Unsupervised Domain Adaptation with Label and Structural Consistency

Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang, *Member, IEEE*

***Abstract*—Unsupervised domain adaptation deals with scenarios in which labeled data are available in the source domain, but only unlabeled data can be observed in the target domain. Since the classifiers trained by source-domain data would not be expected to generalize well in the target domain, how to transfer the label information from source to target-domain data is a challenging task. A common technique for unsupervised domain adaptation is to match cross-domain data distributions, so that the domain and distribution differences can be suppressed. In this paper, we propose to utilize the label information inferred from the source domain, while the structural information of the unlabeled target-domain data will be jointly exploited for adaptation purposes. Our proposed model not only reduces the distribution mismatch between domains, improved recognition of target-domain data can be achieved simultaneously. In the experiments, we will show that our approach performs favorably against state-of-the-art unsupervised domain adaptation methods on benchmark datasets. We will also provide convergence, sensitivity and robustness analysis, which support the use of our model for cross-domain classification.**

***Index Terms*—Domain adaptation, unsupervised, structure discovery, label propagation.**

I. Introduction

IN many real-world classification tasks, one cannot expect that the data to be recognized always exhibit the same or similar distribution as the training data does. The distribution mismatch between training and test data typically comes from the fact that such data are collected from different domains (e.g., videos captured by cameras at different views, images taken by cameras with different resolutions, etc.) [32], [25]. For the above scenarios, classifiers learned from training data cannot be expected to generalize well when recognizing test data.

To address the aforementioned problems, researchers advance the idea of *domain adaptation* and aim at associating cross-domain data for recognition purposes. If the difference between source and target domains can be eliminated, test data observed in the target domain can be recognized by source-domain training data accordingly. Thus, domain adaptation and its applications has been widely exploited in computer vision [25], [20], [11] and machine learning [31], [9], [22] communities.

Depending on the availability of labeled data in the target domain, domain adaptation approaches can be generally divided into two different categories. For *semi-supervised domain adaptation* [25], [5], one can collect source-domain labeled data in advance, but only a small amount of labeled data can be observed in the target domain. Given such cross-domain data and label information, the task is to recognize the remaining target-domain data. On the other hand, *unsupervised domain adaptation* [11], [20] deals with totally unlabeled target-domain data, with only labeled data available in the source domain. In this paper, we focus on unsupervised domain adaptation.

Among existing domain adaptation methods, the most common strategy is to derive feature representations for reducing the domain differences [20], [31], [22], [11], so that recognition can be performed in the resulting feature spaces. While some advocated the adaptation of marginal distributions across data domains [22], [31], several works have been proposed to further adapt both marginal and conditional distributions for improved performance [35], [20]. It is worth noting that, however, adaptation of conditional distributions is not trivial for unsupervised domain adaptation problems. This is because that, only unlabeled data can be observed in the target domain. Therefore, how to properly transfer the source-domain label information to the target domain for associating cross-domain data becomes a challenging task.

As noted above, we particularly address the unsupervised domain adaptation problem in this paper. As illustrated in Figure 1, we propose to exploit the *structural* information of target-domain data, together with the *label* information transferred from the source domain for performing domain adaptation. Based on *Maximum Mean Discrepancy* (MMD) [13], we utilize the above information and approach domain adaptation by solving a *label-propagation* based optimization task, aiming at matching cross-domain marginal and conditional feature distributions. As verified in our experiments, the proposed method not only exhibits improved domain adaptation ability, it also outperforms several state-of-the-art unsupervised domain adaptation approaches in terms of cross-domain visual classification performance.

We now summarize the contributions of this paper:

- We propose to exploit the local structural information of the target-domain data, together with the label information inferred from the source domain, for matching cross-domain feature distributions. As a result, improved adap-

C.-A. Hou is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 (e-mail: c.a.andyhou@gmail.com).

Y.-H. H. Tsai is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: y.h.huberttsai@gmail.com).

Y.-R. Yeh is with the Department of Mathematics, National Kaohsiung Normal University, Kaohsiung 824, Taiwan (e-mail: yryeh@nknu.edu.tw).

Y.-C. F. Wang is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: ycwang@citi.sinica.edu.tw).

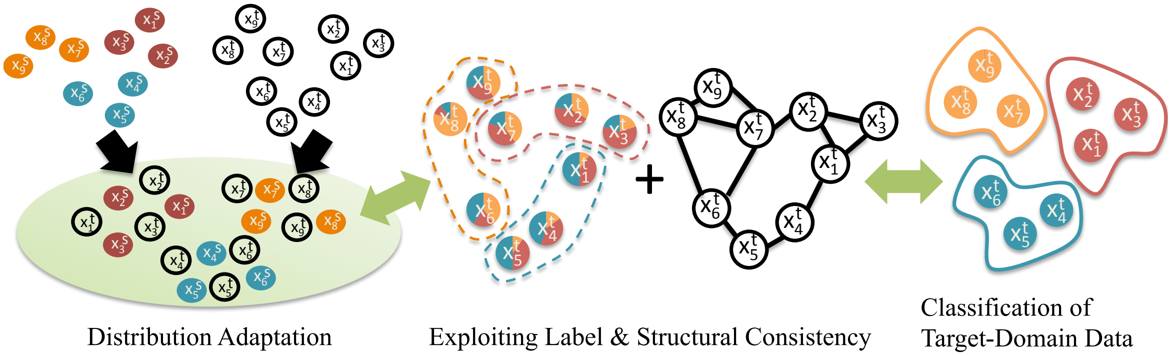

Distribution Adaptation
+
Exploiting Label & Structural Consistency
Classification of Target-Domain Data

Fig. 1. Overview of our proposed method for unsupervised domain adaptation. Note that different colors indicate the class labels, while $\mathbf{x}^s$ and $\mathbf{x}^t$ denote data in source and target domains, respectively.

tation and classification performance can be achieved. (Section III)

- Compared to several state-of-the-art unsupervised domain adaptation methods like [22], [20], [21], improved performance on several cross-domain classification tasks can be obtained by our method. Our approach also performs favorably against the case in which a substantial amount of ground truth target-domain labels are observed for matching cross-domain feature distributions. (Section IV)
- Additional experiments on convergence analysis, parameter sensitivity, and robustness to initialization errors are provided. These supporting materials further verify the effectiveness of our proposed model for unsupervised domain adaptation. (Section IV)

II. Related Works

Semi-supervised domain adaptation: To associate source and target domain data during the learning process, semi-supervised domain adaptation allows the users to collect either a small amount of target-domain labeled data [15], [23], [29] or a number of cross-domain instance pairs [27], [28], [16]. For example, Shekhar et al. [29] addressed cross-domain image classification problems by constructing a common domain-adaptive dictionary through the labeled data in the target domain. By exploiting cross-domain data correspondence information, Huang and Wang [16] proposed to learn a coupled dictionary for relating source and target domains for classification and synthesis purposes.

With the use of cross-domain label or correspondence information, adaptation problems with large domain differences (e.g., cross-pose face recognition [28], [27]) or those dealing with distinct features across domains (e.g., image-to-text classification [34]) can be possibly solved. Nevertheless, for semi-supervised domain adaptation tasks, the performance would be highly dependent upon the amount of label/correspondence information available.

