text{if } i,j \leq n_S \text{ and } z_i = z_j = c \\ \frac{1}{n_T^c n_T^c} & \text{if } i,j > n_S \text{ and } z_i = z_j = c \\ \frac{-1}{n_S^c n_T^c} & \text{if } \begin{cases} i \leq n_S, j > n_S \\ i > n_S, j \leq n_S \end{cases} \text{and } z_i = z_j = c \\ 0 & \text{otherwise.} \end{cases}$$

We note that, the constraints in (1) can be rewritten as $\mathbf{A}^\top \mathbf{X} \mathbf{H} \mathbf{X}^\top \mathbf{A} = \mathbf{I}$, where $\mathbf{H} = \mathbf{I}_{n_S+n_T} - \frac{1}{n_S+n_T}\mathbf{1}_{n_S+n_T}$ and $\mathbf{1}$ is a matrix which all element equals to 1.

With the above derivations, we turn the original problem of (1) into the following compact version:

$$\min_{\mathbf{A}=[\mathbf{A}_S;\mathbf{A}_T]} \sum_{i=0}^{C} tr(\mathbf{A}^\top \mathbf{X} \mathbf{M}_i \mathbf{X}^\top \mathbf{A}) + \lambda \left(\|\mathbf{A}_S\|^2 + \|\mathbf{A}_T\|^2 \right)$$
$$\text{s.t.} \quad \mathbf{A}^\top \mathbf{X} \mathbf{H} \mathbf{X}^\top \mathbf{A} = \mathbf{I}. \tag{6}$$

To solve the transformation $\mathbf{A}$, we first derive the Lagrange function of (6), and take the derivative of this Lagrange function with respect to $\mathbf{A}$ and set it to zero. This allows us to solve $\mathbf{A}$ by determining the d_K *smallest* eigenvectors of the following generalized eigenvalue decomposition problem:

$$(\mathbf{X} \sum_{i=0}^{c} \mathbf{M}_i \mathbf{X}^\top + \mathbf{R})\mathbf{A} = \psi \mathbf{X} \mathbf{H} \mathbf{X}^\top \mathbf{A}, \tag{7}$$

where $\mathbf{R} \in \mathbb{R}^{n_S+n_T} = \begin{pmatrix} \lambda \mathbf{I}_{n_S} & \mathbf{0}_{n_S\times n_T} \\ \mathbf{0}_{n_T\times n_S} & \lambda \mathbf{I}_{n_T} \end{pmatrix}$ and ψ is the Lagrange multipliers.

3.4. G-JDA for Semi-Supervised HDA

When applying G-JDA for solving semi-supervised HDA problems, we need to assign pseudo labels for unlabeled target-domain data, so that the matching of cross-domain class-conditional data distributions can be performed. To start this adaptation process, we simply consider all the source and target-domain data with ground truth label information to initialize the transformation $\mathbf{A}$ from (7). Thus, no predicted labels for $\{\mathbf{x}_U^i\}_{i=1}^{n_U}$ will be utilized. Once this initialization stage is complete, we use the observed $\mathbf{A}_S, \mathbf{A}_T$ to project source and target-domain data into the resulting common feature space. In this space, we learn linear SVMs [17] using projected labeled cross-domain data for predicting the pseudo labels $\{y_U^i\}_{i=1}^{n_U}$ for the unlabeled target-domain instances. With all the pseudo-labels are obtained, we then update the transformations $\mathbf{A}_S, \mathbf{A}_T$ by solving (7). In other words, we iterate between the learning of transformation $\mathbf{A}$ and updating of the pseudo labels $\{\hat{y}_U^i\}_{i=1}^{n_U}$ until convergence. The use of our proposed G-JDA for HDA is now summarized in Algorithm 1.

