This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

Manifold Transfer Learning Via Discriminant Regression Analysis

Yuwu Lu, *Member, IEEE*, Wenjing Wang, Chun Yuan, Xuelong Li, *Fellow, IEEE*, Zhihui Lai

***Abstract*—In transfer learning, how to effectively transfer useful information from the source domain to the target domain is crucial. In this paper, we propose a novel transfer learning method for image classification, named manifold transfer learning via discriminant regression analysis (MTL-DRA), to transfer the local geometry structure information from the source domain to the target domain and ensure that the transform matrix is robust or sparse so that samples from different domains can be well combined. In MTL-DRA, we encode discriminant information of the source domain to the target domain by introducing between- and within-class graphs to preserve within-class similarity and reduce between-class similarity. With different norms as constraints, MTL-DRA overcomes the disturbance of noise and avoids negative transfer learning. To improve the robustness of MTL-DRA, we encode a nuclear norm constraint and propose robust MTL-DRA (RMTL-DRA). We analyzed the convergence and complexity of the two proposed methods. To verify the performance of the proposed methods, we conducted extensive experiments on five public image benchmarks. The experimental results show that the proposed methods outperform state-of-the-art transfer learning methods.**

***Index Terms*—Manifold, transfer learning, regression, discriminant, image classification.**

I. Introduction

With the development of computer science, human lives, work, and entertainment have all become inseparable from the Internet, which generates a massive amount of complex data every day. By sorting out, analyzing, and mining these data, we can obtain important information about the communication, interests, habits, and hobbies of users in order to make decisions. Data mining and machine learning technologies have been widely used to deal with massive data in different fields.

Machine learning has been widely used in image recognition, speech recognition, clustering learning, and other fields [1]-[5]. However, traditional data analysis methods require sufficient labeled training data to obtain a better decision-making model, and then use this model to test and predict new data. The traditional data classification methods must satisfy the assumption that the training samples and test samples must be independent and identical distribution (IID). However, in practice, this assumption is always difficult to fulfill. For example, using a set of labeled childhood facial images to recognize their adult faces is a highly challenging task.

Different from traditional machine learning methods, transfer learning does not need to satisfy the IID assumption and can migrate existing knowledge from the source domain to the target domain to solve a problem. Transfer learning uses the existing knowledge of different domains to solve different but related problems. For example, Peng et al. [6] proposed a transfer learning method named active transfer learning (ATL) to solve the negative transfer problem. Wang et al. [7] proposed a softly associative transfer learning method for cross-domain text classification, in which two nonnegative matrix tri-factorizations are combined in a joint optimization framework. Extreme learning machines (ELMs) often must satisfy the assumption that the training data and test data are from identical distributions, but this assumption is often violated. To solve the problem in ELMs, Chen et al. [8] proposed an ELM-based space learning method, domain space transfer ELM (DST-ELM), to deal with unsupervised domain adaptation problems. To address unsupervised domain adaptation problems, Gholami et al. [9] proposed a probabilistic latent variable model, in which the categorization task is tackled from different domains. Zhang et al. [10] proposed a guide subspace learning (GSL) method for unsupervised domain adaptation. To handle nonlinear domain shift, they extended GSL into a kernel case and proposed a nonlinear GSL (NGSL). To consider the class prior for domain adaptation, Wang et al. [11] proposed a class-specific reconstruction transfer learning (CRTL) method, which fully exploited the inter-class independency and intra-class dependency. To encode the data locality structure and avoid a negative transfer effect, Zhang et al. [12] proposed a manifold criterion guided transfer learning (MCTL) method.

Manuscript received Nov. 05, 2019. This work was supported by the National Natural Science Foundation of China (Grant Nos. U1833101, 61761130079, 61732011), the Guangdong Basic and Applied Basic Research Foundation (2019A1515011493), the Tencent “Rhinoceros Birds”-Scientific Research Foundation for Young Teachers of Shenzhen University, the Natural Science Foundation of Shenzhen University (No. 2019046), and the Science Foundation of Shenzhen (Grant Nos. JCYJ20160422144110140, JCYJ20190809172201639).

Y. Lu, W. Wang, and Z. Lai are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518055, PR China, Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China, and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China. (e-mail: luyuwu2008@163.com; wangwenjing2018@email.szu.edu.cn; lai_zhi_hui@163.com).

C. Yuan is with the Tsinghua Shenzhen International Graduate School and Peng Cheng Laboratory, Shenzhen, 518055, China. (e-mail: yuanc@sz.tsinghua.edu.cn).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi’an 710072, P. R. China. (e-mail: xuelong_li@nwpu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

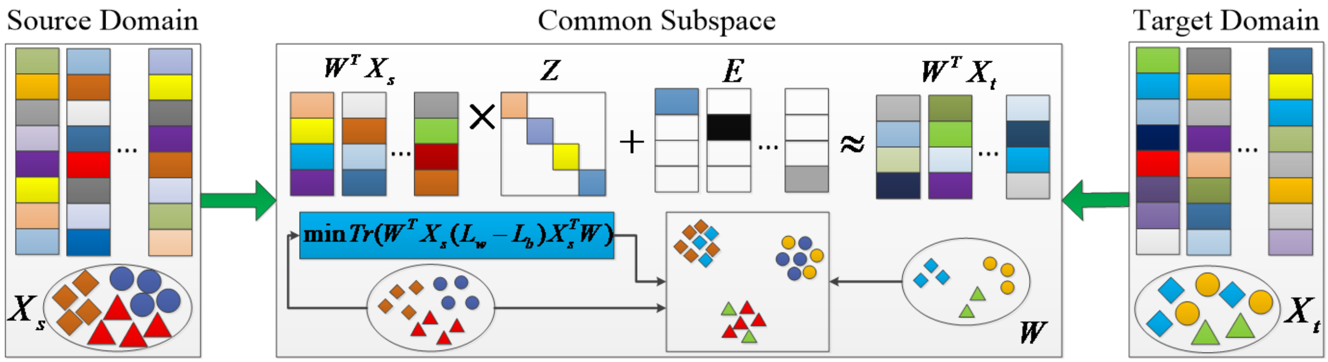


Fig. 1. Schematic of the proposed MTL-DRA method. MTL-DRA iteratively learns a latent subspace W that satisfies $W^T X_t = W^T X_s Z$, Z is a coefficient matrix. In the learned subspace, data samples from the same class are more compact and the data samples from different classes are far from each other as far as possible.