Unsupervised domain adaptation: Many real-world classification problems deal with training and test data collected from different domains. Since it is often expensive to collect labeled data in the target domain or cross-domain data pairs for training purposes, this results in the challenging task of unsupervised domain adaptation. For unsupervised domain adaptation, the users are able to collect labeled data in the source domain, but only unlabeled test data can be observed in the target domain (and no cross-domain instance pair is available during training either) [11], [9], [12], [10], [8].

Due to the absence of target-domain label information, an underlying observation for unsupervised domain adaptation is that the source and target domains have different data distributions but still exhibit relatedness [2]. As a result, with the goal to eliminate the bias across different domains [32], recent works advance Maximum Mean Discrepancy (MMD) [13] to match cross-domain data distributions [17], [22], [20], [9], [1], [10]. Nevertheless, modeling data distributions across domains is not a trivial task. Previously, researchers chose to model and match cross-domain marginal distributions and assume the conditional distributions are the same across domains (i.e., covariate shift [30]). For example, Huang et al. [17] proposed kernel mean matching by weighting source-domain data, so that the mean difference between cross-domain data can be minimized. Pan et al. [22] proposed Transfer Component Analysis (TCA) to determine a low-dimensional embedding for cross-domain data, so that matching of cross-domain data can be performed accordingly. Based on TCA, Long et al. [21] further presented Transfer Feature Matching (TJM), which combines instance reweighting and distribution adaptation techniques for improved adaptation.

However, adapting marginal distributions only might not be sufficient to associate cross-domain data for classification purposes. To address this issue, Long et al. [20] proposed Joint Distribution Adaptation (JDA) to match both marginal and conditional data distributions in the derived common feature space. Since no label information can be observed in the target domain, they applied the prediction outputs of source-domain classifiers as the pseudo labels of target-domain data, while the classifiers were updated during their adaptation process.

Since the direct use of such pseudo labels might not be preferable due to possible domain mismatch, we not only transfer label information but further exploit target-domain structural information during the adaptation process. Inspired by recent semi-supervised learning techniques presented in [7], [24], [36], we approach the original problem of unsupervised domain adaptation by solving a label-propagation based optimization task. This allows us to better associate cross-domain marginal and conditional feature distributions, while additional robustness is introduced for alleviating possible errors due to domain mismatch. By jointly solving the adaptation and recognition tasks in a unified framework, improved recognition performance can be expected in the target domain.

III. Our Proposed Method

A. Motivation

We start from the problem definition, and introduce the notations which will be used in the following of this paper. Let $\mathcal{D}_S = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_M^s, y_M^s)\} = \{\mathbf{X}_S, \mathbf{y}_S\}$, where $\mathbf{X}_S \in \mathbb{R}^{d \times M}$ represents M d-dimensional data in the source domain, and each entry in $\mathbf{y}_S \in \mathbb{R}^{M \times 1}$ indicates the corresponding label (from 1 to C). On the other hand, we have N unlabeled instances observed in the target domain (with the same feature dimension), i.e., $\mathcal{D}_T = \{\mathbf{x}_n^t\}_{n=1}^N = \mathbf{X}_T \in \mathbb{R}^{d \times N}$. Thus, we determine the cross-domain data matrix as $\mathbf{X} = [\mathbf{X}_S, \mathbf{X}_T] \in \mathbb{R}^{d \times (M+N)}$. By assuming that both source and target domains contain data of the same C classes of interest, the goal of our work is to predict the label vector $\mathbf{y}_T \in \mathbb{R}^{N \times 1}$ for classification purposes, while each element in $\mathbf{y}_T$ is the assigned class label for the corresponding instance in the target domain.

In this paper, we perform transfer feature learning for unsupervised domain adaptation (i.e., only labeled and unlabeled data are available in source and target domains, respectively). We not only eliminate domain differences for associating cross-domain data, we also need to leverage label information from source to target domains for recognition purposes. To address the above issues, we propose and integrate two components highlighted below, which will be detailed in Sections III-B and III-C, respectively:

i) Adaptation of joint feature distributions. Let $P_S(\mathbf{X}_S)$ and $P_T(\mathbf{X}_T)$ as marginal distributions of data in source and target domains, respectively, and we have $P_S(\mathbf{y}_S|\mathbf{X}_S)$ and $P_T(\mathbf{y}_T|\mathbf{X}_T)$ as the corresponding conditional distributions. As noted in [22], [20], we typically have $P_S(\mathbf{X}_S) \neq P_T(\mathbf{X}_T)$ and $P_S(\mathbf{y}_S|\mathbf{X}_S) \neq P_T(\mathbf{y}_T|\mathbf{X}_T)$. Thus, our goal is to match both cross-domain marginal and conditional distributions, so that recognition of target-domain data can be performed accordingly. Minimizing the differences between cross-domain marginal and conditional distributions effectively matches the joint distribution of $P(\mathbf{X}, \mathbf{y})$. Following JDA [20], we do not approximate and align $P(\mathbf{y})$ due to the assumption of equal prior probabilities for the UDA problem of interest. To be more precise (and as seen in our experiments), we consider the same label numbers across domains, with equal numbers of instances to be observed across different categories.

ii) Exploitation of cross-domain data with label and structural consistency. We advance the technique of label propagation [36] for domain adaptation. More specifically, we utilize label information inferred from the source domain and observe the target-domain data structure for performing adaptation. This allows us to tackle the unsupervised domain adaptation problem with improved recognition of target-domain data.

B. Distribution Adaptation

As highlighted in Section III-A, the primary goal of this work is to match both marginal and conditional feature distributions of cross-domain data, so that data in the target domain can be classified accordingly. However, as noted in [20], since the modeling of conditional distributions $P(\mathbf{y}_S|\mathbf{X}_S)$ and $P(\mathbf{y}_T|\mathbf{X}_T)$ is not explicitly applicable, an alternative way is to observe and adapt class-conditional distributions $P(\mathbf{X}_S|\mathbf{y}_S)$ and $P(\mathbf{X}_T|\mathbf{y}_T)$ based on their sufficient statistics.

In our work, we aim at determining a feature transformation Φ for cross-domain data, so that both $P_S(\Phi(\mathbf{X}_S)) \approx P_T(\Phi(\mathbf{X}_T))$ and $P_S(\Phi(\mathbf{X}_S)|\mathbf{y}_S) \approx P_T(\Phi(\mathbf{X}_T)|\mathbf{y}_T)$ can be satisfied. For simplicity, we apply empirical criteria of MMD [13] for adapting the above distribution. To be more precise, we need to minimize the difference between feature distributions is calculated by the distance between data means in a reproducing kernel Hilbert space (RKHS):

$$\begin{aligned} &Dist(P_S(\mathbf{X}_S), P_T(\mathbf{X}_T)) + Dist(P_S(\mathbf{X}_S|\mathbf{y}_S), P_T(\mathbf{X}_T|\mathbf{y}_T)) = \\ &\left\| \frac{1}{M} \sum_{i=1}^{M} \phi(\mathbf{x}_i^s) - \frac{1}{N} \sum_{j=1}^{N} \phi(x_j^t) \right\|_{\mathcal{H}}^2 + \\ &\sum_{c=1}^{C} \left\| \frac{1}{|\mathcal{D}_S^{(c)}|} \sum_{\mathbf{x}_i^s \in \mathcal{D}_S^{(c)}} \phi(\mathbf{x}_i^s) - \frac{1}{|\hat{\mathcal{D}}_T^{(c)}|} \sum_{\mathbf{x}_i^t \in \hat{\mathcal{D}}_T^{(c)}} \phi(\mathbf{x}_i^t) \right\|_{\mathcal{H}}^2 , \end{aligned} \tag{1}$$

where $Dist$ measures the distance between feature distributions, and ϕ is the feature transformation induced by universal kernels. In (1), we have $\mathcal{D}_S^{(c)} = \{\mathbf{x}_i^s : y_i^s = c\}$ indicate source-domain data of class c, and $\hat{\mathcal{D}}_T^{(c)} = \{\mathbf{x}_i^t : \hat{y}_i^t = c\}$ as those in the target domain with the same predicted label.

