Incomplete Multisource Transfer Learning

Zhengming Ding, *Student Member, IEEE*, Ming Shao, *Member, IEEE*, and Yun Fu, *Senior Member, IEEE*

***Abstract*—Transfer learning is generally exploited to adapt well-established source knowledge for learning tasks in weakly labeled or unlabeled target domain. Nowadays, it is common to see multiple sources available for knowledge transfer, each of which, however, may not include complete classes information of the target domain. Naively merging multiple sources together would lead to inferior results due to the large divergence among multiple sources. In this paper, we attempt to utilize incomplete multiple sources for effective knowledge transfer to facilitate the learning task in target domain. To this end, we propose an incomplete multisource transfer learning through two directional knowledge transfer, i.e., cross-domain transfer from each source to target, and cross-source transfer. In particular, in cross-domain direction, we deploy latent low-rank transfer learning guided by iterative structure learning to transfer knowledge from each single source to target domain. This practice reinforces to compensate for any missing data in each source by the complete target data. While in cross-source direction, unsupervised manifold regularizer and effective multisource alignment are explored to jointly compensate for missing data from one portion of source to another. In this way, both marginal and conditional distribution discrepancy in two directions would be mitigated. Experimental results on standard cross-domain benchmarks and synthetic data sets demonstrate the effectiveness of our proposed model in knowledge transfer from incomplete multiple sources.**

***Index Terms*—Cross domain/source, incomplete multisource, transfer learning.**

I. Introduction

TRANSFER learning [1], [2] has attracted considerable interests as it is able to well tackle learning tasks with limited or no labeled data in the training stage. In a word, transfer learning adapts well-established knowledge from source domain to boost the unlabeled target learning, where two domains have different distributions/feature spaces. In general, conventional transfer learning algorithms [3]–[10] tend to either extracting domain-invariant representation or adapting classifiers to mitigate the marginal distribution (MD) or conditional distribution (CD) difference across two domains. In reality, however, we could always face such challenges that multiple source data sets are accessible [11]–[15], but no one could cover all categories of the target data set. See example in Fig. 1; the target domain contains object images from Amazon Web site while source domains include object images taken by Web camera (low resolution), captured with digital camera (high resolution), and images from Caltech-256 object data set. Amazon Web site has three categories: “keyboard, mouse, and computer,” while Caltech-256, digital camera, and Web camera only covers “keyboard and mouse,” “mouse and computer,” and “keyboard and computer,” respectively. To the best of our knowledge, transfer learning with incomplete multiple sources (TL-IMS) is under insufficient exploration currently in machine learning and computer vision fields.

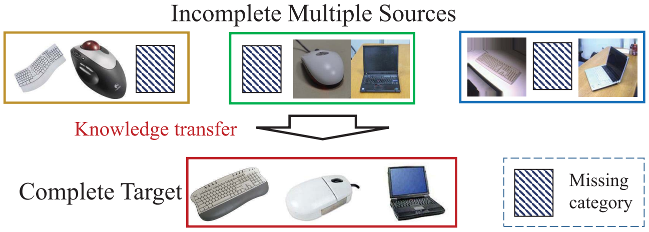


Fig. 1. Illustration of “transfer learning with incomplete multiple sources” problem (TL-IMS), where we have multiple sources (three sources here) but each single source has certain missing categories compared with the target domain.

When multiple sources are available, previous multisource transfer learning [5], [16]–[19] focuses on extracting domain-free representations from multiple sources rather than simply merging them together. In general, there are two strategies to deal with multisource transfer learning. One strategy is to reweight various sources in order to adapt the rich yet complex information among sources to boost the target learning [5], [16], [17]. Another successful strategy is to exploit multitask framework to joint multiple sources to guide the knowledge transfer [5], [18]. Nonetheless, all these methods assume complete multiple sources and may fail to transfer knowledge from incomplete multisource cases.

Recently, low-rank modeling [20] has been successfully applied to conventional transfer learning [5], [7], [21]. Existing low-rank transfer learning benefits the knowledge transfer with locality aware reconstructions, meaning only appropriate knowledge in local neighborhood is transferred from one

Manuscript received January 2, 2016; revised July 28, 2016; accepted October 7, 2016. Date of publication November 14, 2016; date of current version January 16, 2018. This work was supported in part by the NSF IIS Award under Award 1651902, in part by the NSF CNS Award under Award 1314484, in part by the Office of Naval Research (ONR) Award under Award N00014-12-1-1028, in part by the ONR Young Investigator Award under Award N00014-14-1-0484, and in part by the U.S. Army Research Office Young Investigator Award under Award W911NF-14-1-0218.

Z. Ding is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: allanding@ece.neu.edu).

M. Shao is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, Dartmouth, MA 02747 USA (e-mail: mshao@umassd.edu).

Y. Fu is with the Department of Electrical and Computer Engineering, College of Computer and Information Science, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2618765

2162-237X © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

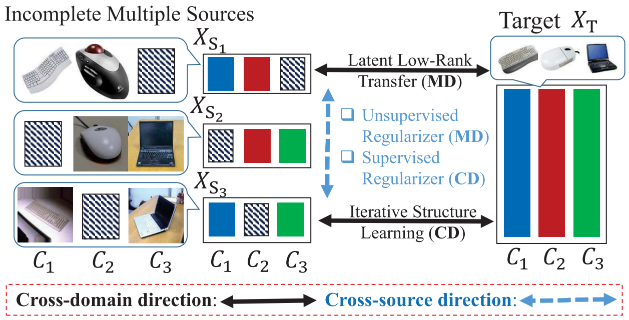

Incomplete Multiple Sources
X_{S_1}
X_{S_2}
X_{S_3}
Latent Low-Rank Transfer (MD)
Unsupervised Regularizer (MD)
Supervised Regularizer (CD)
Iterative Structure Learning (CD)
Target X_T
C_1 C_2 C_3
C_1 C_2 C_3
C_1 C_2 C_3
Cross-domain direction:
Cross-source direction:

Fig. 2. Framework of the proposed algorithm. Each color represents one class. None of single source domain $\{X_{S_1}, X_{S_2}, X_{S_3}\}$ can cover all the labels of target X_T. In cross-domain direction, *latent low-rank subspace transfer* with *iterative structure learning* is developed to couple each source and target domain, and therefore, the MD and CD discrepancy between sources and target domains would be both minimized. In cross-source direction, two novel regularizers are introduced to align multiple sources and compensate for the missing classes.

domain to another. For example, Shao *et al.* [7] developed a generalized low-rank transfer subspace learning algorithm, which explicitly imposed low-rank constraint on the data reconstruction from source to target in a latent space. This paper subtly links transfer learning and generalized subspace learning; however, it is originally designed for single source transfer, which ignores key factors of incomplete multisource scenario. For incomplete multisource transfer learning (IMTL), the missing categories in each source could be treated as the latent information, which should be well uncovered, so that we could transfer more knowledge to facilitate the target learning.

Latent structure discovery plays an important role in various real-world applications, which aims to approximate certain unobservable factors [22]–[26]. In general, researchers exploit greedy search, inference, or approximation algorithms to infer reasonable values for hidden factors, e.g., latent SVM [23], hidden Markov model [24], and missing modality transfer learning [21], [25], [26]. To recover missing knowledge during transfer model training, Ding *et al.* [21], [25] proposed a bidirectional latent low-rank transfer learning, which extended the conventional concept of latent factor. Therefore, it is reasonable to recover the missing categories in the source domain through latent transfer learning.

In this paper, we propose a novel IMTL algorithm through structured latent low-rank constraint and cross-source alignments, whose core idea is to seek an appropriate domain-free subspace where relevant knowledge for target from multiple sources is coupled and reinforced to compensate for any missing data in other sources. In particular, IMTL is designed to minimize the MD and CD discrepancy from two directions: cross-domain transfer and cross-source transfer (Fig. 2).

A. Our Contributions

As we mentioned before, current multisource transfer learning algorithms [5], [16]–[19] assume each single source could cover all the label information of the target data. In reality, we always confront the situations that none of the single source includes complete categories for target data. Thus, traditional multisource transfer learning [5], [16]–[19] may fail to transfer effective knowledge from multiple incomplete sources. To that end, IMTL is required to address the following issues: 1) how to adapt well-labeled knowledge from multiple sources to the target domain and 2) how to align multiple sources to make up for the missing categories. In this paper, we conduct a trial on the incomplete multisource cases to compensate any missing categories to facilitate the target learning tasks in an effective way. In particular, we aim to seek a domain-invariant subspace for multiple domains, where various sources and target data could be well aligned. This paper is partially based on our previous conference paper [27], which also tends to address the ITML but with different strategies. To sum up, the key contributions of this paper are in twofolds.

1) *Cross-Domain Transfer:* An evolutionary model is incorporated to describe the correlation between source and target based on training labels and, therefore, can guide the learning of low-rank reconstruction coefficient matrix in a common subspace in a supervised fashion. Moreover, the missing labels in each source can be implicitly recovered through a latent factor from the target data.
2) *Cross-Source Transfer:* Effective multisource alignment and manifold regularizer are integrated into latent low-rank transfer framework to reduce both MD and CD disparity. Therefore, the same class data from different sources are tightly coupled and jointly transferred to the target domain, even when some categories are missing from one individual source.

The rest of this paper is organized as follows. In Section II, we provide a brief review of related works. Then, we propose our IMTL in Section III with details on latent factor discovery, regularizers, and solutions. Experiments are provided in Section IV and the conclusions are drawn in Section V.

II. Related Work

Two lines of related work are discussed in this section: 1) low-rank transfer subspace learning and 2) multisource transfer learning, and highlight the difference between the proposed algorithm and them.

A. Low-Rank Transfer Subspace Learning

Transfer learning has been witnessed as an attractive technique in many real-world applications, which can be categorized into self-taught learning [28], inductive transfer learning, and transductive transfer learning according to the property of domains and tasks [1], [2]. This paper falls in the transductive transfer learning, since we adapt data from different domains for the same task [3]–[6], [17], [29]. However, none of these methods explicitly tackle the incomplete multisource problem.

