Transfer Learning for Image Classification with Incomplete Multiple Sources

Zhengming Ding[†], Ming Shao[†], and Yun Fu[†‡],
[†]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115
[‡]College of Computer and Information Science, Northeastern University, Boston, MA, 02115
{allanding,mingshao,yunfu}@ece.neu.edu

***Abstract*—Transfer learning plays a powerful role in mitigating the discrepancy between test data (target) and auxiliary data (source). There is often the case that multiple sources are available in transfer learning. However, naively combining multiple sources does not lead to valid results, since they will introduce negative transfer as well. Furthermore, each single source from multiple sources may not cover all the labels of the target data. In this paper, we consider the problem that how to better utilize multiple incomplete sources for effective knowledge transfer. To this end, we propose a Bi-directional Low-Rank Transfer learning framework (BLRT). First, we adapt the conventional low-rank transfer learning to multiple sources knowledge transfer scenario. Second, an iterative structure learning is proposed to better use prior knowledge for transfer learning coefficients matrix. Third, a cross-source regularizer is added to couple the same labels from multiple incomplete sources, so that they could jointly compensate missing data from other sources. Experimental results on three groups of databases including face and object images have demonstrated that our method can successfully inherit knowledge from incomplete multiple sources and adapt to the target data successfully.**

I. Introduction

Transfer learning has already attracted considerable interest in the field of computer vision and pattern recognition, as it can address learning problems with insufficient labeled/training data. In brief, transfer learning borrows well-learned knowledge from source domain to facilitate learning problem in the target domain. Conventional transfer learning algorithms consider modifying either representation of the data or adapting classifiers, or both of them [1] to alleviate the divergences across two domains, i.e., source and target. However, in reality we always confront the situation that more than one source data are accessible, but none of them cover all labels information of the target data. In this paper, we formally define it as *Transfer Learning with Incomplete Multiple Sources* (**TL-IMS**) problem (Fig. 1).

Recent research activities on multi-source transfer learning concentrates on seeking for a better knowledge representation from multiple sources rather than simply merging all knowledge from different sources together. One way is to reweight different sources to align them well so that the rich yet complex knowledge can be successfully transferred. For example, Tan et al. proposed a Multi-Transfer framework for transfer learning with multiple views and multiple sources by extending co-training [2]. Another promising way is to use multi-task framework to correlate different sources to jointly organize knowledge transfer. For example, He et al. proposed a graph-based multi-task multi-view learning method for classification task [3].

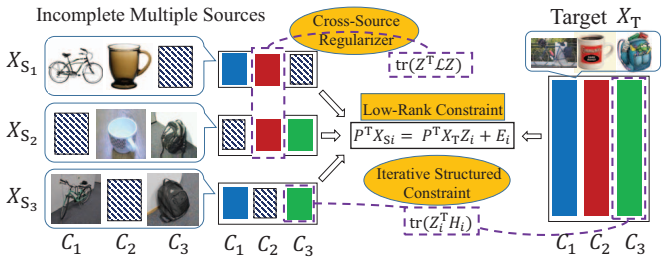


Fig. 1. Framework of the proposed algorithm. None of the source domains ($X_{\mathrm{S}} = \{X_{\mathrm{S}_1}, X_{\mathrm{S}_2}, X_{\mathrm{S}_3}\}$) can cover all the labels in the target domain (X_{T}). Take the visual domain object database for example. Each color means one category. Each source is reconstructed by the target in an appropriate common subspace P using *low-rank constraint* $P^{\mathrm{T}}X_{\mathrm{S}_i} = P^{\mathrm{T}}X_{\mathrm{T}}Z_i + E_i$. We aim to use one class in X_{T} to reconstruct the same class in X_{S_i} by adding the *iterative structured constraint* $\mathrm{tr}(Z_i^{\mathrm{T}}H_i)$, where H_i is the structure constraint for i-th source. For example, the green part in each source is reconstructed by the green part in the target. Moreover, a *cross-source regularizer* $\mathrm{tr}(Z^{\mathrm{T}}\mathcal{L}Z)$ is added to align the same category in multiple sources, where $\mathcal{L}$ is the cross-source constraint matrix. For example, the red parts in different sources should share their structure information.

Most recently, low-rank constraint on matrix [4] has been popular in transfer learning [5], [6], [7] to facilitate the knowledge transfer. Along this line, Jhuo et al. presented a multi-source domain adaption by imposing low-rank constraints on data reconstruction from target to source, and all the transfer learning coefficient matrices from different sources [7]. Compared with [7], low-rank transfer subspace learning algorithms [8], [9], [5] are proposed to align source and target data in a learned common subspace or multiple subspaces, where the curse of dimensionality could be well handled. The very benefit of low-rank constraint in transfer learning is its local-aware reconstruction by uncovering the global structure within the data. That is only appropriate knowledge is transferred from one local space in the source/target to another local space in the target/source.

In this paper, we propose a novel Bi-directional Low-Rank Transfer learning framework (BLRT) with iterative discriminative structures and cross-source alignment towards Transfer Learning with Incomplete Multiple Sources (**TL-IMS**) (See Fig. 1). To the best of our knowledge, this is the first time

978-1-5090-0620-5/16/$31.00

to consider **TL-IMS** under cross-source regularization, and iterative structure learning framework. The core idea of our algorithm is to learn an appropriate domain-invariant subspace where relevant knowledge from different sources are coupled and reinforced to compensate any missing data with the same labels. In addition, the prior structure information is embedded in the transfer learning coefficient matrix to guide the low-rank reconstruction process. The main contributions of the paper are summarized as follows:

- An iterative structure learning approach is integrated to the conventional low-rank transfer subspace learning with incomplete multiple sources, where source data are well reconstructed by the target data in an appropriate common subspace in a supervised fashion. The prior structural information as well as low-rank constraint guarantees data from each single source is accurately aligned with the target data.
- A cross-source regularizer is integrated into low-rank transfer learning framework, so that data from different sources with the same labels are tightly coupled. As a result, the knowledge from different sources are jointly transferred to the target domains. In addition, the missing data from certain sources can be effectively compensated through relevant data from other sources after coupled by the proposed regularizer.