For unsupervised domain adaptation, a major challenge is to transfer the label information from source to target domains when reducing domain biases. Later in Section III-C, we will explain how we determine the label information for target-domain data, so that the adaptation of the above conditional distributions can be achieved.

By applying the kernel tricks, we can rewrite (1) as $tr(\mathbf{KL}) + \sum_{c=1}^{C} tr(\mathbf{KL}_c)$, where $\mathbf{K} \in \mathbb{R}^{(M+N)\times(M+N)}$ indicates the kernel matrix of data matrix $\mathbf{X}$, and

$$L_{ij} = \begin{cases} \frac{1}{M^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_S \\ \frac{1}{N^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_T \\ \frac{-1}{MN}, & \text{otherwise.} \end{cases}$$

$$(L_c)_{ij} = \begin{cases} \frac{1}{|D_S^{(c)}|^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_S^{(c)} \\ \frac{1}{|\hat{D}_T^{(c)}|^2}, & \mathbf{x}_i, \mathbf{x}_j \in \hat{\mathcal{D}}_T^{(c)} \\ \frac{-1}{|D_S^{(c)}||\hat{D}_T^{(c)}|}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_S^{(c)}, \mathbf{x}_j \in \hat{\mathcal{D}}_T^{(c)} \\ \mathbf{x}_i \in \hat{\mathcal{D}}_T^{(c)}, \mathbf{x}_j \in \mathcal{D}_S^{(c)} \end{cases} \\ 0, & \text{otherwise.} \end{cases}$$

As noted in [22], the above optimization problem with respect to $\mathbf{K}$ requires high computational costs. In our work, we utilize the empirical kernel map [26] as suggested in [22], [20] to rewrite the $\mathbf{K}$ into $(\mathbf{K}\mathbf{K}^{-1/2})(\mathbf{K}^{-1/2}\mathbf{K})$. Then, we derive a lower-dimensional space in terms of $\mathbf{K}$ by determining the projection matrices $\mathbf{W}'$ and $\mathbf{W}$ (both of size $(M+N) \times k$, and $k << d$). Now, the kernel matrix can be written as $(\mathbf{K}\mathbf{K}^{-1/2}\mathbf{W}')(\mathbf{W}'^{\top}\mathbf{K}^{-1/2}\mathbf{K}) = \mathbf{K}\mathbf{W}\mathbf{W}^{\top}\mathbf{K}$, where $\mathbf{W} = \mathbf{K}^{-1/2}\mathbf{W}'$. As a result, by replacing the original $\mathbf{K}$ with $\mathbf{K}\mathbf{W}\mathbf{W}^{\top}\mathbf{K}$, the original objective function of (1) turns into:

$$\begin{aligned}&\min_{\mathbf{W}} tr(\mathbf{W}^{\top}\mathbf{K}\mathbf{L}\mathbf{K}^{\top}\mathbf{W}) + \sum_{c=1}^{C} tr(\mathbf{W}^{\top}\mathbf{K}\mathbf{L_c}\mathbf{K}^{\top}\mathbf{W}) + \lambda\|\mathbf{W}\|_F^2 \\ &\text{s.t. } \mathbf{W}^{\top}\mathbf{K}\mathbf{H}\mathbf{K}^{\top}\mathbf{W} = \mathbf{I}.\end{aligned} \tag{2}$$

It can be seen that, the first two terms in (2) are associated with the adaptation of marginal and conditional distributions, respectively. The sum of the these two terms corresponds to the MMD distance between the cross-domain data. The third term in (2) regularizes the projection $\mathbf{W}$, weighted by parameter λ. The centering matrix $\mathbf{H}$ in the constraint of (2) is defined as $\mathbf{H} = \mathbf{I} - \frac{1}{M+N}\mathbf{1}$, in which $\mathbf{1}$ is the matrix of ones. As noted in [22], [20], adding this constraint would preserve the data variance after adaptation, which implies and introduces additional data discriminating ability into the learned model $\mathbf{W}$. By applying Lagrange techniques, we can rewrite the objective function of (2) into the following Lagrangian function:

$$\begin{aligned}\mathcal{L}(\mathbf{W}, \mathbf{\Psi}) \equiv\ & tr(\mathbf{W}^{\top}(\mathbf{K}\mathbf{L}\mathbf{K}^{\top} + \mathbf{K}\sum_{c=1}^{C}\mathbf{L}_c\mathbf{K}^{\top} + \lambda\mathbf{I})\mathbf{W}) \\ & + tr((\mathbf{I} - \mathbf{W}^{\top}\mathbf{K}\mathbf{H}\mathbf{K}^{\top}\mathbf{W})\mathbf{\Psi}),\end{aligned} \tag{3}$$

where $\mathbf{\Psi}$ is a diagonal matrix with Lagrange Multipliers (i.e., $\mathbf{\Psi} = diag(\psi_1, ..., \psi_k) \in \mathbb{R}^{k\times k}$). By setting the derivative of (3) with respective to $\mathbf{W}$ equal to zero, we approach the original optimization problem by solving the following generalized eigen-decomposition problem:

$$(\mathbf{K}\mathbf{L}\mathbf{K}^{\top} + \mathbf{K}\sum_{c=1}^{C}\mathbf{L}_c\mathbf{K}^{\top} + \lambda\mathbf{I})\mathbf{W} = \mathbf{K}\mathbf{H}\mathbf{K}^{\top}\mathbf{W}\mathbf{\Psi}. \tag{4}$$

Taking the $k-$smallest eigenvectors from (4) would satisfy (2), which determines the optimal solution of $\mathbf{W}$ (recall that $k << d$). Once $\mathbf{W}$ is obtained, we project cross-domain data into the resulting k-dimensional latent space i.e., $\mathbf{Z} = \mathbf{W}^{\top}\mathbf{K} = [\mathbf{Z}_S, \mathbf{Z}_T] \in \mathbb{R}^{k\times(M+N)}$, where $\mathbf{Z}_S$ and $\mathbf{Z}_T$ represent the transformed data projected from source and target domains, respectively. In other words, the data matrix $\mathbf{Z}$ can be viewed as adapted cross-domain data with matched marginal and conditional distributions.

C. Exploiting Label and Structural Consistency for Unsupervised Domain Adaptation

Due to the lack of label information in the target domain, matching of cross-domain conditional distributions is a challenging task. Without proper prediction of class labels for the target-domain data $\hat{\mathcal{D}}_T$, adapting the conditional distributions for cross-domain data cannot be achieved.