Transfer subspace learning has demonstrated with promising results in transfer learning tasks by bridging distribution gaps between two domains in a common feature space [7], [17], [21]. The common feature space is usually obtained by conventional subspace learning methods to simultaneously address the problems of *curse of dimensionality* [30]

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

and *distribution gap*. In the learned feature space, it becomes easier to adapt two domains and pass on knowledge from one to another to boost performance. In this paper, we equip subspace learning with two directional knowledge transfer to mitigate both MD and CD disparity among different incomplete sources and target data, which, to the best of knowledge, has never been explored previously.

Furthermore, low-rank modeling [20] has been extensively exploited in transfer learning to mitigate the MD discrepancy between domains [7], [21], [25]. The low-rank constraint enforced on the reconstruction coefficients across two domains is able to reveal intrinsic data structure, which can guide the conventional transfer subspace learning. When data are limited in recovering the underlying structure, however, mining latent knowledge from insufficient observed data become necessary [22]. Different from existing methods along this line, we incorporate iterative structure learning to reveal the low-rank structure of the coefficient matrices of multiple sources in a semisupervised way. In addition, the latent factors and complete target data can jointly recover the missing labels of each single source. In this way, our method has better chance to transfer knowledge from incomplete multiple sources to the unlabeled target domain.

B. Multisource Transfer Learning

Multisource transfer learning is always a hot topic, since there are generally multiple sources available for knowledge transfer in the target learning [19]. Although multiple sources would bring more knowledge, they further lead to a challenging transfer learning problem, since multiple sources have a large divergence within them. To this end, there are a lot of approaches proposed to address the multisource scenario in many real-world applications [5], [11]–[15], [18], [31].

Multitask learning is one popular technique for multisource transfer learning [18], where multiple sources are well aligned under multitask strategy. Along this line, Jhuo *et al.* [5] developed a multisource domain adaptation algorithm with low-rank constraint, which further aligns multiple source by adding a rank constraint on multiple rotated sources together. However, unlike this paper, no supervised knowledge is considered in [5]. Moreover, current multisource transfer learning all assumes the source is complete.

Different from the existing models, our proposed algorithm tends to seek a latent domain-free shared subspace, in which prior structure and cross-source regularizers are developed to couple multiple sources during knowledge transfer. Compared with [5], our model could release the computation burden by designing an efficient term. Furthermore, we propose two cross-source alignment regularizers to further explore the supervised knowledge from multiple sources and intrinsic information of target. This paper is based on our conference paper [27]; however, this paper develops a more efficient framework to tackle with the ITML. Differently, we deploy latent low-rank transfer learning guided by iterative structure learning to transfer knowledge from each single source to target domain, which reinforces to compensate for any missing data in each source by the complete target data. Second, through the effective multisource alignment, the learned nonlinear source features are very discriminative. Thus, it could well serve the role as a powerful dictionary. In this way, the learned projection could carry discriminative knowledge from multiple sources.

TABLE I

NOTATIONS AND DESCRIPTIONS

Notation	Description
$X_{\mathrm{S},i}$	the i-th source domain feature
$Y_{\mathrm{S},i}$	low-dimensional features for i-th source domain
X_{T}	target domain feature
Y_{T}	pre-learned low-dimensional feature of target domain
P	Domain-invariant Linear Projection
$Z_i/\mathcal{Z}_i$	the i-th low-rank reconstruction coefficients
L_i	the i-th latent low-rank recovering matrix
H_i	the i-th iterative structure matrix
E_i	the i-th sparse error matrix
n_{s_i}, n_t	# i-th source/target examples
d, p	# original/low-dimensional features

III. Incomplete Multisource Transfer Learning

In this section, we provide the IMTL approach for effective and robust multisource knowledge transfer. First of all, we start with the variable definitions of terminologies. For clarity, Table I lists the frequently used notations.

For IMTL scenarios, target domain data $X_T \in \mathbb{R}^{d\times n_t}$ include C classes but unlabeled, where d is the dimensionality of original space and n_t is the sample size of the target domain; K sources $X_S = [X_{S_1}, \cdots, X_{S_K}]$ also have C classes, but none of the single source $X_{S_i} \in \mathbb{R}^{d\times n_{s_i}}$ (n_{s_i} is the size of the ith source) can cover all the C classes in the target domain. Each source and target data are distributed differently, i.e., $X_{S_i} \subsetneq \mathrm{span}(X_T)$. Motivated by recent proposed transfer subspace learning [7], [25], we also devote to seek a latent space shared by sources and target domains, where the distribution divergence across multiple sources and target could be mitigated, thus the discriminative knowledge in multiple sources could be adapted to facilitate the target learning. In the following, we will discuss our IMTL in detail.

A. Effective Incomplete Multisource Alignment

Since multiple sources may have the different distributions, it is essential to align them well, so that they could contribute effective knowledge to boost the target learning. Following traditional multisource transfer learning [5], [17], we aim to align multiple sources in a domain-free low-dimensional space. Suppose $Y_{S_i} \in \mathbb{R}^{p\times n_{s_i}}$ is the learned low-dimensional feature for the ith source. In particular, we develop two graphs to effectively learn the low-dimensional feature $Y_S = [Y_{S_1}, \cdots, Y_{S_K}] \in \mathbb{R}^{p\times n_s} (n_s = \sum_i n_{s_i})$ in a nonlinear fashion, which can be detailed as follows:

$$Y_S = \arg\min_{Y_S} \frac{\mathrm{tr}\big(Y_S^\top S_w Y_S\big)}{\mathrm{tr}\big(Y_S^\top S_b Y_S\big)} \tag{1}$$

where $S_w \in \mathbb{R}^{n_s\times n_s}$ and $S_b \in \mathbb{R}^{n_s\times n_s}$ are the graph Laplacian of two graphs, i.e., within-class graph and between-class graph

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

on all sources data X_S, respectively [32]. The goal is to learn discriminative low-dimensional features, which can preserve more within-class compactness while keeping between-class discriminability. In particular, entries of affinity matrices for two graphs are defined as follows:

$$G_w^{jk} = \begin{cases} \exp\left(-\dfrac{\|x_j - x_k\|^2}{2\sigma^2}\right), & \text{if } x_j \text{ and } x_k \\ & \text{have the same label} \\ 0, & \text{otherwise} \end{cases}$$

$$G_b^{jk} = \begin{cases} \exp\left(-\dfrac{\|x_j - x_k\|^2}{2\sigma^2}\right), & \text{if } x_j \text{ and } x_k \\ & \text{have different labels} \\ 0, & \text{otherwise} \end{cases}$$

where $x_{k/j}$ values are the k/jth sample of X_S, respectively. In particular, σ is the so-called bandwidth for Gaussian kernel (in the paper, we set $\sigma = 10$ for simplicity). $S_{w/b} = G_{w/b} - D_{w/b}$, where $D_{w/b}$ values are diagonal matrices with the ith element as $D_{w/b}^{ii} = \sum_j G_{w/b}^{ij}$. In particular, Y_S can be effectively obtained by solving the following eigendecomposition problem:

$$S_w Y_S = \vartheta S_b Y_S \tag{2}$$

where Y_S values are the eigenvectors corresponding to the minimum p eigenvalues.

Notably, this is very similar to the idea of linear discriminative analysis extended to multiview learning. However, (1) falls in the nonlinear dimensionality reduction category. With the learned low-dimensional features Y_S, the distribution divergence of multiple sources in the original space could be mitigated in the low-dimensional space. In Section III-B, we will present the novel latent low-rank transfer learning for incomplete multiple sources.

B. Cross-Domain Knowledge Transfer Through Latent Low-Rank Constraint and Iterative Structure Learning

In IMTL, each source cannot cover all the categories in the target. Therefore, recovering missing source data becomes necessary for effective knowledge transfer. Next, we will present our cross-domain transfer learning in detail.

To recover the missing data $Y_{S_i}^u$ in each Y_{S_i}, we first assume it is observable. Then, we optimize our objective by considering all source data: $Y_{S_i} = [Y_{S_i}^o, Y_{S_i}^u]$, and derive the formulation by assuming $Y_{S_i}^u$ is missing in the ITML ($Y_{S_i}^o$ indicates the observable source data). In particular, in our problem, target data can be reconstructed by each source data in a shared domain-invariant subspace $P \in \mathbb{R}^{d\times p}$ through a low-rank constraint, each of which can be seen as a unique learning task. Therefore, multitask learning framework could be exploited into our knowledge transfer problem. Given a learned P for target data, we can formulate a naive multiple sources transfer learning framework as

$$\begin{aligned} &\min_{\mathcal{Z}_i} \sum_{i=1}^{K} \operatorname{rank}(\mathcal{Z}_i) \\ &\text{s.t. } P^\top X_T = Y_{S_i}\mathcal{Z}_i, \quad i = 1, \ldots, K \end{aligned} \tag{3}$$

where rank$(\cdot)$ is the rank operator of matrix and $\mathcal{Z}_i \in \mathbb{R}^{n_{s_i}\times n_t}$ is the ith low-rank reconstruction matrix, which guides locality aware reconstruction across target and each source. The rank minimization is a well-known NP-hard problem. Recent studies [20], [25] relax the rank minimization to its convex surrogate, that is, nuclear norm. Hence, (3) can be converted into its equivalent optimization

$$\begin{aligned} &\min_{\mathcal{Z}_i} \sum_{i=1}^{K} \|\mathcal{Z}_i\|_* \\ &\text{s.t. } P^\top X_T = Y_{S_i}\mathcal{Z}_i, \quad i = 1, \ldots, K \end{aligned} \tag{4}$$

where $\|\cdot\|_*$ is nuclear norm calculating the sum of singular values of a matrix.