The rest sections of this paper are organized as follows. In Section II, we present a brief discussion of the related works and highlight the difference between theirs and ours. Then we propose our novel incomplete multi-source transfer learning for image classification in Section III. Experimental evaluations on three different datasets are reported in Section IV, which is followed by the conclusion in Section V.

II. Related Work

Transfer learning has been extensively discussed recently due to its prevalence in computer vision, machine learning, and data mining community [1]. It can be categorized by domains and tasks. Among them, transductive transfer learning, especially domain adaptation [7] uses different data domains for the same task. For example, we would like to conduct object recognition on both Amazon website images, and photos captured by digital cameras. Therefore, we need to adapt the either source or target data, or both of them to have similar feature distributions. Research works along this line can be found in [7], [10], [11].

Recently, subspace learning has been explored in transfer learning [12], [10], [13], [11] to mitigate the divergence between source and target domains. The reason is when data is in high-dimensional space, it usually fails to learn discriminant models due to the *curse of dimensionality* [14] as well as expensive computational cost. After projecting data to low-dimensional space, we can adapt both source and target data, to pass on the well-established subspace to the target data. In this paper, we also adopt transfer subspace learning for **TL-IMS**, but incorporate iterative structure learning and a cross-source regularizer as well.

Low-rank constraint is another factor that can guide knowledge through data reconstruction [7], [8], [6], [9]. It was developed from low-rank representation [4] where reconstruction coefficients matrix is used for revealing data structure, especially when the data are lying in multiple subspaces. Among them, Latent Low-rank Transfer Subspace Learning (L^2TSL) [9], [6] considers both subspace learning and low-rank constraint in missing modality transfer learning. Specifically, L^2TSL aims to seek a common subspace in which the target data and source data are well aligned under the low-rank constraint. Differently, our method can tackle incomplete multiple sources transfer learning through additional prior structure information and cross-sources regularizer.

Multi-task multi-view learning is another potential framework for multi-source learning [3], where multiple sources can be coupled under multi-task framework. Along this line, Jhuo et al. proposed a robust domain adaptation framework through low-rank constraint [7], which can solve the multiple sources transfer learning under multi-task framework. Different from these methods, our method aims to find an appropriate domain-invariant subspace, where prior structure information and cross-source regularizer are adopted for multiple sources transfer learning. In this way, we reduce the computation burden by an efficient regularizer, compared to [7]. In addition, explicit structure modeling through iterative updating assists in learning a better transfer learning coefficients for multiple sources.

III. Bi-directional Low-Rank Transfer Learning for Incomplete Multiple Sources

In Transfer Learning with Incomplete Multiple Sources (**TL-IMS**), suppose target data $X_{\mathrm{T}} \in \mathbb{R}^{d \times n_t}$ have l classes, where d is the original dimension of the data and n_t is the sample size of target; K different sources $X_{\mathrm{S}} = [X_{\mathrm{S}_1}, \cdots, X_{\mathrm{S}_K}]$ also has l classes, but none of the single source can cover all the classes ($X_{\mathrm{S}_i} \in \mathbb{R}^{d \times n_{si}}$ and $n_s = \sum_{i=1}^{K} n_{si}$). As illustrated in Fig. 1, each source data misses certain labels. Under this situation, transfer learning should be able to handle two problems: (1) how to transfer supervised knowledge from multiple sources to the target data; (2) how to couple multiple sources to compensate the missing data and avoid negative transfer. In the following parts, we propose two complement models: "supervised knowledge transfer with iterative structure learning", and "regularized low-rank transfer learning" to address these problems, respectively.

A. Supervised Knowledge Transfer with Iterative Structure Learning

Suppose each source X_{S_i} covers partial classes of the target X_{T}, and $X_{\mathrm{S}_i} \subsetneq \mathrm{span}(X_{\mathrm{T}})$. Inspired by recent work in transfer subspace learning [8], we aim to find an appropriate common subspace $P \in \mathbb{R}^{d \times p}$ (p is the reduced dimension) shared by sources and target domains, so that $P^{\mathrm{T}} X_{\mathrm{S}_i} \subseteq \mathrm{span}(P^{\mathrm{T}} X_{\mathrm{T}})$. Therefore, we can use reconstruction with extra constraint to guide knowledge transfer from each source to the target. In [8], [9], a low-rank constraint is enforced to guarantee accurate

alignment between sources and target data. Then, a discriminative subspace learned by sources data can be employed by the target data for feature extraction, or dimensionality reduction.

Specifically, in our problem, each source data can be reconstructed by the target in a shared subspace through a low-rank constraint. Then we can formulate a naive multiple source transfer learning framework in a multi-task scheme:

$$\begin{aligned} &\min_{Z_i} \textstyle\sum_{i=1}^{K} \|Z_i\|_*, \\ &\text{s.t. } P^{\mathrm{T}} X_{\mathrm{S}_i} = P^{\mathrm{T}} X_{\mathrm{T}} Z_i, \quad i = 1, \cdots, K, \end{aligned} \tag{1}$$

where $\|\cdot\|_*$ is matrix nuclear norm, $Z_i \in \mathbb{R}^{n_t \times n_{si}}$ is the low-rank coefficients and i indexes the i-th source. As low-rank constraint uncovers the structure information when transferring knowledge from the sources to the target, the reconstruction process will guide the knowledge transfer in a non-trivial way.