The above methods neglect the discriminant information of the data. Many supervised methods have thus been proposed [13]-[17]. To overcome the limitations in designing and comparing different tasks in self-supervised learning, Noroozi et al. [13] presented a framework for self-supervised learning by transferring knowledge. To improve the weakly supervised detection for the object categorization task, Li et al. [14] proposed a more reasonable and robust objectness transfer method for mixed supervised detection. Han et al. [15] proposed a sparse multi-label transfer learning method to learn a sparse linear embedding space from the source domain. To predict image memorability, Jing et al. [16] proposed a novel framework called multiview transfer learning from external sources (MTLES) to enhance representation ability of visual features. Zhang et al. [17] proposed an L_{21} norm-based discriminative kernel transfer learning (DKTL) method for the recognition task that can learn discriminative subspaces simultaneously using the domain-class-consistency metric and the representation of the robust transfer model.

However, the above-mentioned methods are sensitive to noise in the data and cannot directly divide the noise from the data. Recently, low-rank representation [18]-[20] has attracted much attention due to its global learning ability. Low-rank learning methods have global recovery ability and can remove noise directly [21]-[25]. For example, Jing et al. [21] proposed a transductive low-rank multi-view regression (TLRMVR) to boost the performance of micro-video popularity prediction. To better tackle micro-video multi-label classification tasks, Su et al. [22] proposed a low-rank regularized deep collaborative matrix factorization (LRDCMF) method. To enhance the robustness of preserving projection methods, Lu et al. [23] proposed a nuclear norm-based two-dimensional locality preserving projection (NN-2DLPP) for image classification. To predict the compatibility scores of fashion outfits, Jing et al. [24] proposed a transductive low-rank hypergraph regularizer multiple-representation learning (LHMRL) framework fashion compatibility prediction. To learn the discriminant low-rank representation and robust projecting subspace in a supervised manner, Li et al. [25] proposed a method that uses the least squares regularization.

To transfer the local geometry structure information and ensure that the transform matrix is robust or sparse, we propose a novel transfer learning method, named manifold transfer learning via discriminant regression analysis (MTL-DRA) for image classification, in which we encode discriminant information of the source domain to the target domain by introducing between- and within-class graphs. With different norms as constraints, MTL-DRA overcomes the disturbance of the noise and avoids the negative transfer. We also encode a nuclear norm constraint and propose robust MTL-DRA (RMTL-DRA) to improve the robustness of MTL-DRA. The convergence and complexity analysis of the two proposed methods are given. To verify their performance, we conducted extensive experiments on five public image benchmarks and compared the proposed methods to state-of-the-art transfer learning methods. Fig. 1 shows the schematic of MTL-DRA.

The main contributions of the paper are as follows.

1) The MTL-DRA method is proposed for image classification. By introducing within- and between-class graphs from the source domain, MTL-DRA ensures that data samples from the same class are more compact and the data samples from different classes are far from each other as far as possible.

2) MTL-DRA uses the L_{21} and L_1 norms as a sparse constraint for the learned projection matrix and the assumed noise matrix, respectively. In this way, MTL-DRA can join sparsity together and enhance the robustness of the algorithm. Furthermore, without the disturbance of noise, MTL-DRA can effectively avoid negative transfer.

3) Based on MTL-DRA, we use the nuclear norm instead of the L_{21} norm as a constraint of the projection matrix and propose a robust extension, robust MTL-DRA (RMTL-DRA).

4) The convergence proof and time complexity of both the proposed methods are provided in detail.

The rest of the paper is organized as follows. We present the motivation, objective function, optimization, convergence proof, and complexity analysis of the proposed MTL-DRA in Section II. The objective function, optimization, convergence, and complexity analysis of RMTL-DRA are given in Section III. Experiments and comparisons are reported in Section IV. Finally, the conclusion of the paper is given in Section V.

II. Manifold Transfer Learning Via Discriminant Regression Analysis

The motivation for proposing the manifold transfer learning via discriminant regression analysis (MTL-DRA) is first given.

Then, the formulation, optimization, solution, convergence, and complexity analysis are discussed.

A. Motivation

Traditional machine learning methods often demand that the data are IID, which in many applications cannot be ensured. As transfer learning methods can migrate the information in the source domain to the target domain for certain tasks, IID data are not required. Thus, transfer learning has attracted much attention. However, for noisy data, how to avoid the influence of the noise and migrate effective and discriminative information is crucial for transfer learning methods. Even though many discriminant transfer learning methods have been proposed [13]-[17], these methods do not consider both the local geometry structure and global data structure.

In this paper, to transfer the local geometry structure and local structure information of the data and ensure that the transform matrix is robust or sparse, we propose a novel transfer learning method, named manifold transfer learning via discriminant regression analysis (MTL-DRA) for image classification. By introducing between- and within-class graphs of the data in the source domain, we can transfer the discriminant information of the source domain to the target domain in MTL-DRA. To avoid transferring negative information to the target task, we use low-rank learning to remove the disturbance of the noise in the source domain data. To enhance the robustness of MTL-DRA, we use the nuclear norm instead of the L_{21} norm as a constraint of the transform matrix and propose robust MTL-DRA (RMTL-DRA). The convergence and complexity of the two proposed methods are analyzed in detail.

B. Formulation of the Problem

In transfer learning, how to transfer useful knowledge from the source domain to the target domain is vital. Most transfer subspace methods neglect manifold discriminant information of the data, which can greatly improve the performance in image classification.

Given the source domain data $X_s=[x_s^1,\dots,x_s^{n_s}]\in R^{m\times n_s}$, where x_s^i ($i=1,\dots n_s$) are vector samples from the source domain, their corresponding label matrix $Y=[y_1,\dots,y_{n_s}]\in R^{c\times n_s}$, and the target domain data $X_t=[x_t^1,\dots,x_t^{n_t}]\in R^{m\times n_t}$, where x_t^i ($i=1,\dots n_t$) are vector samples from the target domain; m is the dimension of each sample. Assume that the target data can be reconstructed by the source data in a common subspace, that is:

$$\min_{W,Z} \| W^T X_t - W^T X_s Z \|_F^2 , \tag{1}$$

where $W\in R^{m\times c}$ denotes a transformation matrix and $Z\in R^{n_s\times n_t}$ is a reconstruction matrix.

To ensure that the reconstruction coefficient matrix Z has a block-wise structure, we rewrite (1) as

$$\min_{W,Z} rank(Z), \quad \text{s.t. } W^T X_t = W^T X_s Z . \tag{2}$$

As the rank minimization problem in (2) is NP-hard, we relax it as

$$\min_{W,Z} \|Z\|_*, \quad \text{s.t. } W^T X_t = W^T X_s Z . \tag{3}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix.