Assuming the above-mentioned objective function has a unique solution, then we can derive $P^\top X_T \subseteq \operatorname{span}(Y_{S_i})$ in subspace P. Suppose $[P^\top X_T, Y_{S_i}] = U\Sigma V^\top$ and $V = [V_{\mathrm{T}}; V_{S_i}]$, where $P^\top X_{\mathrm{T}} = U\Sigma V_{\mathrm{T}}^\top$ and $Y_{S_i} = U\Sigma V_{S_i}^\top$, then we can immediately deduct the constraint as $U\Sigma V_T^\top = U\Sigma V_{S_i}^\top \mathcal{Z}_i$. Therefore, we have

$$\begin{aligned} &\min_{\mathcal{Z}_i} \sum_{i=1}^{K} \|\mathcal{Z}_i\|_* \\ &\text{s.t. } V_T^\top = V_{S_i}^\top \mathcal{Z}_i, \quad i = 1, \ldots, K \end{aligned}$$

whose optimal low-rank representation is $\mathcal{Z}_i^* = V_{S_i} V_T^\top = [V_{S_i}^o; V_{S_i}^u]V_T^\top$, where $V_{S_i}^o$ and $V_{S_i}^u$ are row partitions of V_{S_i}. The constraint can be rewritten into

$$\begin{aligned} P^\top X_T &= Y_{S_i}\mathcal{Z}_i^* = \left[Y_{S_i}^o, Y_{S_i}^u\right]\mathcal{Z}_i^* \\ &= \left[Y_{S_i}^o, Y_{S_i}^u\right]\left[V_{S_i}^o; V_{S_i}^u\right]V_T^\top \\ &= Y_{S_i}^o\left(V_{S_i}^o V_T^\top\right) + U\Sigma\left(V_{S_i}^u\right)^T V_{S_i}^u V_T^\top \\ &= Y_{S_i}^o Z_i + \left(U\Sigma\left(V_{S_i}^u\right)^\top V_{S_i}^u \Sigma^{-1} U^\top\right) P^\top X_T \end{aligned}$$

where $L_i = U\Sigma(V_{S_i}^u)^\top V_{S_i}^u \Sigma^{-1} U^\top$ is encouraged to be low-rank to recover the structure of $Y_{S_i}^u$. Since we assume Y_{S_i} and $P^\top X_T$ are drawn from the same collection of low-rank subspaces, the union of the subspaces has a rank of r. Hence, we could derive that $\operatorname{rank}(Z_i) \leq r$ and $\operatorname{rank}(L_i) \leq r$. From the above-mentioned deduction, it is known that even if Y_{S_i} has unobserved data $Y_{S_i}^u$, i.e., $Y_{S_i} = Y_{S_i}^o$, we can still recover it by imposing additional constraints, namely

$$\begin{aligned} &\min_{Z_i,\ L_i} \sum_{i=1}^{K} (\|Z_i\|_* + \|L_i\|_*) \\ &\text{s.t. } P^\top X_T = Y_{S_i} Z_i + L_i P^\top X_T, \quad i = 1, \ldots, K. \end{aligned} \tag{5}$$

From geometrical point of view, (5) actually presents a way to reconstruct projected target data $P^\top X_T$ through two directions: column ($Y_{S_i} Z_i$) and row ($L_i P^\top X_T$). While the column reconstruction is usually recognized as the dictionary learning, we refer the latter part, row reconstruction as latent factors. When some categories in Y_{S_i} are missing, i.e., some columns are empty, it is beneficial to reconstruct along the rows of $P^\top X_T$. Such benefits have been discussed in [22] and [33], which manages to recover the missing data from the data itself. Interesting examples in visual recognition can

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

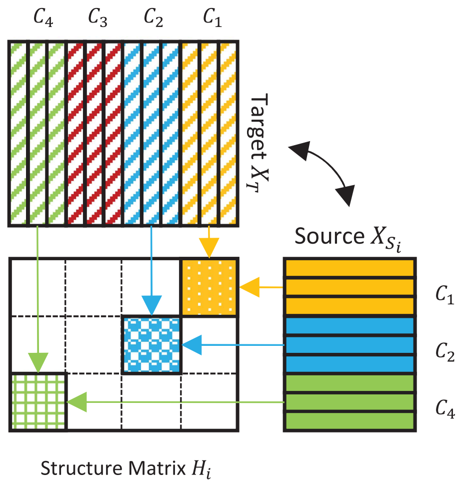


Fig. 3. Illustration of structure matrix H_i for pseudolabeled target X_T with four classes and the ith labeled source X_{S_i} with three classes. H_i only has positive values at the positions where X_T and X_{S_i} share the same labels, otherwise H_i is 0. The same color denotes the same category.

be found from the experiments in [22] and [33], where the column space represents the principle features while the row space corresponds to the key object parts and is usually discriminative for recognition task. Differently, our algorithm aims to recover the missing data through two directions for source data for knowledge transfer.

From now on, we only provide the latent transfer model based on data distributions in an unsupervised fashion. To further exploit supervised information of source and target domains, we could involve the supervised information during knowledge transfer, so that source data with certain categories can only be reconstructed by target data with the same category. In particular, for each source-target reconstruction task, we design a structured term H_i to carry corresponding supervised information [29]. Moreover, the original latent low-rank constraint is further relaxed by incorporated with a sparse error term $E_i \in \mathbb{R}^{p\times n}$. This brings twofold benefits to our model. One is that it converts the original hard constraint to a soft one, which could avoid the potential overfitting issue. The other is that, in practice, term E_i could compensate to remove the data noise if we jointly minimize its $L_{2,1}$-norm [7], [25]

$$\min_{P,Z_i,L_i,H_i,E_i} \sum_{i=1}^{K}\left(\|Z_i\|_*+\|L_i\|_*+\lambda\|E_i\|_{2,1}+\frac{\alpha}{2}\|Z_i-H_i\|_F^2\right)$$
$$\text{s.t. } P^\top X_T = Y_{S_i}Z_i + L_iP^\top X_T + E_i, \quad i=1,\ldots,K \tag{6}$$

where α and λ are the balance parameters, and $\|\cdot\|_F^2$ is the Frobenius norm. However, since we have access to limited or none labeled data of the target domain, a predefined structured term H_i is usually inaccurate, which may further mislead the knowledge transfer. Therefore, we develop to iteratively optimize H_i after each iteration of transfer subspace learning in an EM-like manner. The temporary recognition results in the previous iteration in both source and target domains are utilized as supervised knowledge for the next iteration of knowledge transfer. In the ideal case, H_i will converge after several iterations, and we call this learning process as *iterative structure learning* (see Fig. 3).

Discussions: In general, researchers adopt maximum mean discrepancy (MMD) to address the MD difference [6]. That is to minimize the distance of two domains in the transformed space (reproducing kernel Hilbert space), while we adopt low-rank reconstruction to address the MD divergence, that is, $P^\top X_T \approx Y_{S_i}Z_i$. Each target sample would be close to the same class source samples in the new space, so that we can build a connection to MMD. In this way, we minimize the difference of MD. For the CD, researchers usually use the revised MMD (that is to minimize the mean of two domains for each similar class), or SVM-based transfer learning (which relies on the target labels). Differently, we introduce the iterative structure learning, which encourages target data to be only correlated with the same class source data. Therefore, the CD difference would be mitigated.

C. Cross-Source Knowledge Alignment

Model in (6) transfers knowledge from each single source to target domain independently. Thus, it is essential to couple multiple tasks to guide effective knowledge transfer learning. Recall our latent low-rank constraint $P^\top X_T = Y_{S_i}Z_i + L_iP^\top X_T + E_i$, if we remove the error term, it could be reformulated as $(I-L_i)P^\top X_T \approx Y_{S_i}Z_i$. In this way, we could conclude that rotated low-dimensional target data are reconstructed by each low-dimensional source data under a low-rank constraint. Therefore, $Y_{S_i}Z_i$ can be treated as one version low-dimensional feature of X_T, while $P^\top X_T$ is also one version feature. Interestingly, in (6), we also observe that X_T has $K+1$ versions of low-dimensional representations, i.e., $K+1$ different features. To encourage the consistency among different features, we consider that each feature should well preserve the manifold structure of the target domain. Actually, the unlabeled data in the target domain are capable of revealing the intrinsic structure of the target domain, e.g., the sample variances and manifold structure. Therefore, we define the following manifold regularizer:

$$\mathcal{R}_m(Z_i,P) = \sum_{j=1}^{n_t}\sum_{k=1}^{n_t}\left(\sum_{i=1}^{K}\left(Y_{S_i}Z_i^j - Y_{S_i}Z_i^k\right)^2 + \left(P^\top X_{T,j} - P^\top X_{T,k}\right)^2\right)W_{j,k} \tag{7}$$

where $Z_i^{j/k}$ values are the j/kth column of Z_i, and $X_{T,j/k}$ values are the j/kth column of X_T, while W is the weight matrix of the manifold graph on the target data, whose elements are defined as follows:

$$W_{j,k} = \begin{cases} 1, & \text{if } X_{T,j} \in \mathcal{N}_{\bar{\kappa}}(X_{T,k}) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $X_{T,j} \in \mathcal{N}_{\bar{\kappa}}(X_{T,k})$ means $X_{T,k}$ is the $\bar{\kappa}$ nearest neighbor of the same class data $X_{T,k}$.

Furthermore, we reformulate (7) as follows:

$$\mathcal{R}_m(Z_i,P) = \sum_{i=1}^{K}\operatorname{tr}(Y_{S_i}Z_i\mathcal{L}(Y_{S_i}Z_i)^\top) + \operatorname{tr}\left(P^\top X_T\mathcal{L}X_T^\top P\right) \tag{9}$$

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

where $\mathcal{L} = D - W$ is the graph Laplacian of W [34]. In particular, D is a diagonal matrix and its ith element is calculated as $D_{ii} = \sum_j W_{ij}$.