So far, the transfer learning framework in multi-task scheme only considers the marginal distributions of two domains. In reality, however, we have always access to label information of either source or partial target data, or both of them. In this way, it is essential to pre-load these label knowledge into the multi-task transfer learning model where source data with certain classes/clusters are only reconstructed by target data with the corresponding classes/clusters. Previous works [15], [16] have discussed such similar thought for learning low-rank codings which are guided with such structured low-rank regularizer. In multi-source transfer learning problem, for each source X_{S_i}, we formulate a structured constraint H_i with label information:

$$\begin{aligned} &\min_{Z_i} \sum_{i=1}^{K} \big(\|Z_i\|_* - \alpha \mathrm{tr}(Z_i^{\mathrm{T}} H_i)\big), \\ &\text{s.t. } P^{\mathrm{T}} X_{\mathrm{S}_i} = P^{\mathrm{T}} X_{\mathrm{T}} Z_i, i = 1, \cdots, K, \end{aligned} \tag{2}$$

where α is the balance parameter and $\mathrm{tr}(\cdot)$ is the trace operator of matrix. $\mathrm{tr}(Z_i^{\mathrm{T}} H_i)$ measures the similarity of Z_i and H_i, which tends to be maximized so that Z_i would be close to H_i. However, since we only have limited labels in the target domain, a predefined H_i may be inaccurate under the transfer learning scenario. Unlike [15] that uses fixed well-defined structured coefficients matrix to model relations between inputs and a dictionary, we propose to iteratively update our $H_i \in \mathbb{R}^{n_t \times n_{si}}$ after each round of our algorithm. The temporary classification results on the unlabeled samples in the target domain [17] are immediately adopted as label information to build the structure matrix for next round knowledge transfer. In general, H_i will converge towards ground truth labels iteration by iteration. This is what we called "Iterative Structure Learning". We will show this trend in the experimental section.

To compensate corrupted data or outlier of the source data, we include error terms in Eq. (2). Then the objective function of the multi-source transfer learning model can be rewritten as:

$$\begin{aligned} &\min_{P, Z_i, E_i, H_i} \textstyle\sum_{i=1}^{K} \big(\|Z_i\|_* + \lambda \|E_i\|_{2,1} - \alpha \mathrm{tr}(Z_i^{\mathrm{T}} H_i)\big), \\ &\text{s.t. } P^{\mathrm{T}} X_{\mathrm{S}_i} = P^{\mathrm{T}} X_{\mathrm{T}} Z_i + E_i, \\ &P^{\mathrm{T}} P = \mathrm{I}_p, \ i = 1, \cdots, K, \end{aligned} \tag{3}$$

in which E_i is sparse error term for ith source and λ is the trade-off between E_i and other terms. Specifically, $L_{2,1}$-norm is used to make $E_i \in \mathbb{R}^{p \times n_{si}}$ sample specific, aiming to find the outliers. The orthogonal constraint $P^{\mathrm{T}} P = \mathrm{I}_p$ is imposed to ensure the learned P is an orthogonal projection and $\mathrm{I}_p \in \mathbb{R}^{p \times p}$ is an identity matrix. For simplicity, we define $\Theta(Z_i, E_i, H_i) = \|Z_i\|_* + \lambda \|E_i\|_{2,1} - \alpha \mathrm{tr}(Z_i^{\mathrm{T}} H_i)$.

Low-rank transfer learning makes local-aware reconstruction, that is, the target sample is required to be reconstructed by the samples from one class in source domain. Therefore, the marginal distribution between each source and target is minimized. Furthermore, we introduce an iterative structured term to guide the low-rank transfer learning to mitigate the conditional distribution difference between each source and target domains. To sum up, model (3) is able to minimize the marginal and conditional distributions from cross-domain direction.

B. Cross-source Low-Rank Transfer Learning

Model (3) in the former section is able to deal with multiple sources independently, but does not consider multiple sources jointly. Consider the case in Fig. 1 where $X_{\mathrm{S}_1} = [\mathcal{C}_1, \mathcal{C}_2]$, $X_{\mathrm{S}_2} = [\mathcal{C}_2, \mathcal{C}_3]$ and $X_{\mathrm{S}_3} = [\mathcal{C}_1, \mathcal{C}_3]$ with low-rank reconstruction coefficients $\{Z_1, Z_2, Z_3\}$, each source data has two parts, corresponding to two different labels. Specifically, for $Z_1 = [Z_{11}, Z_{12}], Z_2 = [Z_{22}, Z_{23}], Z_3 = [Z_{31}, Z_{33}]$, the first number in the subscript of Z indexes the source and the second one indexes the class. Intuitively, data with the same label should use similar target data for reconstruction, although they come from different data sources. For example, Z_{12} and Z_{22} are both in category 2, but from different sources. This is another critical structure information that can assist in jointly learning Z_i over different sources. In addition, it compensates the missing data since the labeled data from other sources have already aligned the source and target well under this joint learning framework.

Let us further define $Z = [Z_1, Z_2, \cdots, Z_K]^{\mathrm{T}} \in \mathbb{R}^{n_s \times n_t}$, whose i-th row z_i correlates with the i-th sample x_i from multiple sources $X_{\mathrm{S}} = [X_{\mathrm{S}_1}, X_{\mathrm{S}_2}, \cdots, X_{\mathrm{S}_K}]$. If two source samples x_i and x_j have the same label, they would be reconstructed by similar data from the target domain, which means their low-rank coefficient vectors z_i and z_j should also be similar. Through this observation, we propose an effective graph regularization:

$$\min_{z_i, z_j} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} (z_i - z_j)^2 W_{ij}, \forall i, j, W_{ij} \in \mathcal{W}, \tag{4}$$

where W_{ij} is the weight between the new presentations z_i and z_j from all sources. Specifically, we exploit the binary strategy to define the weight matrix $\mathcal{W}$, in which $W_{ij} = 1$ when x_i and x_j share the same label (With this, z_i and z_j would be similar in the new low-rank representation); and $W_{ij} = 0$ if x_i and x_j are in different classes. Actually, we could adopt other weight definition schemes, e.g., heat kernel, cosine similarity.