To solve the above problem, we propose to take the knowledge which is exploited across domains into the adaptation process, with the goal of suppressing domain biases with cross-domain recognition guarantees. To begin with, we apply SVM-based classifiers [4], [33] for estimating the class posterior probability of the transformed target-domain data $\mathbf{Z}_T$. For the setting of unsupervised domain adaptation, these SVM classifiers are trained by labeled source-domain data in the transformed feature space (i.e., $\mathbf{Z}_S$). Based on the estimated posterior probabilities, we construct an *uncertain label matrix* $\mathbf{Y} \in \mathbb{R}^{N\times C}$ for target-domain instances, in which each entry is defined as:

$$Y_{ij} = \begin{cases} 2p(y_i^t = j|z_i^t) - 1, & \begin{cases} p(y_i^t = j|z_i^t) > \delta \\ p(y_i^t = j|z_i^t) = \max\limits_{c} p(y_i^t = c|z_i^t) \end{cases} \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

It is worth noting that, the posterior probability $p(y_i^t = j|z_i^t)$ indicates how likely the projected target-domain instance z_i^t belongs to class j. Obviously, we have $-1 \leqslant Y_{ij} \leqslant 1$, and a larger Y_{ij} value implies that the instance of interest is of the corresponding class. The parameter δ controls the number of uncertain labels to be transferred from the source domain (i.e., the aggressiveness of label propagation). For simplicity, we set δ equal to the lower quantile (i.e., 25%) of the maximum posterior probabilities observed from each target-domain instance. In other words, 75% of target-domain data will be assigned the predicted uncertain labels for adaptation purposes. Later in the experiments, we will provide additional remarks on our choice of δ.

Once the above uncertain label matrix is constructed, we effectively set a semi-supervised setting for the target-domain data. However, unlike standard semi-supervised learning problems in which a portion of the data are given specific class labels, we do not directly take the labels predicted by source-domain data due to possible domain mismatch. In other words, we cannot directly apply existing semi-supervised techniques, since they cannot deal with data collected from different domains. This is the reason why the use of our uncertain label matrix together with feature adaptation is preferable, which offers additional robustness in adapting and assigning class labels.

In addition to the use of uncertain labels predicted from the source domain, we further take the data structure observed in the target domain into our adaptation process. This allows us to better determine the target-domain labels for improved recognition. To observe target-domain structural information, we advance graph-based semi-supervised learning by constructing a k-nearest neighbors (k-NN) graph over target-domain data [36], [24], [7]. We note that, we choose to construct this k-NN based graph in the transformed space (i.e., $\mathbf{Z}_T$). The use of the transformed space not only allows us to better observe data structural information due to reduced feature dimensions, the learned transformation model $\mathbf{W}$ also exhibits capabilities in eliminating biases across source and target domains. This is why improved recognition of target-domain data can be expected.

Based on the above observations, we calculate the distance between target-domain data pairs as $d(\mathbf{z}_i^t, \mathbf{z}_j^t) = \|\mathbf{z}_i^t - \mathbf{z}_j^t\|$, and apply Gaussian kernels for converting such distances into similarity scores: $s(\mathbf{z}_i^t, \mathbf{z}_j^t) = exp(-d(\mathbf{z}_i^t, \mathbf{z}_j^t)/2\sigma^2)$. With this structural information determined, the k-NN based similarity matrix $\mathbf{E} \in \mathbb{R}^{N\times N}$ can be formulated, in which each entry is:

$$E_{ij} = \begin{cases} s(\mathbf{z}_i^t, \mathbf{z}_j^t), & \text{if } \mathbf{z}_j^t \text{ is one of k-NN of } \mathbf{z}_i^t \text{ and } i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

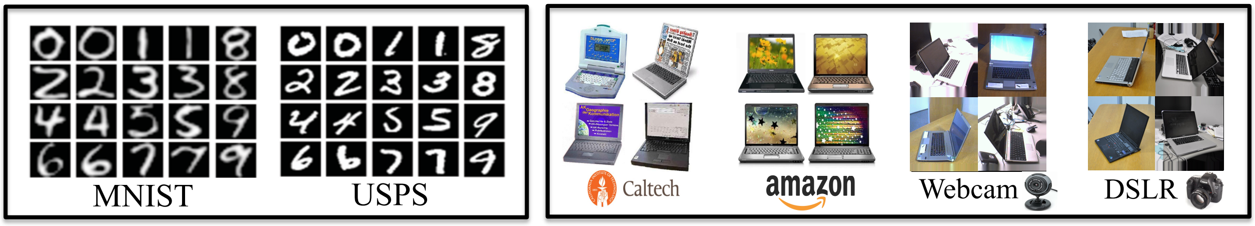


Fig. 2. Example images of (a) *MNIST* + *USPS* datasets and (b) *Caltech-256* + *Office* datasets.

Algorithm 1 Our Proposed Model

```
Input: Kernel matrix $\mathbf{K}$ of cross-domain data, labels $\mathbf{y}_S$ of source-domain data
  1. Initialize $\hat{D}_T^{(c)}$ as $\emptyset$
  while not converged do
    2. $\mathbf{W} \leftarrow$ Distribution adaptation $(\hat{D}_T^{(c)}, \hat{\mathbf{y}}_T)$ in (4) and let $[\mathbf{Z}_S, \mathbf{Z}_T] = \mathbf{W}^\top \mathbf{K}$
    3. Assign $\mathbf{Y}^{(0)}$ by classifiers trained by $\mathbf{Z}_S$ and (5)
    4. Construct the k-NN graph matrix $\mathbf{E}$ and $\mathbf{S}$ within target domain $\mathbf{Z}_T$
    5. $(\hat{D}_T^{(c)}, \hat{\mathbf{y}}_\mathbf{T}) \leftarrow$ label propagation $(\mathbf{Y}^{(0)}, \mathbf{S})$ in (7)
  end while
  6. $\mathbf{y}_T \leftarrow \hat{\mathbf{y}}_T$
Output: $\mathbf{y}_T$ as labels of target-domain data
```

Once we observe the uncertain label matrix $\mathbf{Y}$ and the target-domain structural similarity matrix $\mathbf{E}$, we apply the technique of label propagation [36] to determine the labels for the target-domain data. More specifically, by constructing $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{E}\mathbf{D}^{-1/2}$ where $\mathbf{D}$ is a diagonal matrix that $d_{ii} = \sum_j E_{ij}$, we update the following equation for propagating the label information in the target domain:

$$\mathbf{Y}^{(t+1)} = \alpha \mathbf{S} \mathbf{Y}^{(t)} + (1-\alpha)\mathbf{Y}^{(0)}.$$

At each iteration, each target-domain instance observes the structural information from its neighbors via $\mathbf{S}$, while retaining the original label information (i.e., soft labels $\mathbf{Y}^{(0)}$ with regularization parameter $\alpha \in (0, 1]$). We note that, for $\mathbf{Y}^{(t)}$, it can be reformulated as

$$\mathbf{Y}^{(t)} = (\alpha \mathbf{S})^t \mathbf{Y}^{(0)} + (1-\alpha) \sum_{i=0}^{t-1} (\alpha \mathbf{S})^i \mathbf{Y}^{(0)}.$$

Since $0 < \alpha \leq 1$ and the eigenvalues of $\mathbf{S}$ are in $[-1, 1]$ (see detailed derivations in [36]), we have

$$\lim_{t\to\infty} (\alpha \mathbf{S})^t = 0, \text{ and } \lim_{t\to\infty} \sum_{i=0}^{t-1} (\alpha \mathbf{S})^i = (\mathbf{I} - \alpha \mathbf{S})^{-1}.$$

Therefore, the optimal label matrix $\mathbf{Y}^*$ can be derived as:

$$\mathbf{Y}^* = (\mathbf{I} - \alpha)(\mathbf{I} - \alpha \mathbf{S})^{-1} \mathbf{Y}^{(0)}. \tag{7}$$

With $\mathbf{Y}^*$ obtained, the final label of each target-domain instance is determined by $\hat{y}_i^t = arg \max_{j \leqslant C} Y_{ij}^*$ based on the winner-take-all strategy.