Discussions: Our manifold regularizer attempts to preserve more intrinsic structure within the target data. From the manifold assumption [35], we could conclude that the CDs $\mathcal{Q}(y_i|x_i)$ and $\mathcal{Q}(y_j|x_j)$ tend to be similar if two data samples $x_i, x_j \in X_T$ are close in the intrinsic geometry of the MDs $\mathcal{P}(x_i)$ and $\mathcal{P}(x_j)$. In this practice, the manifold structure within the target data would be preserved for K different kinds of reconstructed features on different sources and its own representation. In this way, we could uncover more intrinsic structure of the target data during model training. To explore more, the manifold structure of the target would guide the consistency of multiple reconstructed features, so that such strategy would align multiple tasks during knowledge transfer. This can be treated as another way to align multiple sources.

To sum up, the final objective function can be rewritten as follows:

$$\begin{aligned}\min_{\substack{P,Z_i,L_i,\\H_i,E_i}} \sum_{i=1}^{K}\Bigg(&\|Z_i\|_* + \|L_i\|_* + \lambda\|E_i\|_{2,1} + \frac{\alpha}{2}\|Z_i - H_i\|_F^2\\ &+\frac{\gamma}{2}\Big(\mathrm{tr}\big(Y_{S_i}Z_i\mathcal{L}Z_i^\top Y_{S_i}^\top\big) + \mathrm{tr}\big(P^\top X_T\mathcal{L}X_T^\top P\big)\Big)\Bigg)\\ \text{s.t. } P^\top X_T = Y_{S_i}&Z_i + L_iP^\top X_T + E_i\\ &i = 1,\ldots,K,\ P^\top P = I_p \end{aligned} \tag{10}$$

where γ is the balance parameter. Note that the orthogonal constraint $P^\top P = I_p$ is involved to make sure the learned subspace P is valid. I_p represents the identity matrix of size $p \times p$. Therefore, the proposed framework in (10) can minimize the MD and CD disparity from two directions for incomplete multisource problem.

Discussion: Compared with our two previous work [21], [25], which also follow in latent low-rank transfer learning scenario and aim to uncover missing modality knowledge in the target domain, our algorithm is designed to address the IMTL problem. That is, our current algorithm tends to uncover the missing knowledge in sources with the help of the target data. Technically, we also prelearn low-dimensional features for the source domains, compared with [25]. However, we adopt sources to reconstruct the target while [25] adopted the opposite direction. Through the effective multisource alignment, the learned source features Y_S are already discriminative, thus it could well serve the role of a powerful dictionary. In this way, the learned projection could carry more discriminative knowledge from sources. Moreover, we incorporate an iterative structure learning to further transfer more discriminative information.

Second, the manifold regularizer aims to uncover the intrinsic structure of the target; however, our novel manifold regularizer is different from previous work [21], [25]. Actually, we aim to adopt the same manifold structure to couple $K + 1$ versions of target features; meanwhile, we could well align multiple sources. To this end, our algorithm could transfer more knowledge from multiple sources. For the Fisher regularizer, we aim to align multiple sources, so that the distribution difference across them could be mitigated. In this way, they can effectively boost the target learning. In our new version, we adopt nonlinear dimensionality reduction scheme to prelearn discriminative low-dimensional features for sources, which work as a basis in the low-rank transfer learning constraint.

D. Solving Objective Function

Since we have an iterative structure matrix H_i for each source, we adopt an EM-like refinement strategy to optimize the variables. In particular, in the E-step, we fix H_i and optimize the other variables P, Z_i, L_i, and E_i; in the M-step, we update the iterative structure matrix H_i with the optimized P. Therefore, we could iteratively update two steps until it converges.

Before two-step optimization, we transform problem (10) into its equivalent one by mitigating the orthogonal constraint, following previous work [36] (in fact, we could adopt Manopt [37] to address the optimization of P with orthogonal constraint, but it costs a lot of time). First of all, we optimize the following equation as:

$$\begin{aligned}&\min_{P} \mathrm{tr}\big(P^\top X_T\mathcal{L}X_T^\top P\big)\\ &\text{s.t. } P^\top P = I_p\end{aligned} \tag{11}$$

where $Y_T = P^\top X_T$ is the low-dimensional feature of target domain with its own knowledge. To that end, we can prelearn Y_T then transform problem (10) into

$$\begin{aligned}\min_{\substack{P,Z_i,L_i,\\H_i,E_i}} \sum_{i=1}^{K}\Bigg(&\|Z_i\|_* + \|L_i\|_* + \lambda\|E_i\|_{2,1} + \frac{\alpha}{2}\|Z_i - H_i\|_F^2\\ &+\frac{\gamma}{2}\Big(\mathrm{tr}\big(Y_{S_i}Z_i\mathcal{L}Z_i^\top Y_{S_i}^\top\big) + \|Y_T - P^\top X_T\|_F^2\Big)\Bigg)\\ \text{s.t. } P^\top X_T = Y_{S_i}&Z_i + L_iP^\top X_T + E_i, \quad i = 1,\ldots,K \end{aligned} \tag{12}$$

in which we can observe that the square loss could replace with the graph regularizer in problem (10).

1) E-Step: With the iterative structure H_i fixed, problem (12) could be solved by off-the-shelf algorithms, such as augmented Lagrange methods (ALMs) [38]. However, we need involve extra variables if we apply ALM, which leads to additional complex matrix operations during iterative optimization. To release the computational burden, we apply the first-order Taylor expansion approximation to avoid the original quadratic term, resulting in a simpler optimization.

To address this problem, we convert (12) to the augmented Lagrangian function as

$$\begin{aligned}\sum_{i=1}^{K}\Bigg(&\|Z_i\|_* + \|L_i\|_* + \lambda\|E_i\|_{2,1} + \frac{\alpha}{2}\|Z_i - H_i\|_F^2\\ &+\frac{\mu}{2}\left\|P^\top X_T - Y_{S_i}Z_i - L_iP^\top X_T - E_i + \frac{Q_i}{\mu}\right\|_F^2\\ &+\frac{\gamma}{2}\mathrm{tr}\big(Y_{S_i}Z_i\mathcal{L}Z_i^\top Y_{S_i}^\top\big)\Bigg) + \frac{\gamma}{2}\|Y_T - P^\top X_T\|_F^2\end{aligned}$$

where Q_i values are the Lagrange multipliers and μ is a positive penalty parameter. $\langle\cdot\rangle$ denotes the inner product of two

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

matrices. To solve the above-mentioned problem, we define $\mathcal{F}^i = \mathcal{F}^i(P, Z_i, L_i, E_i, Q_i, \mu) = (\gamma/2)\mathrm{tr}(Y_{S_i} Z_i \mathcal{L} Z_i^\top Y_{S_i}^\top) + (\alpha/2)\|Z_i - H_i\|_F^2 + (\mu/2)\|P^\top X_T - Y_{S_i} Z_i - L_i P^\top X_T - E_i + (Q_i/\mu)\|_F^2$ for simplicity.

Like the conventional ALM, it is impossible to jointly update Z_i, L_i, E_i, and P, but it is still solvable over each of them in leave-one-out fashion. Hence, we address each subproblem iteratively. In detail, we alternately optimize the variables P, Z_i, L_i, and E_i at $t+1$ iteration in the following.

a) Update Z_i:

$$\begin{aligned} Z_{i,t+1} &= \arg\min_{Z_i} \|Z_i\|_* + \mathcal{F}^i(P_t, Z_i, L_{i,t}, E_{i,t}, Q_{i,t}, \mu_t) \\ &= \arg\min_{Z_i} \frac{1}{\eta_{Z_i}^t \mu_t} \|Z_i\|_* + \frac{1}{2}\left\| Z_i - Z_{i,t} + \frac{\nabla_{Z_i}\mathcal{F}_t^i}{\mu_t} \right\|_F^2 \end{aligned} \tag{13}$$

where $\nabla_{Z_i}\mathcal{F}_t^i$ is the partial differential of $\mathcal{F}^i$ with respect to Z_i at time t and $\eta_{Z_i}^t = \|Y_{S,i}\|_2^2$. In particular, we have

$$\begin{aligned} \nabla_{Z_i}\mathcal{F}_t^i = {} & -\mu_t Y_{S_i}^\top \big(P_t^\top X_T - Y_{S_i} Z_{i,t} - L_{i,t} P_t^\top X_T - E_i\big) \\ & - Y_{S_i}^\top Q_{i,t} + \alpha(Z_{i,t} - H_i) + \gamma Y_{S_i}^\top Y_{S_i} Z_{i,t} \mathcal{L}. \end{aligned}$$

We can apply the singular value thresholding (SVT) [39] to tackle with problem (13) effectively. Suppose $U_{Z_i}\Sigma_{Z_i}V_{Z_i}$ are the SVD of matrix $(Z_{i,t} - \nabla_{Z_i}\mathcal{F}_t^i)$, in which $\Sigma_{Z_i} = \mathrm{diag}(\{\sigma_i\}_{1 \le i \le r})$ with singular value σ_i. Therefore, we could obtain the optimal of Z_i at time $t+1$ as $Z_{i,t+1} = U_{Z_i}\Omega_{(\frac{1}{\mu_t})}(\Sigma_{Z_i})V_{Z_i}$, in which $\Omega_{(\frac{1}{\mu_t})} = \mathrm{diag}(\{\sigma_i - \frac{1}{\mu_t}\}_+)$, and a_+ represents the positive part of a [39].