Mathematically, we could further rewrite Eq. (4) as:

$$\begin{aligned}
&\min_{z_i,z_j} \sum_{i=1}^{n_s}\sum_{j=1}^{n_s}(z_i - z_j)^2 W_{ij} \\
= &\min_{z_i,z_j} \sum_{i=1}^{n_s} z_i^{\mathrm{T}} D_{ii} z_i - \sum_{i=1}^{n_s}\sum_{j=1}^{n_s} z_i^{\mathrm{T}} W_{ij} z_j \\
= &\min_{Z} \mathrm{tr}(Z^{\mathrm{T}}(\mathcal{D}-\mathcal{W})Z) \\
= &\min_{Z} \mathrm{tr}(Z^{\mathrm{T}}\mathcal{L}Z),
\end{aligned} \tag{5}$$

in which $\mathcal{L}$ is the so-called graph Laplacian. $\mathcal{D} \in \mathbb{R}^{n_s \times n_s}$ is a diagonal matrix with is diagonal element D_{ii} calculated as the rows sum of $\mathcal{W}$, that is, $D_{ii} = \sum_{j=1}^{n_s} W_{ij}$. Different from conventional graph embedding algorithms, the proposed weight matrix $\mathcal{W}$ carries discriminant information across different sources so that the regularizer would encourage the multiple sources data with the same labels to share the similar representations. Therefore, it could well align multiple sources to transfer more knowledge to facilitate the target learning.

To sum up, we integrate low-rank transfer learning and cross-source alignment into a unified framework by adding $\mathrm{tr}(Z^T\mathcal{L}Z)$ to Eq. (3) and we can achieve the final formulation as follows:

$$\begin{aligned}
&\min_{P,Z,Z_i,H_i,E_i} \sum_{i=1}^{K} \Theta(Z_i, E_i, H_i) + \frac{\beta}{2}\mathrm{tr}(Z^{\mathrm{T}}\mathcal{L}Z) \\
&\text{s.t. } P^{\mathrm{T}}X_{\mathrm{S}_i} = P^{\mathrm{T}}X_{\mathrm{T}}Z_i + E_i, P^{\mathrm{T}}P = \mathrm{I}_p, i = 1, \cdots, K,
\end{aligned} \tag{6}$$

where β is the balance parameter between the cross-source alignment term with other terms. With the objective function, we aim to mitigate the divergence across multi-source and target in two-directional transfer learning: one is cross-domain direction, which adopts a semi-supervised low-rank transfer learning in multi-task fashion to leverage each incomplete source and the target domain; the other is cross-source direction, which is designed to align multiple sources in a supervised way to compensate the missing knowledge in each source. Two directional transfer learning finally yields a discriminative domain-invariant subspace.

C. Solving Objective Function

To address the optimization problem (6), we first convert it to its equivalent optimization problem by introducing relaxing variables J_i and S_i as:

$$\begin{aligned}
\min_{E_i,S_i,J_i,P,Z} &\sum_{i=1}^{K} \left(\|J_i\|_* + \lambda\|E_i\|_{2,1} - \alpha\mathrm{tr}(S_i^{\mathrm{T}}H_i)\right) \\
&+ \frac{\beta}{2}\mathrm{tr}(Z^{\mathrm{T}}\mathcal{L}Z) \\
\text{s.t.} \quad & P^{\mathrm{T}}X_{\mathrm{S}_i} = P^{\mathrm{T}}X_{\mathrm{T}}S_i + E_i, Z_i = J_i, Z_i = S_i, \\
& P^{\mathrm{T}}P = \mathrm{I}_p, i = 1, \cdots, K.
\end{aligned} \tag{7}$$

Specifically, we deploy Augmented Lagrangian Multiplier (ALM) [18], [19], aiming to achieve better convergence. We further transform Eq. (7) to its augmented Lagrangian function as follows:

$$\begin{aligned}
&\sum_{i=1}^{K}\Big(\|J_i\|_* + \lambda\|E_i\|_{2,1} - \alpha\mathrm{tr}(S_i^{\mathrm{T}}H_i) + \mathrm{tr}(V_i^{\mathrm{T}}(Z_i - S_i)) \\
&+\mathrm{tr}(Q_i^{\mathrm{T}}(P^{\mathrm{T}}X_{\mathrm{S}_i} - P^{\mathrm{T}}X_{\mathrm{T}}S_i - E_i)) + \mathrm{tr}(Y_i^{\mathrm{T}}(Z_i - J_i)) + \\
&\frac{\mu}{2}(\|Z_i - J_i\|_{\mathrm{F}}^2 + \|P^{\mathrm{T}}X_{\mathrm{S}_i} - P^{\mathrm{T}}X_{\mathrm{T}}S_i - E_i\|_{\mathrm{F}}^2 \\
&+\|Z_i - S_i\|_{\mathrm{F}}^2)\Big) + \frac{\beta}{2}\mathrm{tr}(Z^{\mathrm{T}}\mathcal{L}Z),
\end{aligned}$$

where Y_i, Q_i, V_i are the three lagrange multipliers while μ is the penalty parameter. Since we have many variables to be optimized in Eq. (7), so that we cannot jointly update them. Fortunately, we could adopt leave-one-out scheme by updating those variables one by one in an iterative fashion [9]. We define the variables at time t as $J_{i,t}, S_{i,t}, E_{i,t}, P_t$ and Z_t. Then we could alternately update the variables J_i, S_i, E_i, P and Z in the $t+1$ iteration as follows:

Update J_i:

$$J_{i,t+1} = \arg\min_{J_i} \frac{1}{\mu_t}\|J_i\|_* + \frac{1}{2}\|J_i - \frac{Z_{i,t} + Y_{i,t}}{\mu_t}\|_{\mathrm{F}}^2. \tag{8}$$

Update S_i:

$$S_{i,t+1} = \mu_t(\mathrm{I}_d + X_{\mathrm{T}}^{\mathrm{T}}P_tP_t^{\mathrm{T}}X_{\mathrm{T}})^{-1}\bar{\mathcal{S}}_{i,t}, \tag{9}$$

where $\bar{\mathcal{S}}_{i,t} = \mu_t Z_{i,t} + \mu_t X_{\mathrm{T}}^{\mathrm{T}}P_t(P_t^{\mathrm{T}}X_{\mathrm{S}_i} - E_{i,t}) + \alpha H_{i,t} + V_{i,t} + X_{\mathrm{T}}^{\mathrm{T}}P_tQ_{i,t}$ and $\mathrm{I}_d \in \mathbb{R}^{d\times d}$.

Update E_i:

$$E_{i,t+1} = \arg\min_{E_i} \frac{\lambda}{\mu_t}\|E_i\|_{2,1} + \frac{1}{2}\|E_i - \widehat{E}_i\|_{\mathrm{F}}^2, \tag{10}$$

where $\widehat{E}_i = (P_t^{\mathrm{T}}X_{\mathrm{S}_i} - P_t^{\mathrm{T}}X_{\mathrm{T}}S_{i,t+1} + Q_{i,t}/\mu_t)$.

Update Z:

$$Z_{t+1} = F_t(\mathrm{I}_{n_s} + \frac{\beta}{\mu_t}\mathcal{L})^{-1}, \tag{11}$$

where $\mathrm{I}_{n_s} \in \mathbb{R}^{n_s\times n_s}$, $F_t = [F_{1,t}, \cdots, F_{K,t}]^{\mathrm{T}}$ and $F_{i,t} = \frac{1}{2}(S_{i,t+1} + J_{i,t+1} - (Y_{i,t} + V_{i,t})/\mu_t)$.

Update P:

$$P = \big(\sum_{i=1}^{K}\bar{X}_i\bar{X}_i^{\mathrm{T}}\big)^{-1}\big(\sum_{i=1}^{K}\bar{X}_i(E_i - \frac{Q_i}{\mu_t})^{\mathrm{T}}\big), \tag{12}$$

where $\bar{X}_i = X_{\mathrm{S}_i} - X_{\mathrm{T}}S_i$ and P can be initialized with traditional subspace learning methods, e.g., PCA [20], LDA [21], LPP [22]. Note that we adopt LDA if there is no special instructions.

Update H: We apply P_{t+1} to reduce the dimensionality of multiple sources and target data, then the nearest neighbor classifier is adopted to predict the labels of unlabeled target by using labeled sources. With pseudo labels of target data, we could build structured matrix following [15].

The details of the algorithm is outlined in **Algorithm 1**. Specially, Eq. (8) can be easily solved by Singular Value Thresholding (SVT) [23], while Eq. (10) can be effectively addressed with the shrinkage operator [24]. For the parameters, we set μ_0, ρ, ϵ and $\mu_{\max}$ empirically following previous works [25], [4], [6], while other three parameters λ, α and β are tuned during the experiments.

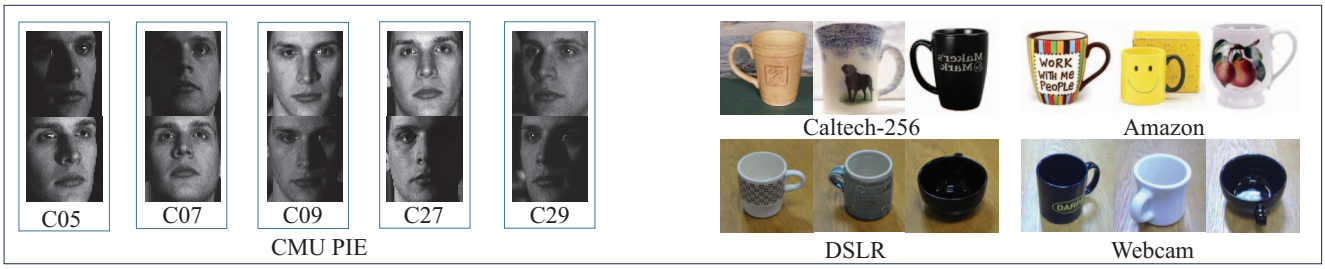


Fig. 2. Samples of three databases. The left one is CMU-PIE face database and the right one includes Caltech-256, Amazon, DSLR and Webcam datasets. Note **Office+Caltech** contains the four object datasets, while **Office** only includes the last three object datasets.