D. Adaptation via Iterative Optimization

Finally, we integrate the techniques and learning models presented in Sections III-B and III-C for performing joint unsupervised domain adaptation and cross-domain recognition. The proposed method is summarized in Algorithm 1. It can be seen that, except for the initialization stage which only adapts marginal distributions of cross-domain data, the optimization process take both marginal and conditional distributions into consideration, while the class labels are updated from the previous adaptation iterations. Later in the experiments, we will show that the proposed method converges to the optimal solution in terms of both MMD and accuracy in few iterations, which verify the effectiveness of the proposed model for domain adaptation and recognition.

IV. Experiments

A. Datasets and Settings

MNIST and USPS: We first consider cross-domain digit recognition using MNIST [19] and USPS [18] datasets. MNIST contains a training set of 60,000 images and a test set of 10,000 images, while each image is of size 28×28 pixels. On the other hand, each image in USPS is of size 16×16 pixels, and a total of 7291 and 2007 images are available for training and testing, respectively. Figure 2(a) shows example images of these two datasets.

Following the setting of [20], we randomly sample 2000 and 1800 images from MNIST and USPS (scaled to the same 16×16 pixels), respectively, and take pixel intensities as the features. Two cross-domain pairs are considered: MNIST → USPS and USPS → MNIST. Take MNIST → USPS for example, we have MNIST as the source domain with 2000 labeled training data, and USPS as the target domain with 1800 instances to be recognized. Similar remarks can be applied to USPS → MNIST.

Caltech-256 and Office: For experiments on cross-domain object recognition, we consider the Caltech-256 [14] and Office [25], [11] datasets. The former consists of object images of 256 categories (with at least 80 instances per category), while the latter contains 31 objects categories collected from three different sub-datasets: Amazon, DSLR, and webcam. Following the same settings applied in [20], [9], [10], we select the 10 over-lapping object categories of Caltech-256 and Office for experiments, and produce four different domains of interest: Caltech (C), Amazon (A), DSLR (D), and webcam

TABLE I

COMPARISONS OF RECOGNITION RATES (%) ON *Caltech256* + *Office* DATASETS.

S → T	C → A	D → A	W → A	A → C	D → C	W → C	A → D	C → D	W → D	A → W	C → W	D → W	*Average*
Direct	91.86	72.13	74.63	82.64	60.20	64.56	81.53	86.62	99.36	74.58	79.66	96.61	80.36
LP [36]	90.92	51.88	60.54	77.56	46.22	50.93	70.06	82.80	98.73	67.80	78.31	84.07	71.65
DASVM [3]	91.75	80.79	79.23	85.84	55.21	78.27	83.44	91.08	100	78.31	72.54	98.31	82.90
TCA [22]	90.21	87.68	82.67	85.04	79.70	77.38	82.16	87.26	98.22	76.94	81.02	97.02	85.44
GFK [11]	87.65	84.96	84.25	79.07	80.41	73.29	79.43	83.06	99.30	76.68	75.08	79.7	81.91
SA [8]	88.95	76.82	74.43	81.01	70.85	70.08	80.70	81.78	99.11	72.61	76.34	70.85	78.63
JDA [20]	92.02	90.28	87.02	86.33	83.88	83.64	88.54	90.36	100	83.78	85.08	97.98	88.91
LM [9]	92.28	-	86.01	84.42	-	70.53	84.71	89.17	99.36	84.07	85.42	-	-
TJM [21]	92.17	88.73	83.51	85.04	81.03	80.14	83.44	87.26	99.36	80.34	81.36	97.02	86.62
Ours*	92.80	92.17	92.07	87.44	86.02	86.02	91.08	93.63	100	87.12	89.67	98.63	91.16
Ours	**94.26**	**92.37**	**93.31**	**87.88**	**86.19**	**87.97**	**94.9**	**95.26**	**100**	**88.81**	**91.18**	**99.32**	**92.62**

TABLE II

COMPARISONS OF RECOGNITION RATES (%) FOR CROSS-DOMAIN DIGIT RECOGNITION. NOTE THAT S AND T DENOTE SOURCE AND TARGET DOMAINS, RESPECTIVELY.

S → T	MNIST → USPS	USPS → MNIST
Direct	50.1	33.2
TCA [22]	52.7	45.7
JDA [20]	68.5	56.0
TJM [21]	63.5	52.7
Ours*	70.6	62.7
Ours	**72.3**	**65.5**

(W). As a result, a total of 12 different cross-domain pairs will be available (e.g., C → A, C → W, etc.).

To describe each object image in the Caltech-256 and Office datasets, we apply the $DeCAF_6$ features [6]. As shown in [6], these features are able to achieve very promising results for image classification. With the use of $DeCAF_6$ features, each image will be converted into a 4096-dimensional representation for training and testing.

It is worth noting that, since only unlabeled (test) data are available in the target domain, one cannot apply cross-validation to select the parameters for the learning models. For fair comparisons, we follow the same parameter settings as [20] did, and set $\lambda = 0.1$ and 1 for digit and object datasets, respectively. When performing data embedding, we choose $k = 100$ as the reduced feature dimension. In addition, we follow the recent works of [20], [22] and apply the linear kernels for constructing the kernel matrix $\mathbf{K}$. For simplicity, we fix the parameter $\alpha = 0.5$ for label propagation, and set the number of neighbors (for the graph-based similarity matrix $\mathbf{E}$) as 15 for all our experiments.

B. Evaluation

For cross-domain digit recognition, we consider the approaches of TCA [22], JDA [20] and TJM [21]. It is worth repeating that, JDA also adapts both marginal and conditional distributions for unsupervised domain adaptation as we do. For baseline approaches, we consider the direct use of SVMs trained by source-domain data in the original feature space (i.e., no domain adaptation). Table II lists the recognition results of cross-domain digit recognition.

Recall that, when using our proposed method, recognition is achieved when the domain adaptation process is complete (i.e., via label propagation). To show that we can also train the SVM classifiers in the derived transformed feature space using projected labeled source-domain data, and apply such classifiers to recognize the projected target-domain data as other recent methods do (e.g., TCA and JDA), we provide additional results of ours in Table II (denoted as Ours*). Nevertheless, as shown in this table, our methods clearly outperformed baseline and state-of-the-art methods for the task of cross-domain digit recognition.