Algorithm 1 Generalized Joint Distribution Adaptation

```
Input: Source-domain data $\mathcal{D}_S = \{\mathbf{x}_S^i, y_S^i\}_{i=1}^{n_S}$, labeled target-domain data $\mathcal{D}_L = \{\mathbf{x}_L^i, y_L^i\}_{i=1}^{n_L}$, unlabeled target-domain data $\{\mathbf{x}_U^i\}_{i=1}^{n_U}$, dimension $d_K$, and parameter $\lambda$
Initialize $\mathbf{A}_S, \mathbf{A}_T$ in (7) using $\mathcal{D}_S$ and $\mathcal{D}_L$
while not converge do
    Learn linear SVMs using projected labeled cross-domain data
    Update pseudo labels $\{\hat{y}_U^i\}_{i=1}^{n_U}$
    Calculate $\{\mathbf{M}_i\}_{i=0}^{c}$ and solve (7) for updating $\mathbf{A}_S, \mathbf{A}_T$
end while
Output: $\mathbf{A}_S, \mathbf{A}_T$, and $\{\hat{y}_U^i\}_{i=1}^{n_U}$
```

4. EXPERIMENTS

4.1. Datasets and Parameter Settings

To evaluate the adaptation and classification performance of our proposed G-JDA, we consider two benchmark datasets for experiments: **Office + Caltech-256** [2, 18] and **Multilingual Reuters Collection** [19, 20]. The former is for cross-domain object recognition, and the latter is for cross-lingual text categorization.

The **Office** dataset [2] contains three sub-datasets: Amazon(A) (images from the Internet), Webcam (W) (low-resolution images captured by webcams), and DSLR (D) (high-resolution images captured by digital cameras). Each sub-dataset has 31 categories of daily office objects. On the other hand, **Caltech-256** (C) [18] consists of 256 object categories. Following [2], we select *ten* overlapping object categories from **Office** and **Caltech-256** for experiments. Moreover, two types of features are used: *SURF* [21] and *DeCAF* [22]. The *SURF* feature is extracted from each image and converted into a Bag-of-Words (BOW) model using a codebook of 800 visual words. The latter is a novel deep-learning based feature with 4096 dimensionality.

Multilingual Reuters Collection [19, 20] is the dataset typically applied for multi-lingual text categorization. It contains about 11K articles from 6 categories in 5 languages (English, French, Italian, German, and Spanish). Following previous HDA works [16, 13, 7], we describe each article by BOW with TF-ITF, and perform PCA with 60% energy preserved. The resulting dimensions of the five languages English, French, Italian, German, and Spanish are 1131, 1230, 1417, 1041, and 807, respectively.

As for the parameters, we fix the regularization parameter $\lambda = 1$ and feature dimension $d_K = 80$ in our work. To compare our performance with other HDA approaches, we consider the baseline approach of SVM$_t$, which simply learns standard SVMs using the labeled target-domain data

Table 1. Classification results (%) with standard deviations for cross-feature object recognition.

S,T	SVM$_t$	DAMA	HFA	MMDT	G-JDA
DeCAF$_6$ to *SURF*					
A, A	38.9±0.7	40.6±0.6	43.5± 0.5	45.0±0.7	**50.3±0.7**
W, W	52.3±1.2	57.0±0.9	62.8±0.7	57.7±0.8	**63.8±0.9**
C, C	27.6±0.6	28.8±0.6	31.7±0.6	30.8±0.7	**33.7±0.8**
SURF to *DeCAF*$_6$					
A, A	82.3±0.9	86.7±0.5	87.8±0.3	86.5±0.5	**92.3±0.2**
W, W	84.7±1.0	87.3±0.9	89.2±0.8	88.2±0.6	**89.4±0.9**
C, C	70.3±1.3	72.8±0.8	77.3±0.7	76.4±0.7	**86.7±0.5**

Table 2. Classification results (%) with standard deviations for object recognition across datasets and features.

S,T	SVM$_t$	DAMA	HFA	MMDT	G-JDA
DeCAF$_6$ to *SURF*					
A, D	51.3±0.8	51.7±0.9	56.8±0.6	53.9±0.6	**56.9±0.7**
W, D		56.2±0.9	**56.5±0.5**	52.3±0.9	55.5±0.8
C, D		52.8±0.8	56.9±0.6	55.2±0.8	**57.2±1.0**
SURF to *DeCAF*$_6$					
A, D	89.7±0.8	90.5±0.5	91.3±0.6	90.5±0.6	**94.3±0.7**
W, D		89.4±0.6	90.8±0.7	90.8±0.6	**95.0±0.4**
C, D		89.8±0.6	90.6±0.8	91.2±0.6	**92.8±0.8**

(i.e., no adaptation). We also consider three state-of-the-art HDA methods: DAMA [10], HFA [16], and MMDT [12].