b) Update L_i:

$$L_{i,t+1} = \arg\min_{L_i} \frac{1}{\eta_{L_i}^t \mu_t} \|L_i\|_* + \frac{1}{2}\left\| L_i - L_{i,t} + \frac{\nabla_{L_i}\mathcal{F}_t^i}{\mu_t} \right\|_F^2 \tag{14}$$

where

$$\begin{aligned} \nabla_{L_i}\mathcal{F}_t^i = {} & -\mu_t \\ & \times \left(P_t^\top X_T - Y_{S_i} Z_{i,t+1} - L_{i,t} P_t^\top X_T - E_{i,t} + \frac{Q_{i,t}}{\mu_t} \right) X_T^\top P_t \end{aligned}$$

is the partial differential of $\mathcal{F}^i$ with respect to L_i at time t, and $\eta_{L_i}^t = \|P_t^\top X_T\|_2^2$. Problem (14) can also be solved via SVT operator [39] in the same way as problem (13).

c) Update E_i:

$$\begin{aligned} E_{i,t+1} = {} & \arg\min_{E_i} \frac{\lambda}{\mu_t}\|E_i\|_{2,1} \\ & + \frac{1}{2}\left\| E_i - P_t^\top X_T + Y_{S_i} Z_{i,t+1} + L_{i,t+1} P_t^\top X_T - \frac{Q_{i,t}}{\mu_t} \right\|_F^2 \end{aligned} \tag{15}$$

which is easily addressed with [40].

d) Update P:

$$\begin{aligned} P_{t+1} = {} & \arg\min_P \frac{\gamma}{2}\|Y_T - P^\top X_T\|_F^2 + \frac{\mu}{2}\sum_{i=1}^K \\ & \left\| P^\top X_T - Y_{S_i} Z_{i,t+1} - L_{i,t+1} P^\top X_T - E_{i,t+1} + \frac{Q_{i,t}}{\mu_t} \right\|_F^2 \end{aligned} \tag{16}$$

which has a closed-form solution as

$$\begin{aligned} & X_T X_T^\top P \left(\gamma I_p + \mu_t \sum_{i=1}^K (I_p - L_{i,t+1})^\top (I_p - L_{i,t+1}) \right) \\ & \quad = X_T \left(\gamma Y_T^\top + \mu_t \sum_{i=1}^K \bar{Z}_{i,t+1}^\top (I_p - L_{i,t+1}) \right) \end{aligned} \tag{17}$$

where $\bar{Z}_{i,t+1} = Y_{S_i} Z_{i,t+1} + E_{i,t+1} - (Q_{i,t}/\mu_t)$. Then, we can calculate P_{t+1} as

$$P_{t+1} = \left(X_T X_T^\top\right)^{-1} X_T P_a (P_b)^{-1} \tag{18}$$

where

$$P_a = \gamma Y_T^\top + \mu_t \sum_{i=1}^K \bar{Z}_{i,t+1}^\top (I_p - L_{i,t+1})$$

and

$$P_b = \gamma I_p + \mu_t \sum_{i=1}^K (I_p - L_{i,t+1})^\top (I_p - L_{i,t+1}).$$

2) M-Step: When the projection is optimized, we can extract feature for sources and target domain, and then apply the nearest neighbor classifier to predict the label of the target data using the labeled sources data. It is worth noting that, we can generally achieve a more accurate labeling for the target data with a more effective projection. Thus, if we adopt such labeling strategy as the pseudotarget labels to trigger the projection learning in an iterative way, then we can alternatively improve the labeling quality and subspace learning until convergence.

When the source labels and pseudolabels of target are available, we can update $H_i \in \mathbb{R}^{n_{s_i} \times n_t}$ to guide knowledge transfer during low-rank reconstruction. In particular, the element $H_i^{j,k}$ denotes the element of the jth row and the kth column in H_i, which is optimized as

$$H_i^{j,k} = \frac{\delta\big(P^\top X_{S_i,j}, P^\top X_{T,k}\big)}{\sum_{l_j = \bar{l}_k} \delta\big(P^\top X_{S_i,j}, P^\top X_{T,k}\big)} \tag{19}$$

where l_j means the label of the jth sample in the ith source X_{S_i} and $\bar{l}_k$ denotes that the pseudolabel of $X_{T,k}$ is the kth sample of X_T. And $\delta(P^\top X_{S_i,j}, P^\top X_{T,k}) = \exp(-\|P^\top X_{S_i,j} - P^\top X_{T,k}\|^2/2\omega^2)$, where we set $\omega = 5$ in our experiment.

The details of E-step and M-step are summarized in Algorithm 1, where the parameters μ, ρ, ϵ, and $\mu_{\max}$ are set empirically [20], [41]. In particular, μ and ϵ are set as small values from 10^{-6} to 10^{-3}, and ρ controls the step size, which is usually set from 1.1 to 1.3, while $\mu_{\max}$ is usually set as 10^6 to control the penalty term. Other balance parameters λ, α, and γ for error term, iterative structured term, and cross-source alignment, respectively, are tuned in the experiment.

E. Complexity Analysis

In this section, we show the complexity analysis of our algorithm (Algorithm 1).

First of all, incomplete multisource alignment costs for the eigendecomposition of (2), which needs $O(n_s^3)$ for S_b and S_w, are both $n_s \times n_s$ matrices. It can be reduced to $O(n_s^{2.376})$ using

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Solution to Problem (Eq. (12))

```
Input: $X_T, X_{S_i}, \alpha, \gamma, \lambda, \mathcal{L}$
Initialize: $Z_{i,0} = L_{i,0} = E_{i,0} = Q_{i,0} = 0, t = 0,$
            $\mu_0 = 10^{-6}, \rho = 1.2, \mu_{\max} = 10^6, \epsilon = 10^{-6}$.

while not converged do
    1. Update $Z_{i,t+1}$ using Eq. (13) by fixing others;
    2. Update $L_{i,t+1}$ using Eq. (14) by fixing others;
    3. Update $E_{i,t+1}$ using Eq. (15) by fixing others;
    4. Update $P_{t+1}$ using Eq. (17) by fixing others;
    5. Update $H_{i,t+1}$ using Eq. (19) by fixing others;
    6. Optimize the multipliers $Q_{i,t+1}$ using
          $Q_{i,t+1} = Q_{i,t}$
          $+ \mu_t(P_{t+1}^\top X_T - Y_{S_i} Z_{i,t+1} - L_{i,t+1} P_{t+1}^\top X_T - E_{i,t+1})$
    7. Optimize the parameter $\mu_{t+1}$ using $\mu_{t+1} = \min(\rho\mu_t, \mu_{\max})$;
    8. Check the convergence conditions
          $\|P_{t+1}^\top X_T - Y_{S_i} Z_{i,t+1} - L_{i,t+1} P_{t+1}^\top X_T - E_{i,t+1}\|_\infty < \epsilon$.
    9. $t = t + 1$.
end while

Output: $Z_i, L_i, E_i, P, H_i$.
```

the Coppersmith–Winograd algorithm [42]. Also, we prelearn Y_T in (11), which costs $O(n_t^{2.376})$.

Second, the main time-consuming components of E-Step are: trace norm computation in Steps 1 & 2 and projection learning in Step 4. In particular, each trace norm solved by SVD computation in Step 1 takes $O(n_{s_i}^2 n_t)$ for $Z_i \in \mathbb{R}^{n_{s_i} \times n_t}$. In general, n_t has the smaller order of magnitude compared with n_{s_i}. When n_{s_i} is very large, this step would be very expensive. Fortunately, according to [20, Th. 4.3], the SVD for Z_i could be speeded up to $O(p^2 n_{s_i})$ where p is the dimensionality of Y_{S_i} and it is usually a small one. For Step 2, it could take $O(p^3)$, since $L_i \in \mathbb{R}^{p\times p}$ (note that we could ignore Step 2, since p is usually a very small number). Moreover, the optimized projection P needs $O(d^3 + d^2 p + 2p^3) \approx O(d^3)$.

Finally, the main consuming of updating H_i is to match cross-domain data with the same label. In general, this step would cost $O(n_t n_{s_i})$.

In fact, we could adopt parallel computing technique to further reduce the computational cost. To sum up, the total cost of our algorithm is about $O(n_s^{2.376} + n_t^{2.376} + \mathcal{T}_a(n_{s_i} n_t + p^2 n_{s_i} + d^3))$, where $\mathcal{T}_a$ is the iteration number.

F. Generalization Bound Analysis

In this section, we provide the analysis of our ITML's generalization error bound on the target domain [6], [43]. First of all, we define the induced prediction function as $\theta : X \to \{0, 1\}$ and the true labeling function as $h : X \to \{0, 1\}$. Therefore, the expected error of θ on target domain is calculated as

$$\epsilon_t(\theta) = \mathbb{E}_{x\sim P_t}[|h(x) - \theta(x)|].$$

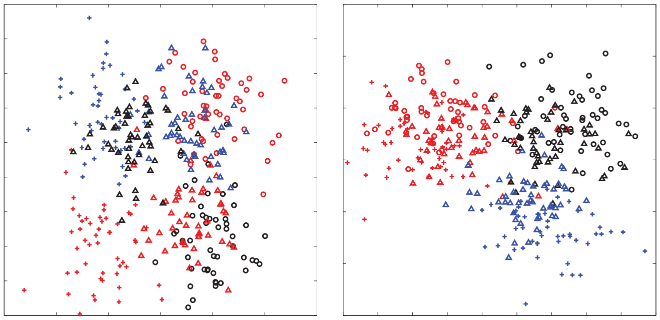


Fig. 4. Left: data distributions before transfer. Right: data distributions after transfer. Note that data in the same shape are from the same domain, and those of the same color have the same label. Cross + and circle ○ points are two-source domains while triangle △ points are target domain.

Similarly, the expected error of f in the ith source domain can be calculated, which is defined as

$$\epsilon_{s_i}(\theta) = \mathbb{E}_{x\sim P_{s_i}}[|h(x) - \theta(x)|].$$

Theorem 1: Assuming that the hypothesis space including θ is with VC-dimension as d, the expected error of θ on the target domain is bounded by each source with probability $1-\delta$ as

$$\epsilon_t(\theta) \le \hat{\epsilon}_{s_i}(\theta) + \sqrt{\frac{4}{n_{s_i}} d \log\frac{2en_{s_i}}{d} + \log\frac{4}{\delta}} + \mathcal{D}(\mathcal{S}_i, \mathcal{T}) + \Lambda$$

where $\hat{\epsilon}_{s_i}(\theta)$ is the empirical error of θ on the ith source domain, and $\Lambda = \inf_\theta[\epsilon_{s_i}(\theta) + \epsilon_t(\theta)]$.