For cross-domain object recognition, we consider another state-of-the-art method of Landmarks (LM) [9], [10]. LM can only be applied to 9 out of 12 cross-domain pairs. As explained in [10], LM requires a sufficient amount of source-domain data for adaptation, and it cannot be applied to the cases when DSLR is applied as the source domain. DASVM [3] is also compared with our proposed method. It first derives the SVM models using labeled source-domain data, and this model will be updated using the observed target-domain data and their pseudo labels for both adaptation and classification purposes. Other state-of-the-art methods like Geodesic Flow Kernel (GFK) [11] and Subspace Alignment (SA) [8] are also compared with our proposed method. Note that LP [36] indicates the results using only label propagation.

We compare the recognition performance of different methods in Table I. Our proposed method does not have this limitation. More importantly, from the results presented in this table, we see that our method significantly outperformed all others in all cases.

C. Evaluation in Imbalanced Cross-domain Data

For imbalanced cross-domain data, we may assume either a single domain comprises of multi sub-domains, or imbalanced label numbers across domains.

We first present our experiments of imbalanced class numbers across domains. It is worth noting that, for unsupervised domain adaptation, label information can only be obtained from the data in the source domain. Thus, we do not consider the case of unseen object categories (and data) in the target

TABLE III
PERFORMANCE OF IMBALANCED UDA ON OFFICE+CALTECH WITH MULTIPLE SUB-DOMAINS.

S → T	Direct	DASVM	TCA	GFK	SA	JDA	LM	TJM	Ours
A, C, D → W	81.69	81.36	86.10	79.78	78.10	92.88	90.51	92.88	**93.22**
A, C, W → D	96.18	94.90	97.45	84.85	89.04	97.45	94.27	**98.73**	**98.73**
C, D, W → A	82.88	85.91	92.28	84.91	88.47	92.69	93.01	91.65	**93.95**
A, D, W → C	78.01	78.36	84.42	79.69	81.11	88.25	84.42	83.53	**88.78**
D, W → A, C	70.49	65.83	81.69	77.06	73.80	87.84	81.55	74.10	**89.57**
C, W → A, D	86.19	84.93	93.36	84.82	88.74	93.18	91.84	92.83	**95.07**
C, D → A, W	83.64	86.75	92.58	83.59	87.80	92.34	90.66	91.70	**94.01**
A, W → C, D	81.02	78.67	86.25	78.48	82.73	88.98	85.31	84.69	**90.23**
A, D → C, W	80.89	82.51	85.40	78.40	83.07	89.00	85.33	84.91	**89.99**
A, C → D, W	64.82	66.15	83.63	77.73	78.16	86.73	87.39	86.73	**88.50**
average	80.58	80.54	88.32	80.93	83.10	90.93	88.43	88.17	**92.20**

TABLE IV
CROSS-DOMAIN CLASSIFICATION ON OFFICE+CALTECH WITH IMBALANCED CROSS-DOMAIN LABEL NUMBERS. NOTE THAT THE SOURCE-DOMAIN LABEL NUMBER IS FIXED AT 10, WHILE THE TARGET-DOMAIN LABEL NUMBER $C_T \leq 10$.

C_T	Direct	TCA	GFK	SA	JDA	TJM	Ours
3	77.38	86.71	74.45	63.45	82.29	80.53	**93.27**
4	77.34	87.22	76.52	67.09	84.42	82.29	**93.94**
5	77.26	87.29	78.34	69.96	86.74	82.73	**94.10**
6	75.47	86.32	78.71	71.07	86.00	81.77	**92.36**
7	76.10	86.53	80.99	73.31	88.35	83.09	**92.52**
8	75.41	86.63	80.55	75.69	89.52	83.56	**92.60**
9	75.36	86.18	81.31	78.24	89.94	84.32	**92.52**
10	80.36	85.44	81.91	78.63	88.91	86.62	**92.62**

TABLE V
RECOGNITION PERFORMANCE ON A → W WHEN CHOOSING DIFFERENT VALUES OF NEIGHBORS (K) AND σ FOR CONSTRUCTING OUR K-NN GRAPH. NOTE THAT THE RESULT WITH OUR DEFAULT PARAMETER CHOICE IS SHOWN IN RED.

k	$\log_2(\sigma)$						
	-3	-2	-1	0	1	2	3
5	85.08	85.42	84.75	84.75	84.75	84.75	84.41
10	86.78	91.19	90.85	89.49	89.49	89.49	89.49
15	90.51	90.85	88.81	87.80	87.46	87.80	87.80
20	90.51	90.85	87.46	86.10	86.10	86.10	86.10
25	90.51	90.85	86.44	85.42	85.42	85.42	85.42
30	90.51	89.15	85.76	84.41	84.41	84.41	84.41

domain. In other words, we always have the target-domain labels as a subset of those in the source domain.

For the dataset of Office+Caltech, we fix the source-domain label number as 10, while that in the target domain ranging from $C_T = 3$ to 10. Table IV lists and compares the results of different domain adaptation approaches. For the sake of simplicity, we only present the average results of 12 cross-domain pairs in Table IV. And, since DASVM cannot produce satisfactory results on such imbalanced settings, we do not include their results in this table. Nevertheless, from Table IV, it is clear that our proposed method consistently outperformed others over different C_T numbers. Thus, the use of our approach for imbalanced domain adaptation can be verified.

For unsupervised domain adaptation with multiple sub-domains, we consider 10 cross-domain data pairs from the dataset of Office+Caltech (see Table III) for evaluation. For each row in Table III, either domain consists of multiple sub-domain data, and complete results of different approaches are listed. From this table, we see that our approach performed favorably against state-of-the-art unsupervised domain adaptation methods. Therefore, the effectiveness of our approach for mix-domain UDA problems can be successfully verified.

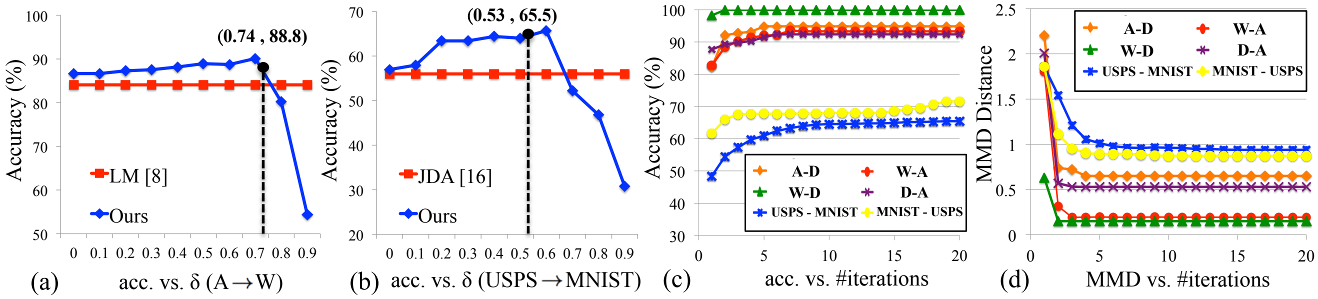
D. Parameter Sensitivity, Convergence and Computation Time

Recall that, δ in (5) determines the aggressiveness of label propagation by controlling the number of uncertain labels to be transferred from the source domain. In our experiments, we fixed δ as the lower quantile (i.e., 25%) of the maximum posterior probabilities observed from target-domain data, which allows 75% of target-domain data to be assigned uncertain labels during each iteration. Although a larger δ value would imply fewer (and less noisy) target-domain data with uncertain labels for propagation, the adaptation capability would be limited due to less information adapted from the source domain. Thus, the choice of δ is a trade-off between adaptation and propagation.