4.2. Object Recognition Across Feature Spaces

We first consider object recognition across feature spaces, i.e., *SURF* and *DeCAF*) in source and target domains, respectively. We only show the results of Amazon, Webcam, and Caltech, since the size of DSLR is much smaller than others. Following the settings in [7, 12, 16], 20 and 3 images per object category are chosen as labeled images in source and target domains, respectively. The remaining images in the target domain are unlabeled and to be recognized.

We list the average recognition results of 20 random trials in Table 1. It is obvious that, without performing adaptation, SVM$_t$ was not able to achieve satisfactory performance. Our G-JDA, on the other hand, reported promising results, and performed favorably against DAMA, HFA, and MMDT.

4.3. Object Recognition Across Datasets and Features

We now consider a more challenging object recognition task using data across datasets and feature spaces. Following the settings of 4.2, we choose Amazon, Webcam, and Caltech as the source domains, and DSLR as the target domain due to its limited size. The results from 20 trials are shown in Table 2. Again, from this table, our proposed G-JDA achieved the highest average accuracy among all the methods. Thus, the effectiveness of our G-JDA for cross-domain object recognition can be successfully verified.

4.4. Cross-Lingual Text Categorization

To evaluate our adaptation and recognition performance on the **Multilingual Reuters Collection** dataset, we follow [7, 16] and consider English, French, Italian, and German as the source domains. Thus, Spanish is viewed as the target domain. For the source domain, we randomly choose 100 articles per category as labeled articles. As for the target-domain, we choose numbers of $\{10,20\}$ per category as labeled data, and randomly choose 500 articles per category in the remaining subset as unlabeled data. The average results of 20 trials are presented listed in Table 3. By comparing the results shown in Table 3, it is clear that our G-JDA again performed favorably against the baseline and state-of-the-art HDA approaches. Thus, the above experiments support the use of our G-JDA for cross-lingual text categorization.

4.5. Analysis on Convergence and Parameter Sensitivity

Finally, we discuss the issues of optimization convergence and parameter sensitivity. Considering the cross-feature object recognition using Caltech, in which *SURF* and *DeCAF* features are applied for describing source and target-domain data, we present the recognition performance over the number of iterations in Figure 2(a). From this figure, we see that our algorithm converged within 3 iterations, and thus our G-JDA is computationally feasible for performing such HDA tasks.

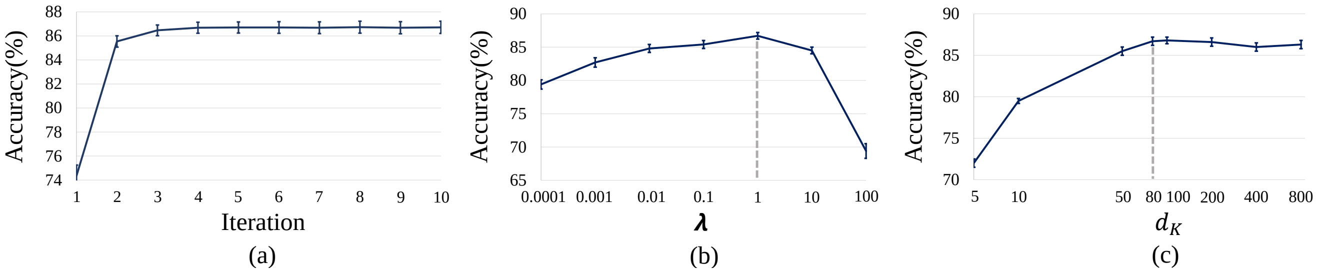
To evaluate the parameter sensitivity, we further plot the recognition performance versus different λ values in Figure 2(b). We also consider the performance over different dimensionality numbers d_K, and show the results in Figure 2(c). From the above figures, it can be seen that our default choices of $\lambda = 1$ and $d_K = 80$ are reasonable. Nevertheless, we did not fine tune such parameters for each test in our experiments, and thus both the effectiveness and robustness of our G-JDA can be successfully verified.