To alleviate the disturbance of the noise in data, we introduce a noise matrix E with a sparse constraint in (3), and have

$$\min_{W,Z,E} \|Z\|_* + \alpha\|E\|_1, \quad \text{s.t. } W^T X_t = W^T X_s Z + E . \tag{4}$$

where α is a parameter and $\|\cdot\|_1$ denotes the L_1 norm of a matrix.

To encode manifold information of the source data, we introduce the within class graph and between class graph as defined in the graph embedding formulation [26]. The definitions of the within class graph and between class graph are as follows:

$$L_w = D_w - S_w, \; L_b = D_b - S_b , \tag{5}$$

where S_w and S_b are defined as follows:

$$S_w = \sum_i \sum_{i\in N_{k_1}^+(j) \text{ or } j\in N_{k_1}^+(i)} \left\| w^T x_i - w^T x_j \right\|^2 \tag{6}$$

and

$$S_b = \sum_i \sum_{(i,j)\in P_{k_2}(c_i) \text{ or } (i,j)\in P_{k_2}(c_j)} \left\| w^T x_i - w^T x_j \right\| \tag{7}$$

where $N_{k_1}^+(i)$ indicates the index set of the k_1 nearest neighbors of the sample x_i in the same class and $P_{k_2}(c)$ is a set of data pairs that are the k_2 nearest pairs among the set $\{(i,j), i\in\pi_c, j\notin\pi_c\}$, where π_c denotes the index set. D_w is the diagonal matrix whose entries are column sums of S_w, and D_b is the diagonal matrix whose entries are column sums of S_b. To maintain the geometric structure of the data, we can combine the intrinsic graph and the penalty graph by introducing the following regularization term,

$$Tr(W^T X_s (L_w - L_b) X_s^T W) , \tag{8}$$

where $Tr(\cdot)$ denotes the trace of a matrix.

Combining (4) and (8), we obtain

$$\min_{W,Z,E} \|Z\|_* + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W),$$
$$\text{s.t. } W^T X_t = W^T X_s Z + E , \tag{9}$$

where α and β are parameters.

To enhance the sparsity among rows of the transform matrix W in (9), we use the L_{21} norm as a sparse constraint of W to obtain

$$\min_{W,Z,E} \|Z\|_* + \|W\|_{21} + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W),$$
$$\text{s.t. } W^T X_t = W^T X_s Z + E . \tag{10}$$

As a result, the objective function of MTL-DRA is defined as:

$$\min_{W,Z,E} \frac{1}{2}\phi(W,Y,X_s) + \|Z\|_* + \|W\|_{21} + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W),$$
$$\text{s.t. } W^T X_t = W^T X_s Z + E , \tag{11}$$

where $\phi(W,Y,X_s)$ is a regression discriminant subspace learning function. The definition of $\phi(W,Y,X_s)$ is

$$\phi(W,Y,X_s) = \|W^T X_s - Y - B\odot M\|_F^2, \text{ s.t. } M\geq 0 . \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

The definition of B is given in (13), where B_{ij} is the i th row and j th column element.

$$B_{ij} = \begin{cases} +1, & y_i = j \\ -1, & otherwise \end{cases}. \quad (13)$$

The first term in (11) performs regression learning to learn a subspace in which the distance between different classes is as large as possible. Thus, the discriminative ability is enhanced. With different norms as constraints, the three terms in MTL-DRA can transfer the sparsity and global learning ability to the target task and avoid negative transfer. The last term in (11) encodes a within-class graph and a between-class graph, which preserves the similarity of the data from the same class and suppresses the similarity of the data from different classes. By learning the transform matrix, MTL-DRL can transform the local geometry structure and global structure in the source domain to the target domain, ensure a compatible representation, and reduce discrepancies in the data.

C. Optimization

Since the optimization problem (11) is non convex, we solve it by fixing other variables to update one variable. We can convert (11) to

$$\begin{aligned} &\min_{W,Z,Z_1,E,M} \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \|Z_1\|_* + \|W\|_{21} \\ &+\alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W) \end{aligned},$$

$$\text{s.t. } W^T X_t = W^T X_s Z + E,\ Z = Z_1,\ M \ge 0. \quad (14)$$

To solve (14), we introduce a property of the nuclear norm by the following Lemma.

Lemma 1 [27] For any matrix $Z_1 \in R^{n_s \times n_t}$, the following holds:

$$\| Z_1 \|_* = \min_{A,D} \frac{1}{2}(\| A \|_F^2 + \| D \|_F^2),\ \text{ s.t. } Z_1 = AD. \quad (15)$$

If the rank of Z_1 is $r \le \min(n_s, n_t)$, then the minimum solution above is attained as a factor decomposition $Z_1 = A_{n_s \times r} D_{r \times n_t}$.

Based on Lemma 1, an equivalent representation of (14) is:

$$\begin{aligned} &\min_{W,Z,Z_1,A,D,E,M} \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \frac{1}{2}(\| A \|_F^2 + \| D \|_F^2) \\ &+\|W\|_{21} + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W) \end{aligned},$$

$$\text{s.t. } W^T X_t = W^T X_s Z + E,\ Z = Z_1,\ Z_1 = AD,\ M \ge 0. \quad (16)$$

We solve (16) by minimizing the following augmented Lagrange multiplier function

$$\begin{aligned} L &= \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \frac{1}{2}(\| A \|_F^2 + \| D \|_F^2) + \|W\|_{21} + \alpha\|E\|_1 \\ &+\beta Tr(W^T X_s (L_w - L_b) X_s^T W) + Tr(M_1^T (W^T X_t - W^T X_s Z - E)) \\ &+Tr(M_2^T (Z - Z_1)) + Tr(M_3^T (Z_1 - AD)) + \frac{\mu}{2}(\| W^T X_t - W^T X_s Z - E \|_F^2 \\ &+ \| Z - Z_1 \|_F^2 + \| Z_1 - AD \|_F^2) \end{aligned},$$

$$\text{s.t. } M \ge 0, \quad (17)$$

where M_1, M_2, and M_3 are Lagrange multipliers and $\mu > 0$ is a parameter. The main steps for solving (17) are as follows.

Step 1 (Update W): Fixed variables other than W, we have

$$\begin{aligned} &\min_W \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W) \\ &+\|W\|_{21} + \frac{\mu}{2}\| W^T X_t - W^T X_s Z - E + \frac{M_1}{\mu} \|_F^2 \end{aligned}. \quad (18)$$

Let $U = Y + B\odot M$, $R = X_s (L_w - L_b) X_s^T$, $Q = X_t - X_s Z$ and $Q_1 = E - \frac{M_1}{\mu}$. Setting to zero the partial derivative of (18) with respect to W gives

$$X_s X_s^T W - X_s U^T + 2\beta R^T W + \mu Q Q^T W - \mu Q Q_1^T + GW = 0, \quad (19)$$

where $G_{ii} = 1/\| W_{i,:} \|_2$, and $W_{i,:}$ denotes the i-th row of the matrix W.