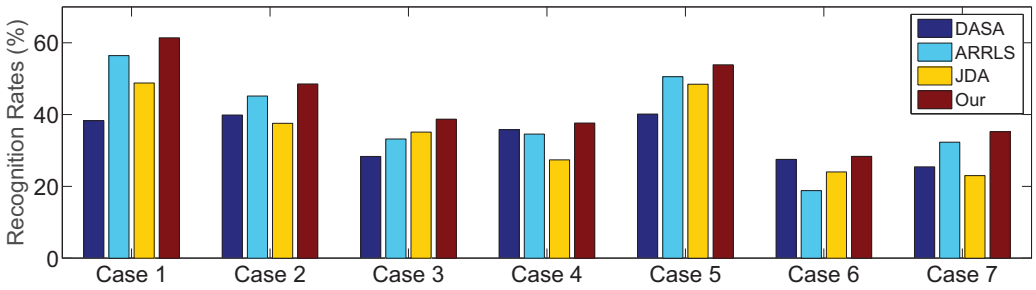


Fig. 3. Recognition results of 7 cases on CMU-PIE cross-pose face database, where Case 1: $\{C09, C05\} \rightarrow C07$, Case 2: $\{C07, C05\} \rightarrow C09$, Case 3: $\{C07, C05\} \rightarrow C09$, Case 4: $\{C09, C05, C29\} \rightarrow C07$, Case 5: $\{C09, C05, C29\} \rightarrow C27$, Case 6: $\{C09, C05, C27\} \rightarrow C29$ and Case 7: $\{C09, C05, C07\} \rightarrow C29$. (Best viewed in color)

Algorithm 1 Solving Problem (III-C)

Input: $X_{\mathrm{T}}, X_{\mathrm{S}_i}, \lambda, \alpha, \beta, i \in [1, K]$

Initialize: $S_{i,0} = J_{i,0} = 0, E_{i,0} = 0, Y_{i,0} = 0, V_{i,0} = 0, P_0,$
$Q_{i,0} = 0, \mu_0 = 10^{-6}, \mu_{\max} = 10^6, \rho = 1.1, \epsilon = 10^{-6}.$

while not converged **do**

1. Update $J_{i,t+1}$ via Eq. (8) by treating others as constant.
2. Update $S_{i,t+1}$ via Eq. (9) by treating others as constant.
3. Update $E_{i,t+1}$ via Eq. (10) by treating others as constant.
4. Update Z_{t+1} via Eq. (11) by treating others as constant.
5. Update P_{t+1} via Eq. (12) by treating others as constant, then $P_{t+1} \leftarrow$ **orthogonal**(P_{t+1}).
6. Optimize the lagrange multipliers $Y_{i,t+1}, Q_{i,t+1}, V_{i,t+1}$ via
$Y_{i,t+1} = Y_{i,t} + \mu_t(Z_{i,t+1} - J_{i,t+1});$
$V_{i,t+1} = V_{i,t} + \mu_t(Z_{i,t+1} - S_{i,t+1});$
$Q_{i,t+1} = Q_{i,t} + \mu_t(P_{t+1}^{\mathrm{T}} X_{\mathrm{S}_i} - P_{t+1}^{\mathrm{T}} X_{\mathrm{T}} S_{i,t+1} - E_{i,t+1}).$
7. Optimize the penalty parameter μ_{t+1} using
$\mu_{t+1} = \min(\rho\mu_t, \max_\mu).$
8. Check the convergence conditions
$\|P_{t+1}^{\mathrm{T}} X_{\mathrm{S}_i} - P_{t+1}^{\mathrm{T}} X_{\mathrm{T}} S_{i,t+1} - E_{i,t+1}\|_\infty < \epsilon,$
$\|Z_{i,t+1} - J_{i,t+1}\|_\infty < \epsilon, \|Z_{i,t+1} - S_{i,t+1}\|_\infty < \epsilon.$
9. $t = t + 1.$

end while

output: $S_i, E_i, J_i, Z, P.$

IV. Experiments

Experiments are conducted on CMU-PIE face database[1] and two visual object databases: **Office**[2]; and **Office+Caltech**[3] (Samples are shown in Fig. 2). We first experiment on three databases by comparing with state-of-the-art methods, and then analyze our method on several properties. Finally, we provide the experimental results on incomplete single source scenario. Note that the arrow "$\rightarrow$" is the direction from "sources" to "target". For instance, "{DSLR, Amazon} $\rightarrow$ Webcam" represents that DSLR and Amazon are two incomplete source domains and Webcam is the target one.

A. Cross-pose Face Database

The five near frontal face subsets of **CMU-PIE** face database $(C05, C07, C09, C27, C29)$ are applied and each subset contains 68 subjects under different illumination variations. We mainly compare with DASA [13], ARRLS [17] and TJM [26], and take the unsupervised setting. That is, we are only accessible to the labels of multiple sources [26]. To build the incomplete sources, we randomly remove 20 subjects when the size of sources is 2, and 30 subjects when the size of sources is 3. We do 20 random selections and average the results of 7 cases in Fig. 3.

B. Visual Object Database

Office object dataset is one popular visual domain adaptation benchmark which is widely used. Specifically, there are three real-world object subsets, i.e, Amazon, Webcam and DSLR. Amazon contains the images downloaded from online, which are usually background clean. Webcam and DSLR

[1]http://vasc.ri.cmu.edu/idb/html/face/
[2]http://www.cs.uml.edu/∼saenko/projects.html#data
[3]http://www-scf.usc.edu/∼boqinggo/domainadaptation.html

TABLE I
RECOGNITION RATE (%) OF 8 ALGORITHMS ON **OFFICE** DATABASE, WHERE D = DSLR, A = AMAZON AND W = WEBCAM.