From Theorem 1, we could observe that the expected error $\epsilon_t(\theta)$ in target domain is bounded if we jointly minimize: 1) the empirical error $\hat{\epsilon}_{s_i}(\theta)$ of the ith source domain; 2) the distribution divergence $\mathcal{D}(\mathcal{S}_i, \mathcal{T})$ across the ith source and target in the low-dimensional space; and 3) the adaptability Λ of h.

In the ITML model, we have the following approaches corresponding to the three factors described earlier. First of all, $\hat{\epsilon}_{s_i}(\theta)$ is explicitly minimized by (1). Second, $\mathcal{D}(\mathcal{S}_i, \mathcal{T})$ is explicitly minimized by latent low-rank reconstruction in (6). Third, Λ is implicitly minimized by manifold regularization in (9), which has been proved by [6].

To sum up, $\epsilon_t(\theta)$ is bounded by each source domain, so $\epsilon_t(\theta)$ is also bounded by the sum of them, corresponding to the multiple sources in our problem.

IV. Experiments

In this section, we will systematically evaluate our proposed algorithm. Before that, we first testify our algorithm with synthetic data. Then, we introduce several real-world data sets and experimental settings. Afterward, we conduct comparison on our proposed method and several transfer learning approaches. Finally, we also testify some properties of our algorithms, e.g., dimensionality influence, parameter influence, training time evaluation, and effectiveness in missing data recovery.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

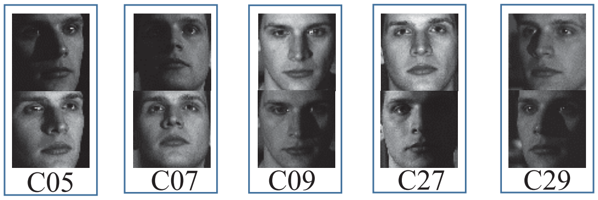


Fig. 5. Examples of five views (C05, C07, C09, C27, and C29) from the same subject in CMU-PIE face data set. We can observe that large dissimilarity exists between different views of the same subject.

Fig. 6. Example images of the headphone category from Office-10+Caltech-10 data set, where the majority of Caltech and Amazon images are from online merchants, while DSLR and Webcam images are captured from office.

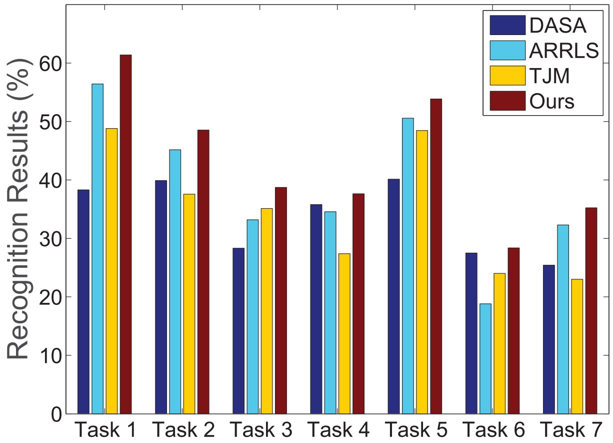


Fig. 7. Average recognition performance of seven tasks on CMU-PIE cross-pose face data set, in which Task 1: $\{C09, C05\} \rightarrow C07$, Task 2: $\{C07, C05\} \rightarrow C09$, Task 3: $\{C07, C05\} \rightarrow C09$, Task 4: $\{C09, C05, C29\} \rightarrow C07$, Task 5: $\{C09, C05, C29\} \rightarrow C27$, Task 6: $\{C09, C05, C27\} \rightarrow C29$, and Task 7: $\{C09, C05, C07\} \rightarrow C29$.

A. Synthetic Experiment

To better understand the insight behind this paper, we conduct an experiment on synthetic data. As shown in Fig. 4 (left), there are two incomplete sources, each with two classes, and single target domain with three classes. All the data are distributed in 2-D space, and generated by Gaussian distributions with different means and covariance matrices.

Before transfer, the knowledge learned from sources cannot be directly applied to target data, since it would group the target data into the wrong categories. After transfer, as shown in Fig. 4 (right), we can observe that the data with the same label of three domains are aligned well. That is, the data in the same color are grouped together. Therefore, the knowledge from the sources can be used to classify the target data. This demonstrates the effectiveness of our proposed method.

B. Real-World Data Sets

In this section, we mainly conduct experiments on three real-world benchmarks: CMU-PIE face data set[1] (Fig. 5) and visual object data sets Office-31[2] and Office-10+Caltech-10[3] (Fig. 6). Note that the arrow "→" indicates the direction of transfer learning from "sources" to "target". For example, "{Webcam, Amazon} → DSLR" means that Webcam and Amazon are the source domains while DSLR is the target one.

CMU-PIE cross-pose face data set is consisted of 68 subjects in total, which has large variances within each subject under various poses, with each under 21 different lighting variations. Five poses, i.e., $C05$, $C07$, $C09$, $C27$, and $C29$ are used in our experiments. These transfer learning algorithms, i.e., DASA [44], ARRLS [6], and TJM [45], are compared in the unsupervised adaptation setting. In other word, only the label information of multiple sources data is accessible during the model training. To construct the incomplete sources situation, 20 subjects out of 68 per pose are randomly removed for the two-source case, while 30 subjects out of 68 are randomly removed for the three-source case. In total, we conduct 20 times and report the average results of seven tasks in Fig. 7.

Office-31 is a standard benchmark for domain adaptation (see samples in Fig. 6), which includes 4,652 images within 31 categories collected from three distinct domains: Amazon (A), which contains images downloaded from amazon.com, and Webcam (W) and DSLR (D), which are taken by Web camera and digital SLR camera in an office with different environment variations, respectively. Office-10+Caltech-10 data sets contain the subsets of common categories from Office-31 and Caltech-256 data sets. The SURF features are used for these two data sets. We mainly evaluate our algorithm by compared with RDALR [5], LTSL [7], TJM [45], GFK [4], DASA [44], ARRLS [6], and SDDL [17]. Furthermore, we also conduct comparison on two multitask learning algorithms, i.e., LGO [46] and CDRL [47]. For LGO and CDRL, we adopt the logistic regression as the loss function to learn the parameters, which could be further used to predict the target labels.

For these two data sets, we adopt the semisupervised setting, where we are accessible to a small number of labeled data in the target domain. We strictly follow the settings in [48] for experiments on Office-31 and Office-10+Caltech-10. To build the incomplete situation, 6 categories out of 31 are randomly removed for each source in Office-31, while 2 categories out of 10 are randomly removed from each source in Office-10+Caltech-10. We also conduct 20 random trials and report the average performance in Tables II and III.

From the results in Fig. 7, Tables II and III, we observe that our proposed algorithm is able to outperform other transfer learning methods. The key reason is that all of them are

[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

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

TABLE II

RECOGNITION PERFORMANCE (%) OF TEN ALGORITHMS ON OFFICE-31 DATA SET, IN WHICH A = AMAZON, D = DSLR, AND W = WEBCAM

	GFK [4]	LTSL [7]	TJM [45]	DASA [44]	ARRLS [6]	RDALR [5]	SDDL [17]	LGO [46]	CDRL [47]	Ours
{A, W} → D	31.32±0.05	34.02±0.02	40.49±0.12	37.30±0.04	44.08±0.05	32.81±0.18	50.38±0.08	33.64±0.08	49.49±0.09	**51.47±0.07**
{A, D} → W	39.65±0.03	37.68±0.02	42.58±0.10	42.45±0.04	56.73±0.09	36.85±0.15	57.43±0.14	39.12±0.09	56.87±0.10	**61.31±0.09**
{D, W} → A	19.50±0.02	18.86±0.14	19.46±0.05	16.41±0.05	18.83±0.04	20.19±0.03	29.23±0.05	22.13±0.06	32.67±0.07	**42.62±0.05**

TABLE III

RECOGNITION PERFORMANCE (%) OF TEN ALGORITHMS ON OFFICE-10+CALTECH-10 DATA SET, IN WHICH A = AMAZON, D = DSLR, C = CALTECH-256, AND W = WEBCAM

	GFK [4]	LTSL [7]	TJM [45]	DASA [44]	ARRLS [6]	RDALR [5]	SDDL [17]	LGO [46]	CDRL [47]	Ours
{A, C, W} → D	23.69±0.02	38.22±0.43	41.92±0.45	50.32±0.27	51.08±0.19	54.90±0.54	65.56±0.34	39.35±0.16	63.87±0.27	**68.31±0.21**
{A, C, D} → W	36.51±0.01	42.20±0.09	33.52±0.33	52.27±0.10	49.63±0.04	61.22±0.58	71.12±0.59	43.23±0.26	69.54±0.28	**74.73±0.16**
{C, D, W} → A	26.14±0.03	30.65±0.04	30.19±0.78	37.91±0.04	31.94±0.12	50.89±0.04	51.39±0.14	31.98±0.31	54.92±0.19	**67.42±0.17**
{A, D, W} → C	18.60±0.01	23.42±0.02	21.88±0.68	32.97±0.07	31.71±0.21	44.92±0.17	34.27±0.27	24.78±0.25	39.63±0.47	**58.12±0.15**

not dealing with incomplete multisource transfer, while our algorithm is designed to guide knowledge in two directional transfer, i.e., cross-domain transfer from each source to target, and cross-source transfer. In particular, in cross-domain direction, we deploy latent low-rank transfer learning guided by iterative structure learning to transfer knowledge from each single source to target domain. This practice reinforces to compensate for any missing data in each source by the complete target data. While in cross-source direction, unsupervised manifold regularizer and effective multisource alignment are explored to jointly compensate for missing data from one portion of source to another. In this way, both MD and CD discrepancy in two directions would be mitigated.