From (19), we have

$$W = (X_s X_s^T + 2\beta R^T + \mu Q Q^T + G)^{-1} (X_s U^T + \mu Q Q_1^T). \quad (20)$$

Step 2 (Update M): Fixed other variables than M in (17), we have

$$\min_{M \ge 0} \left\| V - B\odot M \right\|_F^2, \quad (21)$$

where $V = W^T X_s - Y$.

Considering the i th row and j th column element M_{ij} of M, we have

$$\min_{M_{ij} \ge 0} (V_{ij} - B_{ij} M_{ij})^2. \quad (22)$$

The optimal solution of M_{ij} in (22) is referred as discussed in [28], that is:

$$M_{ij} = \max(B_{ij} V_{ij}, 0). \quad (23)$$

Thus, the optimal solution of (20) with respect to M is

$$M = \max(B\odot V, 0). \quad (24)$$

Step 3 (Update Z): Fixed other variables except Z in (17), we have

$$\min_Z \| W^T X_t - W^T X_s Z - E + \frac{M_1}{\mu} \|_F^2 + \| Z - Z_1 + \frac{M_2}{\mu} \|_F^2. \quad (25)$$

Let $Q_2 = W^T X_t - E + \frac{M_1}{\mu}$, by setting the partial derivative of (25) equal to zero, we have

$$\begin{aligned} &X_s^T W W^T X_s Z - X_s^T W Q_2 + Z - Z_1 + \frac{M_2}{\mu} = 0 \\ &\Rightarrow Z = (X_s^T W W^T X_s + I)^{-1} (X_s^T W Q_2 + Z_1 - \frac{M_2}{\mu}) \end{aligned}. \quad (26)$$

Step 4 (Update Z_1): Fixed other variables except Z_1 in (17), we have

$$\min_{Z_1} \| Z_1 - Z - \frac{M_2}{\mu} \|_F^2 + \| Z_1 - AD + \frac{M_3}{\mu} \|_F^2. \quad (27)$$

By setting the partial derivative of (27) equal to zero, we have

$$2Z_1 - Z - \frac{M_2}{\mu} - AD + \frac{M_3}{\mu} = 0 \Rightarrow Z_1 = \frac{1}{2}(Z + AD + \frac{M_2 - M_3}{\mu}). \quad (28)$$

Step 5 (Update A): Fixed other variables except A, we have

$$\min_A \| A \|_F^2 + \mu \| Z_1 - AD + \frac{M_3}{\mu} \|_F^2. \quad (29)$$

By setting the partial derivative of (29) equal to zero, we have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

$$\begin{aligned}&A+\mu ADD^T-(\mu Z_1+M_3)D^T=0\\&\Rightarrow A=(\mu Z_1+M_3)D^T(I+\mu DD^T)^{-1}\end{aligned}. \tag{30}$$

Step 6 (Update D): Fixed other variables except D, we have

$$\min_D \|D\|_F^2+\mu\|Z_1-AD+\frac{M_3}{\mu}\|_F^2. \tag{31}$$

By setting the partial derivative of (31) equal to zero, we have

$$\begin{aligned}&D+\mu A^TAD-A^T(\mu Z_1+M_3)=0\\&\Rightarrow D=(I+\mu A^TA)^{-1}A^T(\mu Z_1+M_3)\end{aligned}. \tag{32}$$

Step 7 (Update E): Fixed other variables except E in (17), we have

$$\min_E \alpha\|E\|_1+\frac{\mu}{2}\|E-W^TX_t+W^TX_sZ-\frac{M_1}{\mu}\|_F^2. \tag{33}$$

We solve (33) by using the soft thresholding operator $S_\varepsilon[x]=sign(x)\cdot\max(|x|-\varepsilon,0)$ [29]. The closed-form solution of (33) is as follows:

$$E=S_{\frac{\alpha}{\mu}}(W^TX_t-W^TX_sZ+\frac{M_1}{\mu}). \tag{34}$$

Step 8 Multipliers M_1,M_2,M_3 and iteration step-size ρ ($\rho>1$) are updated by

$$\begin{cases}M_1=M_1+\mu(W^TX_t-W^TX_sZ-E)\\M_2=M_2+\mu(Z-Z_1)\\M_3=M_3+\mu(Z_1-AD)\\\mu=\min(\rho\mu,\mu_{\max})\end{cases}. \tag{35}$$

The steps of MTL-DRA are given in Algorithm 1.

Algorithm 1 MTL-DRA

Input: X_s, X_t, Y, B, and parameters α,β in (11);

Initialization: $M=1$; $Z=Z_1=0$; $E=0$, , $M_1=M_2=M_3=0$; $\rho>0$.

repeat

1. Update W by (20);
2. Update M by (24);
3. Update Z by (26);
4. Update Z_1 by (28);
5. Update A by (30);
6. Update D by (32);
7. Update E by (34);
8. Update Lagrange multipliers as follows:
   $M_1=M_1+\mu(W^TX_t-W^TX_sZ-E)$;
   $M_2=M_2+\mu(Z-Z_1)$;
   $M_3=M_3+\mu(Z_1-AD)$.
9. Update μ by $\mu=\min(\rho\mu,\max\mu)$.
10. Update $t=t+1$
11. Obtain the optimal solution (W,Z,E)

Output: W,Z,E

D. Convergence and Complexity Analysis

In this section, we analyze the convergence property of MTL-DRA. By showing that under mild conditions any limited point of the iteration sequence generated by MTL-DRA is a stationary point which satisfies the Karush-Kuhn-Tucker (KKT) conditions [30], we give a weak convergence proof of MTL-DRA. The necessary condition for a local optimal solution is that any converging point must satisfy the KKT conditions to ensure the behavior of MTL-DRA. The convergence proof of MTL-DRA is given in the Appendix.

The major computations in Algorithm 1 are in steps 1, 3, 5, and 6. Steps 1 and 3 have the same complexity, which is at most $O(n_s^3)$. Steps 5 and 6 have the same computational complexity, which is at most $O(r^3)$. Because $r\le\min(n_s,n_t)$, the complexity of Algorithm 1 is at most $O(tn_s^3)$, where t is the number of iterations.