	A, W $\rightarrow$ D	A, D $\rightarrow$ W	D, W $\rightarrow$ A
GFK [10]	31.32$\pm$0.05	39.65$\pm$0.03	19.50$\pm$0.02
LTSL [8]	34.02$\pm$0.02	37.68$\pm$0.02	18.86$\pm$0.14
RDALR [7]	32.81$\pm$0.18	36.85$\pm$0.15	20.19$\pm$0.03
TJM [26]	40.49$\pm$0.12	42.58$\pm$0.10	19.46$\pm$0.05
DASA [13]	37.30$\pm$0.04	42.45$\pm$0.04	16.41$\pm$0.05
ARRLS [17]	44.08$\pm$0.05	56.73$\pm$0.09	18.83$\pm$0.04
SDDL [11]	50.38$\pm$0.08	57.43$\pm$0.14	29.23$\pm$0.05
Ours	**50.59$\pm$0.06**	**60.26$\pm$0.07**	**42.35$\pm$0.03**

include the images captured from low-resolution and high-resolution devices, respectively. In total, Office includes more than 4 thousand images from 31 categories. **Office+Caltech** datasets contains previous Office and Caltech-256, which is also very popular for visual domain adaptation. There are 10 common categories for these four subsets. We follows the settings [10] and adopt SURF features, which are quantized into an 800-bin histogram with codebooks.

In the experiments, we compare our method with LTSL [8], RDALR [7], and GFK [10], TJM [26], DASA [13], ARRLS [17] and SDDL [11]. Assume we have access to a small number of labeled data in the target domain, then it follows semi-supervised domain adaptation setting. We strictly follow the settings in [27] for **Office** and [10] for **Office+Caltech**, respectively. To construct the incomplete environment, we randomly remove 6 categories out of 31 from each source for Office, while we randomly remove 2 categories out of 10 from each source for **Office+Caltech**. 20 rounds of random selections are conducted and the average recognition results are reported in Table I and Table II.

C. Discussion

From the results, we can observe that our method works better than other transfer learning methods. The reason is our method makes full use of the label information to mitigate both marginal and conditional divergences among across sources and across domains. Through two supervised constraints, our method can transfer the label information from source to target and also align incomplete multiple sources to compensate the missing data/labels in certain source.

LTSL and RDALR are both low-rank based algorithms. LTSL is single source method, so that directly combining multiple source together would introduce negative transfer due to the large divergence of multiple sources. RDLAR is multi-source method, but works in an unsupervised way by only considering the data distribution, aiming to find rotations on source domains, which may fail to uncover intrinsic information. Besides, RDALR does not explicitly align the multiple sources in a supervised way, but only adding a low-rank constraint on all the rotated sources. Compared with RDALR, LTSL achieves a slightly better result thanks to the label information and subspace representation. However, LTSL fails to transfer the exact label information to the target domain when labels are available in both domains. On the contrary, our method with iterative structure learning can transfer the label information to the exact target data even when some labels are missing in a single source domain.

GFK and DASA work in a similar way. GFK designs a kernel metric to minimize the divergence of source and target. DASA introduces subspace alignment to deal with the marginal distribution of different domains. TJM jointly matches the features across two domains and re-weights the instances across domains to build an domain-invariant subspace for effective dimensionality reduction. However, all the three cannot make use of the label information and align the multiple sources.

ARRLS also introduces pseudo labels of target domain to minimize the conditional distribution of source and target, together with the manifold regularizer to discover the geometric property of target. However, it is single-source based transfer learning method, which can only naively merge multiple sources. Besides, ARRLS employs the original Maximum Marginal Discrepancy [28] to minimize the marginal distribution, which cannot uncover the global structure of data.

SDDL is a multi-source domain adaptation algorithm, which introduces a common dictionary for the projected data of different domains. The supervised information is introduced to learn the dictionary. However, SDDL connects sources and target data in an undirect way via the common dictionary. Besides, SDDL adopts a sparse representation scheme, and therefore, it cannot uncover the class structure in sources and target domain. Differently, our method directly reconstructs target data through each incomplete source via an iterative structure learning term, and therefore it can transfer more supervised information into the target domain. Especially for incomplete sources, our cross-source regularizer and iteratively structured low-rank constraint would well align the sources and compensate the missing data in one source with the same classes from other sources and target data.

D. Convergence Property

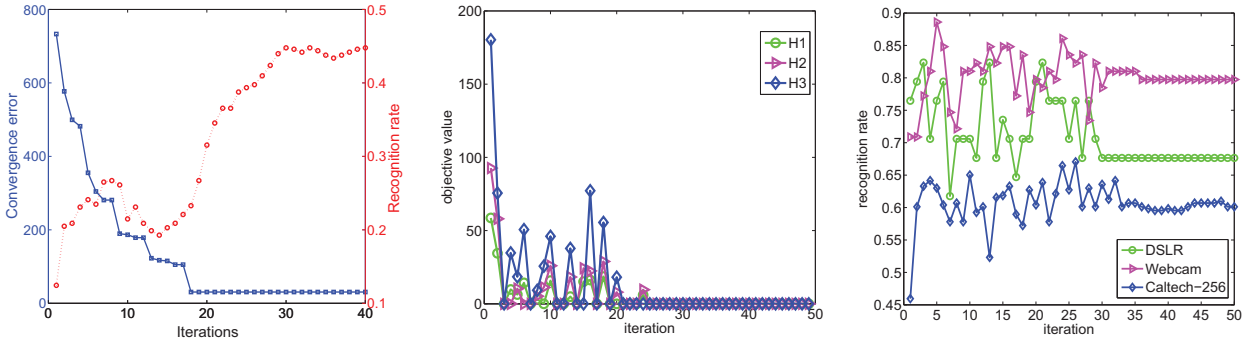
In this section, we evaluate the convergence properties in three parts: convergence of objective function, the feature discriminability of Z and convergence of iteratively structured term H_i.

We first show the convergence of our algorithm and the recognition performance in different iterations (Fig. 4(a)). From the results, we could observe that our algorithm can converge well and also achieve good results with more iterations' optimization (more than 30 iterations). To save computational time, we usually stop the optimization at 40-th iteration.