C. Discussion

We post several observations and discussions on the comparisons.

First of all, low-rank based transfer learning, i.e., LTSL and RDALR, achieves better performance than GFK. LSTL works worse than RDALR. The reason is that LTSL is a single source method, which simply merging multiple sources together. This practice would bring in negative transfer because of the large divergence across multiple sources. However, RDALR does not explicitly couple multiple sources in a supervised way; instead, it adds a rank constraint on the rotated sources together to uncover more shared knowledge. LTSL only adopts an unsupervised low-rank transfer strategy and, therefore, is hard for effective knowledge transfer during data reconstruction.

Second, GFK and DASA are two kernel-based domain adaption algorithms. In particular, GFK aims to build a kernel metric to reduce the domain shift across two domains, while DASA adopts to align two subspaces generated by source and target to mitigate the MD. Moreover, TJM is designed to seek a subspace by simultaneously coupling the cross-domain features and reweighting the instances across domains. ARRLS also incorporates pseudolabels of target data during knowledge transfer; hence, it could mitigate the MD and CD difference at the same time. However, these four methods are all single source transfer method. Differently, our method employs the subspace learning in multitask learning with low-rank constraint, and most importantly adds two regularizers to uncover the label information and manifold structure.

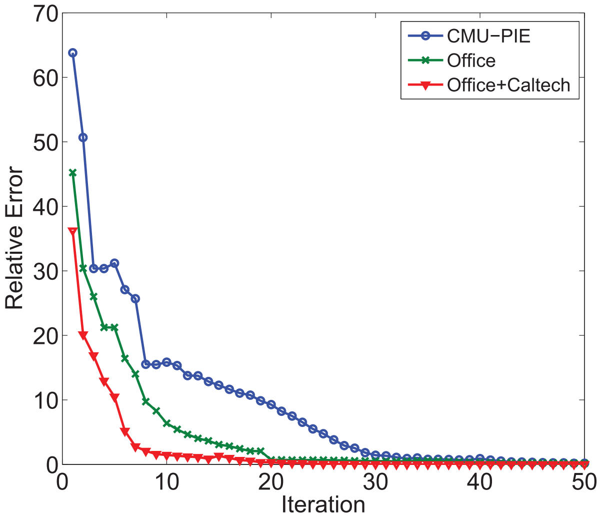


Fig. 8. Convergence curves of our algorithm on three data sets, where CMU-PIE denotes the setting $\{C09, C05\} \rightarrow C07$, Office means $\{A, D\} \rightarrow W$, and Office+Caltech represents $\{A, D, C\} \rightarrow W$. We show the results within first 50 iterations.

Third, SDDL is designed to address the multisource problem, which aims to project different sources into a common space with multiple projections. Moreover, the shared latent space is coupled with a common discriminative dictionary, which implicitly aligns different domains. However, SDDL exploits sparse constraint on data reconstruction, which fails to reveal the intrinsic classwise structure between two domains. Differently, our ITML builds a direct connection between target data and each incomplete source through low-rank constraint with an iterative structure term. Hence, our model could transfer more intrinsic knowledge from each source to the target. In particular, for incomplete sources, the introduced regularizers and latent factors would precisely couple multiple sources to compensate the missing labels with each source.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

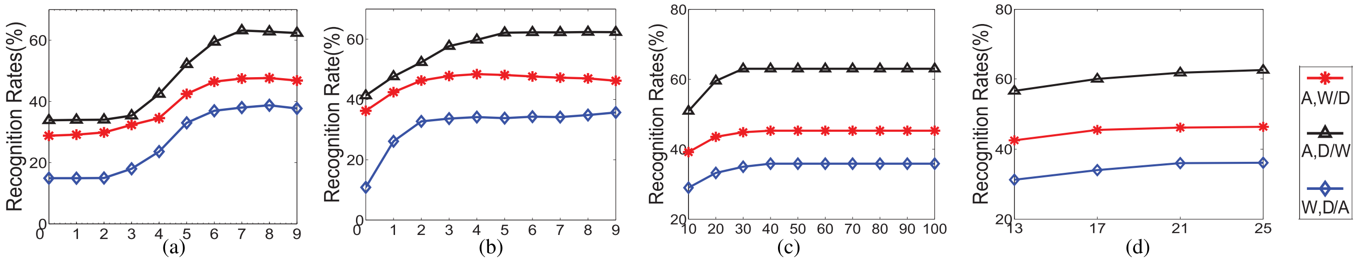


Fig. 9. Parameter evaluations on Office-31 data set, where $W, A/D$ indicates that W, A are the sources, while D is the target. (a) and (b) Influence of α on the iterative structure term and that of γ on the $\mathcal{R}_m(P, Z_i)$ term, where the $\mathbf{X}$-range from 0 to 9 means $[0, 10^{-3}, 10^{-2}, 0.05, 0.1, 0.5, 1, 10, 10^2, 10^3]$, respectively. (c) Influence of various dimensions of the common subspace projection P. (d) Recognition curves with different numbers of classes in each incomplete source.

Finally, LGO and CDRL are two multitask learning algorithms, which aim to uncover the latent shared knowledge and task-specific knowledge. In particular, we treat one domain as one task, so that we can learn shared knowledge and domain-specific knowledge to boost the target learning. CDRL adopts a group strategy to align similar samples across different domains together. Therefore, CDRL could reduce the domain shift more. For these two algorithms, the unlabeled target data cannot be utilized to train the model, that is, only multiple sources and partial labeled target are used to build the model. In this way, they may not well uncover the intrinsic structure of the target domain. Moreover, they ignore the incomplete information as traditional multisource transfer learning does. Therefore, our algorithm can beat them in all the cases.

D. Property Analysis

1) Convergence Analysis: First, we verify the convergence of our proposed algorithm through experiments. Up to now, it is still a challenge to generally guarantee the convergence with more than two blocks of ALM method [20]. Therefore, we empirically show the convergence pf our algorithm. The convergence curve of our algorithm is presented in Fig. 8. From the results, we can notice our method converges after several iterations, which means our algorithm can converge well, especially after 30 iterations.

2) Parameter Analysis: Second, we conduct experiments studying recognition performance under different input parameters α and γ to demonstrate the two incorporated novel terms play critical roles in the model learning. Besides, λ is usually set as a small value according to previous low-rank modeling [7], [20]; therefore, we set $\lambda = 10^{-2}$ for simplicity. We evaluate two parameters independently by fixing the other one. From Fig. 9(a) and (b), we can observe that the performance is worse when any of them is zero. This indicates that both of them are necessary for incomplete multiple sources. And we can observe that the recognition results become stable when both parameters are set in the range of [1, 100]. Therefore, we empirically set the parameters to 10 for Office-31 data set.

3) Dimensionality Influence: Third, we verify the dimension property of the common subspace. In Fig. 9(c), we obtain an initial sharp increase followed with a flat curve. This verifies that our method is effective even in a very low-dimensional feature space. This further demonstrates that transfer learning definitely helps the target learning when involving more source samples.

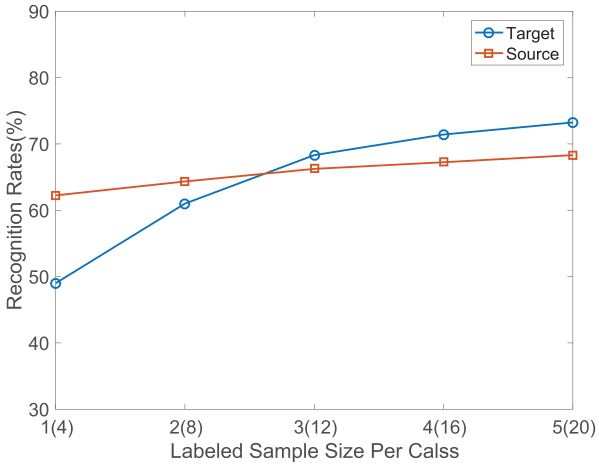


Fig. 10. Sample complexity analysis on $\{A, W, C\} \rightarrow D$ of Office-10+Caltech-10 database, where the x-axis lists the labeled sample size of target (source) per class.

TABLE IV
TRAINING TIME COST OF DIFFERENT ALGORITHMS (SECOND)

	CMU-PIE	Office-31	Office-10+Caltech-10
DASA [44]	20.22	6.12	1.15
LTSL [7]	383.27	104.46	49.65
RDALR [5]	434.12	127.23	57.23
Ours	209.23	76.18	28.76

4) Labeled Sample Complexity Analysis: We further provide the influence of different classes in each source. As shown in Fig. 9(d), the performances increase in three cases with more classes available in the source domains, but very slowly. We can see that more source samples per class would facilitate the target learning. Thus, when we have limited target samples, we can borrow the knowledge from sources to boost the target learning.