III. Robust Manifold Transfer Learning Via Discriminant Regression Analysis

The collected data from different domains may be corrupted by noise in real-world applications. To enhance the robustness of MTL-DRA, we use the nuclear norm as a constraint to the projective matrix in MTL-DRA and propose robust MTL-DRA (RMTL-DRA). In the next sub-sections, we describe the RMTL-DRA method.

A. Formulation of the Problem

Due to their global learning ability, nuclear norm-based methods have been applied in many fields, such as pattern recognition [32], [33] and computer vision [34], [35].

MTL-DRA maps the source domain data and the target domain data to a common subspace. However, if the data are corrupted by noise, then the learning process of the projective matrix is disturbed, and the performance of MTL-DRA is degraded. To avoid this shortcoming and to enhance the robustness of MTL-DRA, we use the nuclear norm instead of the L_{21} norm as a constraint of the transform matrix, and we propose robust manifold transfer learning via discriminant regression analysis (RMTL-DRA). The optimization objective function of RMTL-DRA is defined as:

$$\begin{aligned}&\min_{W,Z,E,M}\frac{1}{2}\|W^TX_s-Y-B\odot M\|_F^2+\|W\|_*+\|Z\|_*\\&+\alpha\|E\|_1+\beta Tr(W^TX_s(L_w-L_b)X_s^TW)\\&\text{s.t. } W^TX_t=W^TX_sZ+E,\ M\ge 0.\end{aligned} \tag{36}$$

where α and β are parameters.

Similar to MTL-DRA, RMTL-DRA aims to learn a compatible representation and reduce discrepancies in the data. If the second regularization term (i.e., $\|W\|_{21}$ in (11) and $\|W\|_*$ in (36)) in both the objective function of MTL-DRA and RMTL-DRA is removed, we can see that MTL-DRA and RMTL-DRA share the same framework. The main difference between MTL-DRA and RMTL-DRA is that: MTL-DRA uses the L_{21} norm as a constraint of the projective matrix W, while RMTL-DRA uses the nuclear norm instead of the L_{21} norm.

B. Optimization and Solutions

To solve the optimization problem (36), we introduce auxiliary variables and obtain

$$\begin{aligned}&\min_{W,W_1,M,Z,Z_1,E}\frac{1}{2}\|W^TX_s-Y-B\odot M\|_F^2+\|Z_1\|_*+\|W_1\|_*\\&+\alpha\|E\|_1+\beta Tr(W^TX_s(L_w-L_b)X_s^TW)\end{aligned},$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

$$\text{s.t. } W^T X_t = W^T X_s Z + E,\ Z = Z_1,\ W = W_1,\ M \ge 0. \quad (37)$$

According to Lemma 1, we can rewrite (37) as

$$\min_{W,W_1,Z,Z_1,A,D,\tilde{A},\tilde{D},E,M} \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \frac{1}{2}(\|A\|_F^2 + \|D\|_F^2) + \frac{1}{2}(\|\tilde{A}\|_F^2 + \|\tilde{D}\|_F^2) + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W),$$

$$\text{s.t. } W^T X_t = W^T X_s Z + E,\ Z = Z_1,\ Z_1 = AD,\ W = W_1,\ W_1 = \tilde{A}\tilde{D},\ M \ge 0. \quad (38)$$

We solve (38) by minimizing the following augmented Lagrange multiplier function

$$\begin{aligned} L = &\frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \frac{1}{2}(\|A\|_F^2 + \|D\|_F^2) \\ &+\frac{1}{2}(\|\tilde{A}\|_F^2 + \|\tilde{D}\|_F^2) + \alpha\|E\|_1 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W) + \\ &Tr(M_1^T(W^T X_t - W^T X_s Z - E)) + Tr(M_2^T(Z - Z_1)) + Tr(M_3^T(Z_1 - AD)), \\ &+Tr(M_4^T(W - W_1)) + Tr(M_5^T(W_1 - \tilde{A}\tilde{D})) + \frac{\mu}{2}(\|W^T X_t - W^T X_s Z - E\|_F^2 \\ &+\|Z - Z_1\|_F^2 + \|Z_1 - AD\|_F^2 + \|W - W_1\|_F^2 + \|W_1 - \tilde{A}\tilde{D}\|_F^2) \\ &\text{s.t. } M \ge 0. \end{aligned} \quad (39)$$

where M_1, M_2, M_3, M_4 and M_5 are Lagrange multipliers and $\mu > 0$ is a parameter. The main steps in solving (39) are as follows.

Step 1 (Update W): Fixed other variables except W in (39), we have

$$\min_W \frac{1}{2}\|W^T X_s - Y - B\odot M\|_F^2 + \beta Tr(W^T X_s (L_w - L_b) X_s^T W) + \frac{\mu}{2}(\|W^T X_t - W^T X_s Z - E + \frac{M_1}{\mu}\|_F^2 + \|W - W_1 + \frac{M_4}{\mu}\|_F^2). \quad (40)$$

Let $Q_3 = W_1 - \frac{M_4}{\mu}$, setting the partial derivative of (40) with respect to W equal to zero gives

$$X_s(X_s^T W - U^T) + 2\beta R^T W + \mu Q Q^T W - \mu Q Q_1^T + \mu W - \mu Q_3 = 0. \quad (41)$$

Thus, we have

$$W = (X_s X_s^T + 2\beta R^T + \mu Q Q^T + \mu I)^{-1}(X_s U^T + \mu Q Q_1^T + \mu Q_3). \quad (42)$$

Step 2 (Update M): Fixed other variables except M in (39), we have

$$\min_{M\ge 0}\|V - B\odot M\|_F^2. \quad (43)$$

The solution of (43) is the same as that of (24).