Figures 3(a) and (b) compare the recognition performance of two example domain pairs over different δ values. It can be seen that, while extreme δ (i.e., close to 0 or 1) cannot achieve satisfactory performance, our δ choice (based on the above guideline) was able to achieve improved results when comparing to state-of-the-art methods. Intuitively, the choice of δ should be domain dependent. For example, if the mismatch between source and target domains is marginal, one would expect that a small δ would be sufficient for performing adaptation. The study of domain biases and its effect on adaptatoin/propagation aggressiveness would be among our future research directions.

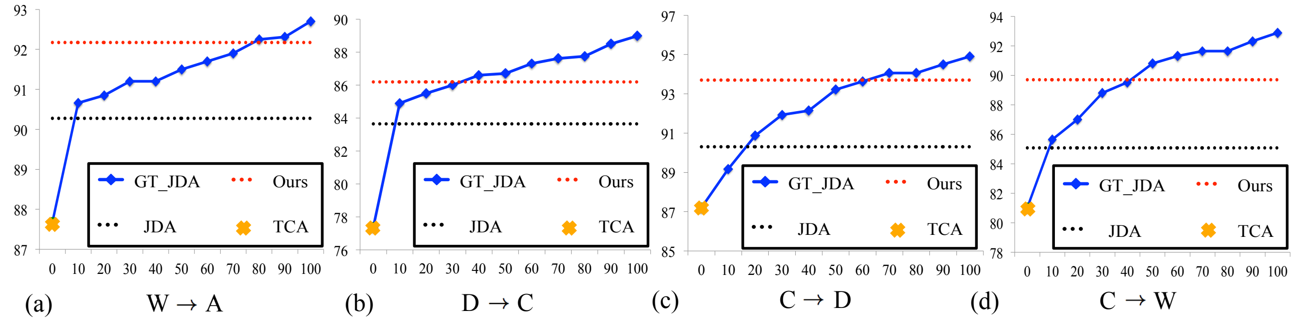
As noted in Section III-D, we iteratively solve the proposed model for adapting cross-domain data. To assess its convergence property, Figures 3(c) and (d) show the recognition accuracy and MMD distance with increasing iteration numbers, respectively. Recall that, the MMD distance is calculated by summing up the first two terms in (2) using target-domain data with ground truth labels. Due to space limit, only selected cross-domain data pairs are presented. From the above figures, we see that both accuracy and distance converged within 5-10 iterations during optimization.

In our experiments, we did not fine tune the parameters



(a) acc. vs. δ (A→W) (b) acc. vs. δ (USPS→MNIST) (c) acc. vs. #iterations (d) MMD vs. #iterations

Fig. 3. Parameter sensitivity and convergence analysis. For the former issue, we show the recognition rates over different δ values on (a) A→W and (b) USPS→MNIST. Note that the vertical dotted lines indicate the δ values determined by our approach (see Section III-C). For convergence analysis, we report (c) recognition accuracy and (d) MMD distance versus the number of iterations over several domain pairs.



(a) W → A (b) D → C (c) C → D (d) C → W

Fig. 4. Evaluation of adaptation capability on (a) W → A, (b) D → C, (c) C → D and (d) C → W. The horizontal axis indicates the percentage of ground truth labeled data utilized in the target domain, and the vertical axis denotes the accuracy. Note that only GT_JDA applies ground truth target-domain labels for matching cross-domain conditional distributions (neither TCA, JDA or ours does so).

TABLE VI

COMPARISONS OF RUNTIME ESTIMATES (IN SECONDS) OF DIFFERENT METHODS FOR A → W. NOTE THE METHODS OF JDA, TJM, AND OURS ALL CONVERGED WITHIN 10 ITERATIONS.

TCA [22]	JDA [20]	LM [10]	TJM [21]	Ours
4.15 (s)	34.12 (s)	1204 (s)	36.17 (s)	45.41 (s)

of "k" and "σ" when constructing the k-NN graph for label propagation. Since there is no labeled data in the target domain, one cannot perform cross-validation to select such parameters. We now provid additional results in Table V, in which our default parameter choice achieved satisfactory performance. As noted in our paper, we fixe k as 15 and σ as $\frac{1}{2}$ in all of our experiments.

Finally, we compare the computation time of different methods on the domain pair of A → W in Table VI. The runtime estimates were performed on an Intel Core i5 PC with 2.6 GHz CPU and 8G RAM. It can be seen that the computation time of our proposed approach (including iterative optimization and label propagation) was comparable to those of state-of-the-art methods, while the recognition performance was greatly improved.

E. Additional Remarks

1) Adaptation capability: As discussed in Section III-C, a major contribution of our work is the ability in exploiting label and structure consistency for unsupervised domain adaptation. This allows us to better match cross-domain conditional distributions for improved classification performance. For verification purposes, we consider an additional JDA-based approach (denoted as GT_JDA), which utilizes different amounts of target-domain instances with their ground-truth labels (not pseudo labels) to construct $\mathbf{L_c}$ in (2). With more labeled target-domain data observed for adaptation, the improvements of cross-domain recognition can be expected.

Figure 4 compares TCA, JDA, our method, and GT_JDA with varying amounts of labeled target-domain instances. From this figure, it can be seen that our method was able to achieve comparable results with GT_JDA using a large amount of ground truth target-domain labels (e.g., those using 30% of ground truth labeled data in the target domain or more). It is worth repeating that, compared to GT_JDA, our method did not consider any labeled data in the target domain during adaptation. Therefore, from the above experiments, the capability of our approach in associating cross-domain data for unsupervised domain adaptation can be successfully verified.

2) Robustness to initialization error: Similar to JDA, or other MMD-based domain adaptation approaches, we apply an iterative optimization process to associate cross-domain data (as shown in Algorithm 1). It would be a crucial issue if the prediction errors observed during initialization would affect the resulting adaptation and classification performance. To verify that our proposed method exhibits sufficient robustness to such initialization errors, we perform additional experiments with different degrees of perturbed errors when initializing our iteration process. To be more specific, instead of applying the

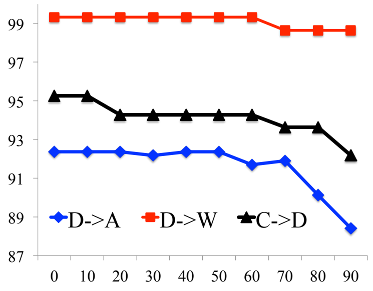


Fig. 5. Robustness analysis. The horizontal axis is the percentage of incorrect labels manually introduced into the first iteration of our adaptation process. The vertical axis denotes the accuracy.

observed uncertain matrix $\mathbf{Y}$, we deliberately and randomly introduce different amounts of prediction errors to the target-domain data in the first iteration of Algorithm 1.