Second, we conduct one more experiment to this claim. In particular, we adopt the setting of $\{A, C, W\} \rightarrow D$ from Office-10+Caltech-10 database. On one hand, we fix the labeled source data as 20 samples per class, and then

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

TABLE V

Recognition Rate (%) on Office-10+Caltech-10 Data Set, Where $A =$ Amazon, $C =$ Caltech-256, $D =$ DSLR, and $W =$ Webcam

	C → D	W → D	A → D	A → W	D → W	C → W
DASA [44]	34.46±0.07	48.09±0.19	34.97±0.06	34.34±0.15	50.31±0.18	39.08±0.15
ARRLS [6]	32.10±0.15	50.83±0.06	28.47±0.09	36.00±0.19	60.20±0.15	37.39±0.08
TJM [45]	30.70±0.05	58.47±0.26	31.34±0.05	38.34 ±0.1	56.91± 0.2	36.63 ± 0.2
	{A, C, W} → D			{A, C, D} → W		
Ours	**65.03±0.20**			**72.58±0.25**		
	C → A	D → A	W → A	W → C	D → C	A → C
DASA [44]	34.51±0.04	32.85±0.04	33.37±0.05	23.91±0.06	23.88±0.07	28.22±0.12
ARRLS [6]	43.26±0.10	32.60±0.06	32.91±0.17	27.67±0.07	28.07±0.08	39.04±0.07
TJM [45]	38.68±0.03	37.54±0.04	36.15±0.07	28.31±0.06	29.03±0.05	34.60±0.05
	{C, D, W} → A			{A, D, W} → C		
Ours	**63.59±0.19**			**57.84±0.18**		

evaluate different labeled target sample sizes per class from 1 to 5. On the other hand, we fix the labeled target data as three samples per class, and then evaluate different labeled source sample sizes per class from 4 to 20. Two results are shown in Fig. 10, where blue curve shows the target evaluation while red curve shows the source evaluation.

From the results (Fig. 10), we could observe more labeled source samples per class would improve the recognition results; however, the increasing speed is slower than involving more labeled target samples. But, we can still conclude that our transfer learning still helps the target learning when we have limited labeled target samples for training.

5) Time Cost: We testify the training cost of various methods, i.e., DASA [44], LTSL [7], RDALR [5], and ours. In particular, we conduct experiments on MATLAB 2014 with an Intel i7-3770 PC of 32-GB memory. The training time is shown in Table IV, in which the unit is *second*. In particular, we have three cases on three databases. For CMU-PIE, we adopt the setting as $\{C09, C05\} \rightarrow C07$, and use $\{A, D\} \rightarrow W$ for Office-31, while $\{A, D, C\} \rightarrow W$ for Office-10+ Caltech-10.

From the results in Table IV, we could observe that our algorithm is more efficient than two low-rank transfer learning algorithms, i.e., LTSL and RDALR. This attributes to the efficient solution to the low-rank coefficients Z_i by avoiding relaxing variables, compared with LSTL and RDALR. Besides, RDALR and LTSL both adopt two-step strategy to address the orthogonal constraint. Moreover, since all three (LTSL, RDALR, and ours) are iteratively optimized, they would spend more than DASA.

E. Incomplete Single Source Comparison

Considerable research efforts have demonstrated that single complete source would achieve a better performance than multiple complete sources [5], [7], [17], as source domains subject to large divergence with target domain will cause negative transfer and hinder the recognition performance. In this experiment, we explore if our algorithm can exploit more from multiple incomplete sources than the single source methods using only one incomplete single source. We compare our method with several single source transfer learning algorithms, e.g., TJM [45], DASA [44], and ARRLS [6] on Office-10+Caltech-10 data set. To build the incomplete source environment, three categories out of ten are randomly removed from each source. Also, we adopt the semisupervised transfer learning setting to evaluate all the algorithms. That means, we are accessible to the labels of the target. We repeat 20 times and average results are shown in Table V.

From the results, we can observe that our algorithm works better than the competitive ones with single source, which indicates the effectiveness of our method. It should be noted that in the Office+Caltech data set, similarity between A and C is relatively high, and that of W and D is high, as shown in the work [4]. Therefore, transfer learning between W and D, or A and C yields very good results. However, including A and C into the knowledge transfer for W or D would introduce negative transfer, especially for the competitive methods, while our method can well handle this to improve the recognition performance by transferring more from multiple sources. This property is very essential, especially for the TL-IMS problem.

V. Conclusion

In this paper, we proposed an IMTL framework with structured latent low-rank constraint and cross-source alignment from two directions. First of all, our method introduced an iterative structure term with a latent factor to conventional low-rank transfer learning framework to facilitate the knowledge transfer from each source. In addition, latent factor would benefit the recovery of missing categories in each source. Secondly, two cross-source regularizers were developed to couple the highly correlated samples of multiple sources and preserve the intrinsic structure of the target data in both supervised and unsupervised fashions. With two directional transfer, both MD and CD differences were mitigated. Experiments on three data sets have shown our design ITML could better tackle with the incomplete multiple sources challenge by compared with other methods.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

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

[2] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2014.

[3] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, Nov. 2014.

[4] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[5] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2168–2175.

[6] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

[7] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *Int. J. Comput. Vis.*, vol. 109, no. 1, pp. 1–20, 2014.

[8] L. Cheng and S. J. Pan, "Semi-supervised domain adaptation on manifolds," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2240–2249, Dec. 2014.

[9] L. Yang, L. Jing, J. Yu, and M. K. Ng, "Learning transferred weights from co-occurrence data for heterogeneous transfer learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2187–2200, Nov. 2016.

[10] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, Aug. 2014.

[11] L. Duan, D. Xu, and I. W. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.

[12] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, "A two-stage weighting framework for multi-source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 505–513.

[13] L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang, "On handling negative transfer and imbalanced distributions in multiple source transfer learning," *Statist. Anal. Data Mining, ASA Data Sci. J.*, vol. 7, no. 4, pp. 254–271, 2014.

[14] L. Ge, J. Gao, and A. Zhang, "Oms-tl: A framework of online multiple source transfer learning," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 2423–2428.

[15] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko, "Discovering latent domains for multisource domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.

[16] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1855–1862.

[17] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 361–368.

[18] J. He and R. Lawrence, "A graph-based framework for multi-task multi-view learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 25–32.

[19] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3150–3157.

[20] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[21] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1192–1198.

[22] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1615–1622.

[23] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1169–1176.

[24] J. Li, A. Najmi, and R. M. Gray, "Image classification by a two-dimensional hidden Markov model," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 517–533, Feb. 2000.

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

[26] C. Jia, Y. Kong, Z. Ding, and Y. R. Fu, "Latent tensor transfer learning for RGB-D action recognition," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 87–96.

[27] Z. Ding, M. Shao, and Y. Fu, "Transfer learning for image classification with incomplete multiple sources," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Vancouver, Canada, 2016. [Online]. Available: http://www.wcci2016.org/

[28] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766.

[29] Z. Ding, M. Shao, and Y. Fu, "Deep low-rank coding for transfer learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3453–3459.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[31] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2439–2446.

[32] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[33] P. Zhou, Z. Lin, and C. Zhang, "Integrated low-rank-based discriminative feature learning for recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1080–1093, May 2016.

[34] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inf. Process. Syst.*, vol. 16. 2004, p. 153.

[35] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[36] D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *Proc. IEEE Int. Conf. Data Mining*, Oct. 2007, pp. 73–82.

[37] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1455–1459, 2014. [Online]. Available: http://www.manopt.org

[38] Z. Lin, M. Chen, and Y. Ma. (Oct. 2010). "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices." [Online]. Available: https://arxiv.org/abs/1009.5055

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

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

[41] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[42] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," in *Proc. 19th Annu. ACM Symp. Theory Comput.*, 1987, pp. 1–6.

[43] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 137–144.

[44] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.

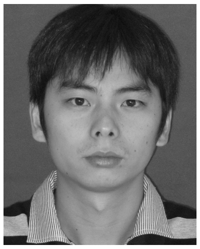
[45] M. Long, J. Wang, G. Ding, J. Sun, and P. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1410–1417.

[46] A. Kumar and H. Daume, "Learning task grouping and overlap in multi-task learning," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1383–1390.

[47] H. B. Ammar, E. Eaton, J. M. Luna, and P. Ruvolo, "Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 3345–3351.

[48] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.



Zhengming Ding (S'14) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China, Chengdu, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.

His current research interests include machine learning and computer vision, specifically involved in developing scalable algorithms for challenging problems in transfer learning and deep learning scenario.

Mr. Ding is an Association for the Advancement of Artificial Intelligence (AAAI) student member. He was a recipient of the Student Travel Grant of International Joint Conference on Artificial Intelligence 16, AAAI 16, ACM Multimedia 14, and IEEE International Conference on Data Mining (ICDM) 14. He was also a recipient of the best paper award International Society for Optics and Photonics (SPIE). He received the National Institute of Justice Fellowship. He has served as the Reviewer for the IEEE journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Ming Shao (S'11–M'16) received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science from Beihang University, Beijing, China, in 2006, 2007, and 2010, respectively, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2016.

He has been a tenure-track Assistant Professor with the College of Engineering, University of Massachusetts Dartmouth, Dartmouth, MA, USA, since 2016. His current research interests include sparse modeling, low-rank matrix analysis, deep learning, and applied machine learning on social media analytics.

Dr. Shao was a recipient of the Presidential Fellowship of the State University of New York at Buffalo from 2010 to 2012 and the best paper award winner of the IEEE ICDM 2011 Workshop on Large Scale Visual Analytics.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA.

He has been an interdisciplinary Faculty Member with the College of Engineering and the College of Computer and Information Science, Northeastern University, Boston, MA, USA, since 2012. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His current research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems.

Dr. Fu is a fellow of the International Association of Pattern Recognition (IAPR), a Lifetime Senior Member of the Association for Computing Machinery (ACM) and the SPIE, a Lifetime Member of the AAAI, the Optical Society Association, and the Institute of Mathematical Statistics, a member of the Global Young Academy, International Neural Network Society (INNS), and a Beckman Graduate Fellow from 2007 to 2008. He received seven Prestigious Young Investigator Awards from National Academy of Engineering, Office of Naval Research, Army Research Organization, Institute of Electrical and Electronics Engineers (IEEE), INNS, University of Illinois Urbana-Champaign, and Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, and Society for Industrial and Applied Mathematics; and three major Industrial Research Awards from Google, Samsung, and Adobe. He serves as an Associate Editor, the Chair, a PC Member, and a Reviewer of many top journals and international conferences/workshops. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS.

Authorized licensed use limited to: Jinan University. Downloaded on August 05,2020 at 07:02:27 UTC from IEEE Xplore. Restrictions apply.