Step 3 (Update W_1): Fixed other variables except W_1 in (39), we have

$$\min_{W_1}\|W_1 - W - \frac{M_4}{\mu}\|_F^2 + \|W_1 - \tilde{A}\tilde{D} + \frac{M_5}{\mu}\|_F^2. \quad (44)$$

Setting the partial derivative of W_1 in (44)) equal to zero, we have

$$2W_1 - W - \frac{M_4}{\mu} - \tilde{A}\tilde{D} + \frac{M_5}{\mu} = 0 \Rightarrow W_1 = \frac{1}{2}(W + \frac{M_4}{\mu} + \tilde{A}\tilde{D} - \frac{M_5}{\mu}). \quad (45)$$

Step 4 (Update Z): Fixed other variables except Z in (39), we have

---

Algorithm 2 RMTL-DRA

Input: X_s, X_t, Y, B, and parameters α, β in (37);

Initialization: $M = 1$; $Z = Z_1 = 0$; $E = 0$, , $M_1 = M_2 = M_3 = M_4 = M_5 = 0$; $\rho > 0$.

repeat

1. Update W by (42);
2. Update M by (24);
3. Update W_1 by (45);
4. Update Z by (26);
5. Update Z_1 by (28);
6. Update A by (30);
7. Update D by (32);
8. Update $\tilde{A}$ by (51);
9. Update $\tilde{D}$ by (53);
10. Update E by (34);
11. Update Lagrange multipliers as follows:
    $M_1 = M_1 + \mu(W^T X_t - W^T X_s Z - E)$;
    $M_2 = M_2 + \mu(Z - Z_1)$;
    $M_3 = M_3 + \mu(Z_1 - AD)$;
    $M_4 = M_4 + \mu(W - W_1)$;
    $M_5 = M_5 + \mu(W_1 - \tilde{A}\tilde{D})$.
12. Update μ by $\mu = \min(\rho\mu, \max\mu)$.
13. Update $t = t + 1$
14. Obtain the optimal solution (W, Z, E)

Output: W, Z, E

$$\min_Z \|W^T X_t - W^T X_s Z - E + \frac{M_1}{\mu}\|_F^2 + \|Z - Z_1 + \frac{M_2}{\mu}\|_F^2. \quad (46)$$

The solution of (46) is the same as that of (26).

Step 5 (Update Z_1): Fixed other variables except Z_1 in (39), we have

$$\min_{Z_1}\|Z_1 - Z - \frac{M_2}{\mu}\|_F^2 + \|Z_1 - AD + \frac{M_3}{\mu}\|_F^2. \quad (47)$$

The solution of (47) is (28).

Step 6 (Update A): Fixed other variables except A in (39), we have

$$\min_A \|A\|_F^2 + \mu\|Z_1 - AD + \frac{M_3}{\mu}\|_F^2. \quad (48)$$

The solution of (48) is (30).

Step 7 (Update D): Fixed other variables except D in (39), we have

$$\min_D \|D\|_F^2 + \mu\|Z_1 - AD + \frac{M_3}{\mu}\|_F^2. \quad (49)$$

The solution of (49) is (32).

Step 8 (Update $\tilde{A}$): Fixed other variables except $\tilde{A}$, we have

$$\min_{\tilde{A}} \|\tilde{A}\|_F^2 + \mu\|W_1 - \tilde{A}\tilde{D} + \frac{M_5}{\mu}\|_F^2. \quad (50)$$

The solution of (50) is the same as (29). Thus, we have

$$\tilde{A} = (\mu W_1 + M_5)\tilde{D}^T(I + \mu\tilde{D}\tilde{D}^T)^{-1}. \quad (51)$$

Step 9 (Update $\tilde{D}$): Fixed other variables except $\tilde{D}$, we have

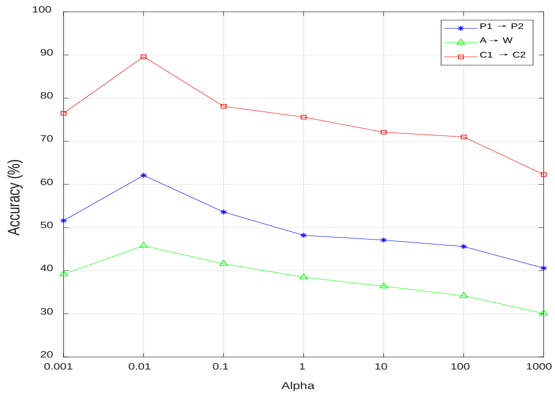
$$\min_{\tilde{D}} \|\tilde{D}\|_F^2 + \mu\|W_1 - \tilde{A}\tilde{D} + \frac{M_5}{\mu}\|_F^2. \quad (52)$$

The solution of (52) is the same as that of (31). Thus, we have

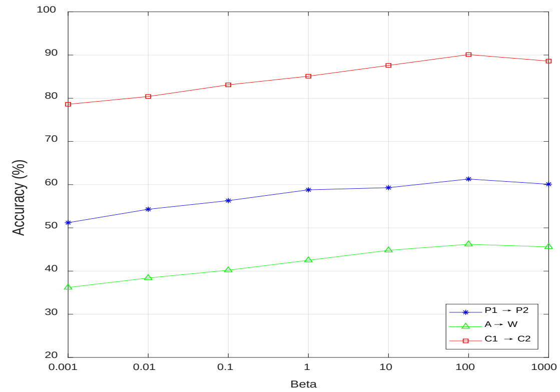
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

TABLE I. TIME CONSUMING (IN SECOND) OF OUR METHODS ON DIFFERENT EVALUATIONS.

Evaluations	$A\rightarrow W$	$M\rightarrow U$	$C1\rightarrow C2$	$P1\rightarrow P4$	$Ar\rightarrow Cl$
MTL-DRA	5.32	10.52	87.64	23.15	53.16
RMTL-DRA	6.15	12.61	95.12	27.61	62.42



(a)



(b)

Fig. 3. Classification accuracy versus (a) α with β fixed and (b) β with α fixed on the P1→P2, A→W, and C1→C2 benchmarks.

$$\tilde{D} = (I + \mu \tilde{A}^T \tilde{A})^{-1} \tilde{A}^T (\mu W_1 + M_5) . \tag{53}$$

Step 10 (Update E): Fixed other variables except E in (39), we have

$$\min_E \alpha \|E\|_1 + \frac{\mu}{2} \| E - W^T X_t + W^T X_s Z - \frac{M_1}{\mu} \|_F^2 . \tag{54}$$

The solution of (54) is (34).

Step 11 Multipliers M_1, M_2, M_3, M_4, M_5 and iteration step-size ρ ($\rho > 1$) are updated by

$$\begin{cases} M_1 = M_1 + \mu(W^T X_t - W^T X_s Z - E) \\ M_2 = M_2 + \mu(Z - Z_1) \\ M_3 = M_3 + \mu(Z_1 - AD) \\ M_4 = M_4 + \mu(W - W_1) \\ M_5 = M_5 + \mu(W_1 - \tilde{A}\tilde{D}) \\ \mu = \min(\rho\mu, \mu_{\max}) \end{cases} . \tag{55}$$

The concrete steps of RMTL-DRA are given in Algorithm 2.

C. Convergence and Complexity Analysis

The convergence of RMTL-DRA is similar to that of MTL-DRA. Due to space limitations, we do not provide the convergence proof for RMTL-DRA.