Figure 5 presents and compares the results. It can be seen that, the introduced errors did not significantly affect the adaptation (and thus recognition) performance. For some dataset pairs (e.g., D → W), only negligible differences can be observed even if 70% of the target-domain data were intentionally assigned incorrect labels in the first iteration. Therefore, we not only verify the robustness of our approach to initialization errors. The above experiments also support the exploitation of both label and structure consistency for unsupervised domain adaptation.

3) Kernel choice: Although only universal kernels like Gaussian RBF are justified in the literature on MMD-based UDA [13], we note that our use of linear kernels is more than matching the data means across domains. This is because that, in addition to minimizing the difference between cross-domain data means, our method also matches the class-conditional means across data domains. Moreover, the constraint in (2) needs to be satisfied for preserving the covariance information of cross-domain data. In order words, subtracting the data means from each domain would not result in zero for our MMD calculation. Recent works of TCA [22], DIP [1], TJM [21] also confirmed show that the use of linear kernels in a MMD-based formulation would achieve satisfactory performance. And, our experiments also show that our proposed method performed favorably against these recent approaches.

Nevertheless, additional experiments using Gaussian RBF kernels are provided in our proposed formulation. Figure 6 compares the performance of the uses of linear v.s. RBF kernels. From this figure, we see that the use of RBF kernels only achieved comparable results as that of linear kernels over different σ^2_{rbf} choices. Therefore, the use of linear kernels for our method would still be preferable.

V. Conclusion

We proposed an unsupervised domain adaptation based on transfer feature learning. In addition to matching both marginal and conditional distributions of cross-domain data, our proposed model further leverages rich label and structural information across domains. This allows us to achieve improved adaptation and recognition of cross-domain data. Our

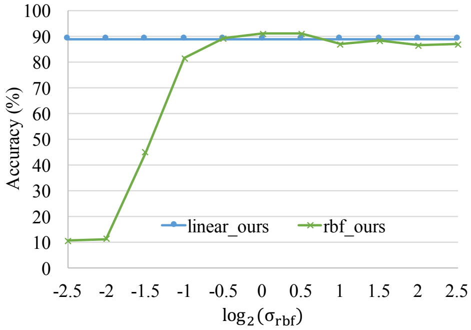


Fig. 6. Performance on A → W of our method with different kernel choices.

experiments on cross-domain digit and object recognition confirmed that our proposed model performed favorably against state-of-the-art domain adaptation methods. Future research directions include landmark (i.e., instance) selection for cross-domain data and domain-adaptive label propagation, which could further improve the domain adaptation and recognition performance.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology under Grant MOST103-2221-E-001-021-MY2.

References

[1] M. Baktashmotlagh et al. Unsupervised domain adaptation by domain invariant projection. In *IEEE ICCV*, 2013. 2, 9

[2] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *AISTATS*, 2010. 2

[3] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE T-PAMI*, 2010. 6

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 4

[5] H. Daumé III, A. Kumar, and A. Saha. Co-regularization based semi-supervised domain adaptation. In *NIPS*, 2010. 1

[6] J. Donahue et al. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014. 6

[7] S. Ebert et al. Extracting structures in image collections for object recognition. In *ECCV*. 2010. 2, 4

[8] B. Fernando et al. Unsupervised visual domain adaptation using subspace alignment. *IEEE ICCV*, 2013. 2, 6

[9] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013. 1, 2, 5, 6

[10] B. Gong, K. Grauman, and F. Sha. Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *IJCV*, 2014. 2, 5, 6, 8

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE CVPR*, 2012. 1, 2, 5, 6

[12] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE ICCV*, 2011. 2

[13] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two sample problem. In *NIPS*, 2007. 1, 2, 3, 9

[14] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 5

[15] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko. Efficient learning of domain-invariant image representations. *ICLR*, 2013. 2

[16] D.-A. Huang and Y.-C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *IEEE ICCV*, 2013. 2

[17] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007. 2

[18] J. J. Hull. A database for handwritten text recognition research. *PAMI*, 16(5):550–554, 1994. 5

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[20] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *IEEE ICCV*, 2013. 1, 2, 3, 4, 5, 6, 8

[21] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer joint matching for unsupervised domain adaptation. In *IEEE CVPR*, 2014. 2, 6, 8, 9

[22] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans Neural Networks*, 2011. 1, 2, 3, 4, 6, 8, 9

[23] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa. Domain adaptive dictionary learning. In *ECCV*. Springer, 2012. 2

[24] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013. 2, 4

[25] K. Saenko et al. Adapting visual category models to new domains. In *ECCV*. 2010. 1, 5

[26] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998. 3

[27] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *IEEE CVPR*, 2011. 2

[28] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE CVPR*, 2012. 2

[29] S. Shekhar et al. Generalized domain-adaptive dictionaries. In *IEEE CVPR*, 2013. 2

[30] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000. 2

[31] S. o. Si. Bregman divergence-based regularization for transfer subspace learning. *IEEE TKDE*, 2010. 1

[32] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE CVPR*, 2011. 1, 2

[33] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *JMLR*, 5(975-1005):4, 2004. 4

[34] Y. Yeh, C. Huang, and Y. Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. 2014. 2

[35] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure. Cross domain distribution adaptation via kernel mapping. In *ACM KDD*, 2009. 1

[36] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003. 2, 3, 4, 5, 6

Cheng-An Hou received the B.S. degree from the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan, in 2013. From 2013 to 2015, he was a Research Assistant with the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan. He is currently a Master Student with Robotics Institute, Carnegie Mellon University, Pittsburgh. His research interests include machine learning, computer vision, and pattern recognition.

Yao-Hung Hubert Tsai received the B.S. degree from the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, in 2014. He is a Research Assistant at the Research Center for Information Technology Innovation (CITI) of Academia Sinica, Taiwan. His research interests include visual domain adaptation, machine learning, optimization, data mining, and pattern recognition.

Yi-Ren Yeh received his M.S. and Ph.D. degrees from the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology in 2006 and 2010, respectively. From August 2008 to May 2009, he was a visiting scholar of CyLab, Carnegie Mellon University, Pittsburgh, USA. He was a postdoctoral research fellow of the Research Center for Information Technology Innovation (CITI) at Academia Sinica and Intel-NTU Connected Context Computing Center at National Taiwan University from 2010 to 2013. He was an assistant professor of Department of Applied Mathematics at Chinese Culture University from 2013 to 2015. He joined Department of Mathematics at National Kaohsiung Normal University as an assistant professor in 2015. His research interests include machine learning, data mining, multimedia content analysis, and data analysis in IoT(Internet of Things).

Yu-Chiang Frank Wang received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2001. From 2001 to 2002, he was a Research Assistant with the National Health Research Institutes, Taiwan. He received the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2004 and 2009, respectively.

Dr. Wang joined the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan, in 2009, where he currently holds the position as the Deputy Director and the Associate Research Fellow. He leads the Multimedia and Machine Learning Laboratory at CITI, and his research interests span the fields of computer vision, pattern recognition, machine learning, and image processing. He serves as an Organizing or Program Committee Member at multiple international conferences, and several of his papers were nominated for the Best Paper Awards. In 2013, he is selected as the Outstanding Young Researcher by the National Science Council of Taiwan.