In addition, to better illustrate the iterative structure learning process H_i and cross-source regularizer, we show the self-similarity structure of Z in different iterations. Specifically, we only use 4 categories from 4DA database with configuration $\{C, D, W\} \rightarrow A$, and adopt LDA as the subspace method. We randomly remove 1 categories out of 4 to construct the incomplete condition. The objective value in Fig. 4(b) is

TABLE II
RECOGNITION RATE (%) OF 8 ALGORITHMS ON **OFFICE+CALTECH** DATABASE, WHERE A = AMAZON, D = DSLR, C = CALTECH-256 AND W = WEBCAM.

	A, C, W → D	A, C, D → W	C, D, W → A	A, D, W → C
GFK [10]	23.69±0.02	36.51±0.01	26.14±0.03	18.60±0.01
LTSL [8]	38.22±0.43	42.20±0.09	30.65±0.04	23.42±0.02
RDALR [7]	33.52±0.33	41.92±0.45	30.19±0.78	21.88±0.68
TJM [26]	50.32±0.27	52.27±0.10	37.91±0.04	32.97±0.07
DASA [13]	51.08±0.19	49.63±0.04	31.94±0.02	31.71±0.01
ARRLS [17]	54.90±0.54	61.22±5.08	50.89±0.04	44.92±1.17
SDDL [11]	65.56±0.34	71.12±5.09	51.39±0.14	34.27±1.27
Ours	**67.12±0.24**	**72.92±0.17**	**64.25±0.11**	**52.87±0.20**



(a) (b) $H_i, i = 1, 2, 3$ (c) Prediction results

Fig. 4. (a) Convergence (blue line) and recognition results (red line) with different iterations in setting "{W, A} → D". (b) Convergence of three H_i , where H_1 is for DSLR, H_2 is for Webcam and H_3 is for Caltech-256. (c) shows the prediction results of three sources.

(a) 1st iteration (b) 10th iteration (c) 25th iteration (d) ground truth

Fig. 5. Self-similarity structure of Z. (a-c) show the updated structure in 1st, 10th and 25th iteration, respectively, and (d) shows the ground truth, where each block represents one category. Note that gray pixel means the weight is 1 and the black one represents 0.

calculated as $\|H_{i,t+1} - H_{i,t}\|_{\mathrm{F}}$ for H_i. Also we evaluate the accuracy of the iteratively structures terms H_i (Fig. 4(c)), since we use pseudo labels of target to build H_i. From the results shown in Fig. 4(b), we can observe each H_i converges well. For each H_i, since the object value in the figure represents the variation of element-wise difference of H_i between two iterations, they finally converge to 0.

Further, we also see the self-similarity structure of Z is close to the ground truth with more iterations, as shown in Fig. 5. The self-similarity structure of Z relates to the discriminability of the reconstruction coefficients Z_i of each source, since we could treat Z_i as the new representation of the i-th source X_i. From the results, we can see the block structure in the 1-st iteration is not obvious. After a few iterations, the learned self-similarity structure converges towards the ground truth block structure, indicating the discriminability of the new features goes up. That is, the discrepancy across multiple sources is mitigated, so that the missing classes in each source could be compensated through other sources.

V. CONCLUSIONS

In this paper, we proposed a Bi-directional Low-Rank Transfer learning method (BLRT) for incomplete multiple sources transfer learning problem. Our BLRT introduced two

novel terms: an iterative structure term to better transfer prior knowledge from each source to the target domain; and a cross-source regularizer, to couple the highly correlated samples in multiple sources to avoid negative transfer and compensate missing data through coupled sources. Experiments on three databases showed our proposed algorithm can well handle the incomplete multiple sources problem to achieve a better performance in image classification.

ACKNOWLEDGEMENT

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, NPS award N00244-15-1-0041, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] B. Tan, E. Zhong, E. W. Xiang, and Q. Yang, "Multi-transfer: Transfer learning with multiple views and multiple sources," in *In Proceedings of the 13rd SIAM International Conference on Data Mining*. SIAM, 2013.

[3] J. He and R. Lawrence, "A graph-based framework for multi-task multi-view learning," in *ICML*, 2011, pp. 25–32.

[4] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[5] Z. Ding and Y. Fu, "Low-rank common subspace for multi-view learning," in *IEEE International Conference on Data Mining*, 2014, pp. 110–119.

[6] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4322–4334, Nov 2015.

[7] I.-H. Jhuo, D. Liu, D. Lee, S.-F. Chang *et al.*, "Robust visual domain adaptation with low-rank reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2168–2175.

[8] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *International Journal of Computer Vision*, pp. 1–20, 2014.

[9] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.

[11] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 361–368.

[12] S. Si, D. Tao, and B. Geng, "Bregman divergence -based regularization for transfer subspace learning," *TKDE*, vol. 22, no. 7, pp. 929–942, 2010.

[13] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars *et al.*, "Unsupervised visual domain adaptation using subspace alignment," in *IEEE International Conference on Computer Vision*, 2013.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[15] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 676–683.

[16] Z. Ding, M. Shao, and Y. Fu, "Deep low-rank coding for transfer learning," in *International Joint Conference on Artificial Intelligence*, 2015, pp. 3453–3459.

[17] M. Long, J. Wang, G. Ding, S. Pan, and P. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, 2013.

[18] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," *Optimization*, pp. 283–298, 1969.

[19] M. R. Hestenes, "Multiplier and gradient methods," *Journal of optimization theory and applications*, vol. 4, no. 5, pp. 303–320, 1969.

[20] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[22] X. He and P. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[23] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[24] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 569–592, 2009.

[25] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *ICML*, 2010, pp. 663–670.

[26] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

[27] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision*. Springer, 2010, pp. 213–226.

[28] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Neural information processing systems*, 2006, pp. 513–520.