The major computations in Algorithm 2 are in steps 1 and 4. These have the same complexity, which is at most $O(n_s^3)$. Thus, the complexity of Algorithm 2 is $O(tn_s^3)$, where t is the number of iterations. Table I shows the actual runtime of our methods on different evaluations.

IV. Experiments

To verify the performance of our proposed methods, we comprehensively compare our methods with state-of-the-art transfer learning methods on five visual benchmark databases including CMU PIE [36], COIL100 [37], MNIST [38]+USPS [39], Office [40] + Caltech-256 [41], and Office + Home [50]. Fig. 2 shows some images of these databases.

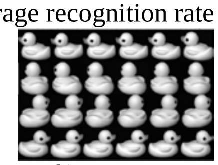
A. Baselines and Setting

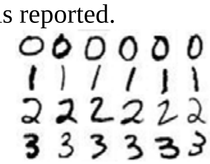
In order to evaluate the superiority of our methods, we compare our methods with the following baselines: nearest neighbor (NN), principle component analysis (PCA) [42], transfer component analysis (TCA) [43], transfer subspace learning (TSL) [44], latent sparse domain transfer (LSDT) [45], joint distribution adaptation (JDA) [46], maximum independence domain adaptation (MIDA) [27], stacked robust adaptively regularized auto-regression (SRARAs) [47], and low-rank constrained latent domain adaptation depression recognition (LDADR) [48].

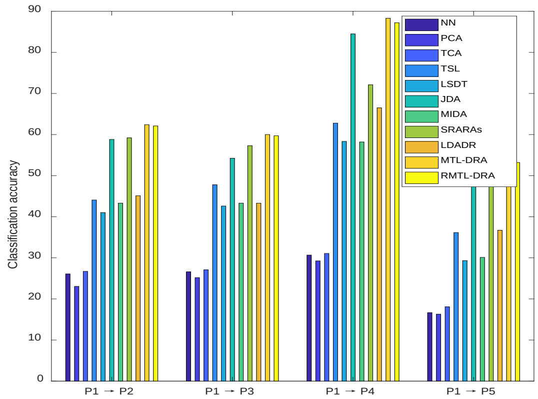
As there are no parameters in NN and PCA, we need not set any values. For other compared methods, we set the parameters as described in their related works. For classification, we used 1-nearest neighbor classifier (NN) to classify the transformation results of the target domain data. The two parameters in both MTL-DRA and RMTL-DRA need to be selected. The two parameters in both methods are selected in {0.001, 0.01, 0.1, 1, 10, 100, 1000}. We evaluate the sensitivity of the parameters α and β in Fig. 3, which shows the classification accuracy versus Fig. 3. (a) α with β fixed and Fig. 3. (b) β with α fixed on three evaluations, i.e., $P1\rightarrow P2$, $A\rightarrow W$, and $C1\rightarrow C2$.

In order to verify the classification performance of the proposed methods, we conducted extensive experiments on three recognition tasks, including face recognition, object classification, and handwritten digit recognition. The CMU PIE database was used to test the performance on face recognition and the COIL 100, Office, and Caltech-256 were used to test the performance on object classification. The USPS and MNIST databases were used to test the performance on handwritten digit recognition. Each experiment was repeated 20 times and the average recognition rate is reported.

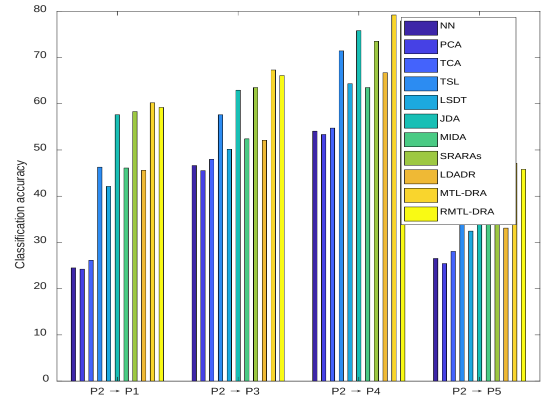

(a) CMU PIE


(b) COIL100

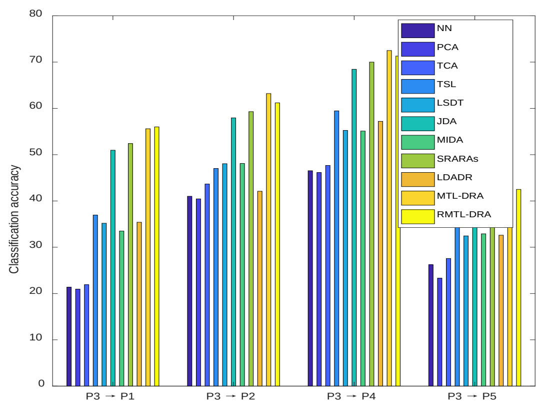

(c) MNIST



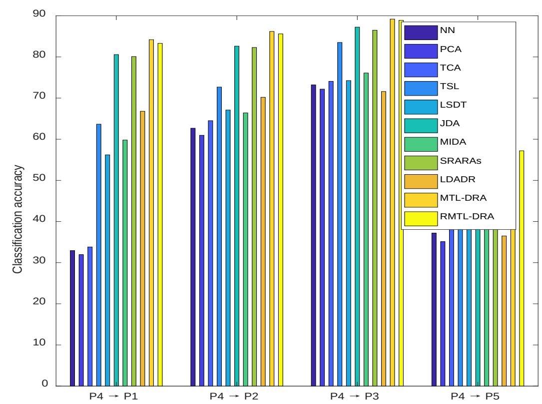
(a)



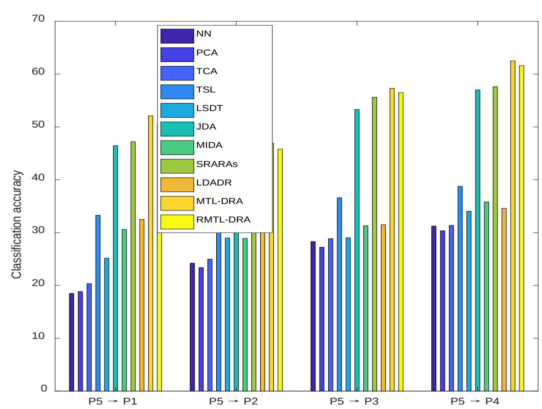
(b)



(c)

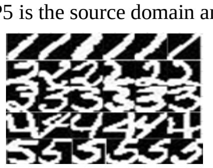


(d)



(e)

Fig. 4. The classification accuracies of NN, PCA, TCA, TSL, LSDT, JDA, MIDA, SRARAs, LDADR, MTL-DRA, and RMTL-DRA on the CMU PIE database. (a) P1 is the source domain and P2-P5 are the target domain, respectively, (b) P2 is the source domain and P1, P3-P5 are the target domain, respectively, (c) P3 is the source domain and P1, P2, P4, and P5 are the target domain, respectively, (d) P4 is the source domain and P1-P3, and P5 are the target domain, respectively, (e) P5 is the source domain and P1-P4 are the target domain, respectively.