5. CONCLUSION

We proposed *Generalized Joint Distribution Adaptation* (G-JDA) for recognizing heterogeneous cross-domain data. Given source-domain labeled data and a small number of labeled data in the target domain, our G-JDA is able to associate heterogeneous cross-domain data in this semi-supervised setting, and classifies the remaining unlabeled data in the target domain. By learning a pair of feature transformation matrices for source and target-domain data, our G-JDA derives a domain-invariant subspace by matching the marginal and conditional distributions of projected cross-domain data. As a result, one can apply labeled data for recognizing the projected unlabeled target-domain data accordingly. Our experiments on several cross-domain classification tasks verified the effectiveness and robustness of our proposed G-JDA for HDA.

Table 3. Classification results with standard deviations for cross-lingual text categorization with Spanish as the target domain.

Source Articles	# labeled target domain data / category = 10					# labeled target domain data / category = 20				
	SVMt	**DAMA**	**HFA**	**MMDT**	**G-JDA**	**SVMt**	**DAMA**	**HFA**	**MMDT**	**G-JDA**
English	67.3±0.6	66.1±0.8	67.7±0.5	68.9±0.6	**69.4±0.8**	74.5±0.4	73.1±0.4	74.3±0.4	75.5±0.4	**76.0±0.7**
France		61.5±1.0	68.1±0.5	69.3±0.6	**70.5±0.7**		72.2±0.5	74.8±0.3	75.3±0.5	**76.8±0.8**
German		63.4±0.8	68.4±0.5	68.7±0.5	**69.6±1.0**		69.3±0.7	74.4±0.4	75.7±0.5	**76.8±0.7**
Italian		65.3±0.8	68.0±0.5	69.5±0.6	**70.1±1.0**		73.0±0.4	75.0±0.4	76.3±0.4	**76.6±0.7**



(a) (b) (c)

Fig. 2. Analysis on (a) iteration, (b) regularization parameter λ, and (c) common subspace dimensionality d_K, respectively.

ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST103-2221-E-001-021-MY2, MOST104-2221-E-017-016, and MOST104-2119-M-002-039.

6. REFERENCES

[1] S. J. Pan *et al.*, "Domain adaptation via transfer component analysis," in *IEEE T-NNLS*, 2011.

[2] K. Saenko *et al.*, "Adapting visual category models to new domains," in *ECCV*, 2010.

[3] M. Long *et al.*, "Transfer feature learning with joint distribution adaptation," in *IEEE ICCV*, 2013.

[4] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," in *IEEE T-PAMI*, 2010.

[5] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for adaptation," in *NIPS*, 2011.

[6] B. Gong *et al.*, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *ICML*, 2013.

[7] W. Li *et al.*, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," in *IEEE T-PAMI*, 2014.

[8] M. Xiao and Y. Guo, "Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation," in *ECML*, 2015.

[9] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *IEEE CVPR*, 2011.

[10] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI*, 2011.

[11] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *IEEE ICDM*, 2010.

[12] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko, "Efficient learning of domain-invariant image representations," in *ICLR*, 2013.

[13] J. T. Zhou *et al.*, "Heterogeneous domain adaptation for multiple classes," in *AISTATS*, 2014.

[14] A. Gretton *et al.*, "A kernel method for the two-sample-problem," in *NIPS*, 2006.

[15] B. Gong *et al.*, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE CVPR*, 2012.

[16] L. Duan *et al.*, "Learning with augmented features for heterogeneous domain adaptation," in *ICML*, 2012.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, 2011.

[18] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[19] M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," in *NIPS*, 2009.

[20] M. Ueffing *et al.*, "NRC's portage system for wmt 2007," in *ACL workshop*, 2007.

[21] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *ECCV*. 2006.

[22] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.