(d) USPS

(e) Office

(f) Caltech-256

Fig. 2. Some image examples of the databases.

B. *Experiments on the CMU PIE Database*

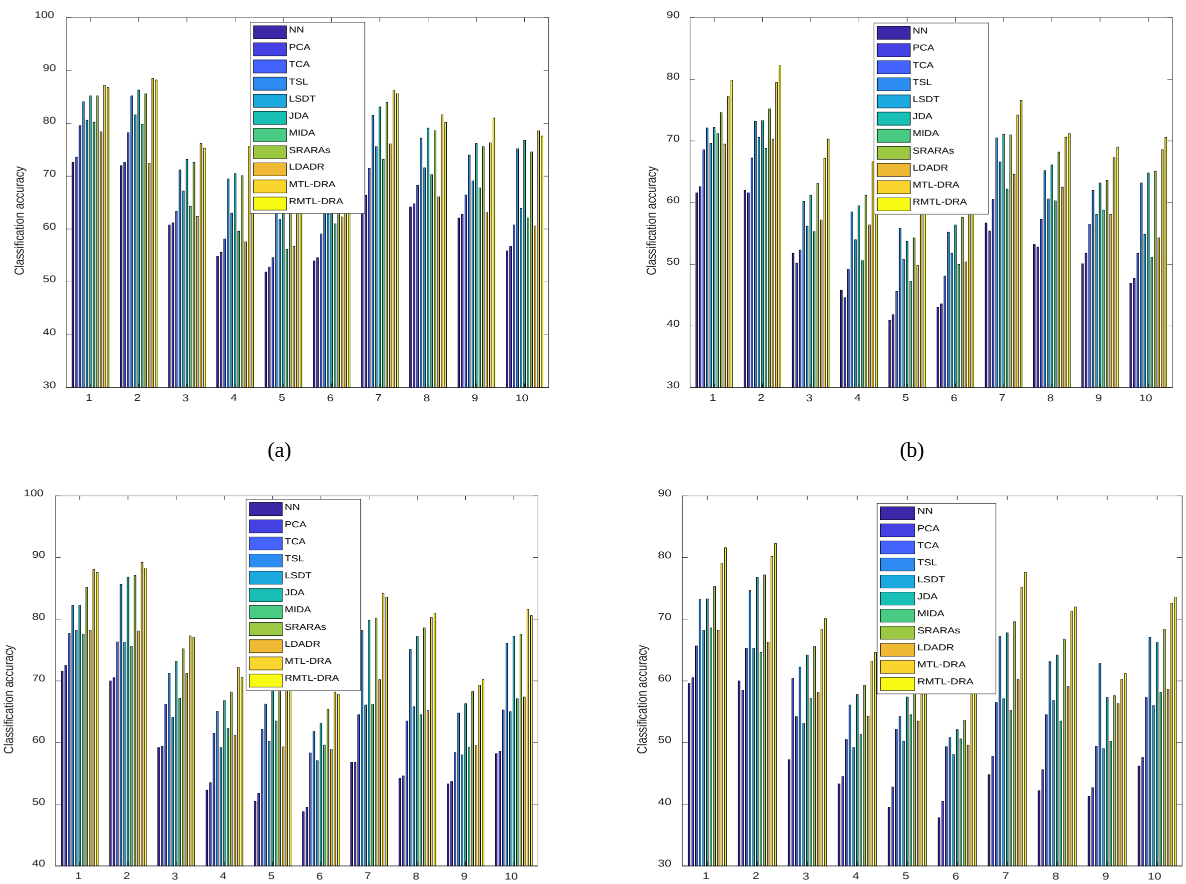
There are 41,368 face images from 68 subjects in the CMU PIE database [36]. All images were captured by 13 synchronized cameras and 21 flashes with “pose”, “illumination”, and “expression” changes. The size of all images is 32×32. In our experiment, we used five pose subsets of PIE to test the performance of each method. The five subsets are constructed by PIE1 (C05, left pose), PIE2 (C07, upward

TABLE II. THE EXPERIMENTAL RESULTS OF ALL THE COMPARED METHODS ON THE COIL100 DATABASE.

Database	NN	PCA	TCA	TSL	LSDT	JDA	MIDA	SRARAs	LDADR	MTL-DRA	RMTL-DRA
$C1\rightarrow C2$	82.6	83.2	86.2	85.2	82.0	86.2	84.3	87.5	84.3	90.1	89.3
$C2\rightarrow C1$	79.8	81.1	84.3	83.1	79.8	85.1	84.6	85.2	81.6	89.2	88.4
Average	81.2	82.2	85.3	84.2	80.9	85.7	84.5	86.4	83.0	89.7	88.9

TABLE III. THE EXPERIMENTAL RESULTS OF ALL THE COMPARED METHODS ON THE COIL100 DATABASE WITH OCCLUSIONS

Database	NN	PCA	TCA	TSL	LSDT	JDA	MIDA	SRARAs	LDADR	MTL-DRA	RMTL-DRA
$C1\rightarrow C2$	71.6	64.2	76.3	76.2	72.2	77.3	75.6	78.1	68.3	79.1	81.2
$C2\rightarrow C1$	69.5	62.1	73.3	74.2	71.1	75.4	74.5	76.5	67.5	78.2	80.3
Average	70.6	63.2	74.8	75.2	71.7	76.4	75.1	77.3	67.9	78.7	80.8



(a) (b)

(c) (d)

Fig. 5 The classification accuracies of NN, PCA, TCA, TSL, LSDT, JDA, MIDA, SRARAs, LDADR, MTL-DRA, and RMTL-DRA on the MNIST and USPS databases. (a) the MNIST databases is the source domain and the USPS database is the target domain (original images); b) the MNIST databases is the source domain and the USPS database is the target domain (all images with 10% pixel corruption); (c) the USPS databases is the source domain and the MNIST database is the target domain (original images); (d) the USPS databases is the source domain and the MNIST database is the target domain (all images with 10% pixel corruption).

pose), PIE3 (C09, downward pose), PIE4 (C27, front pose), PIE5 (C29, right pose) and the face images of each subset with different expression and illumination changes.

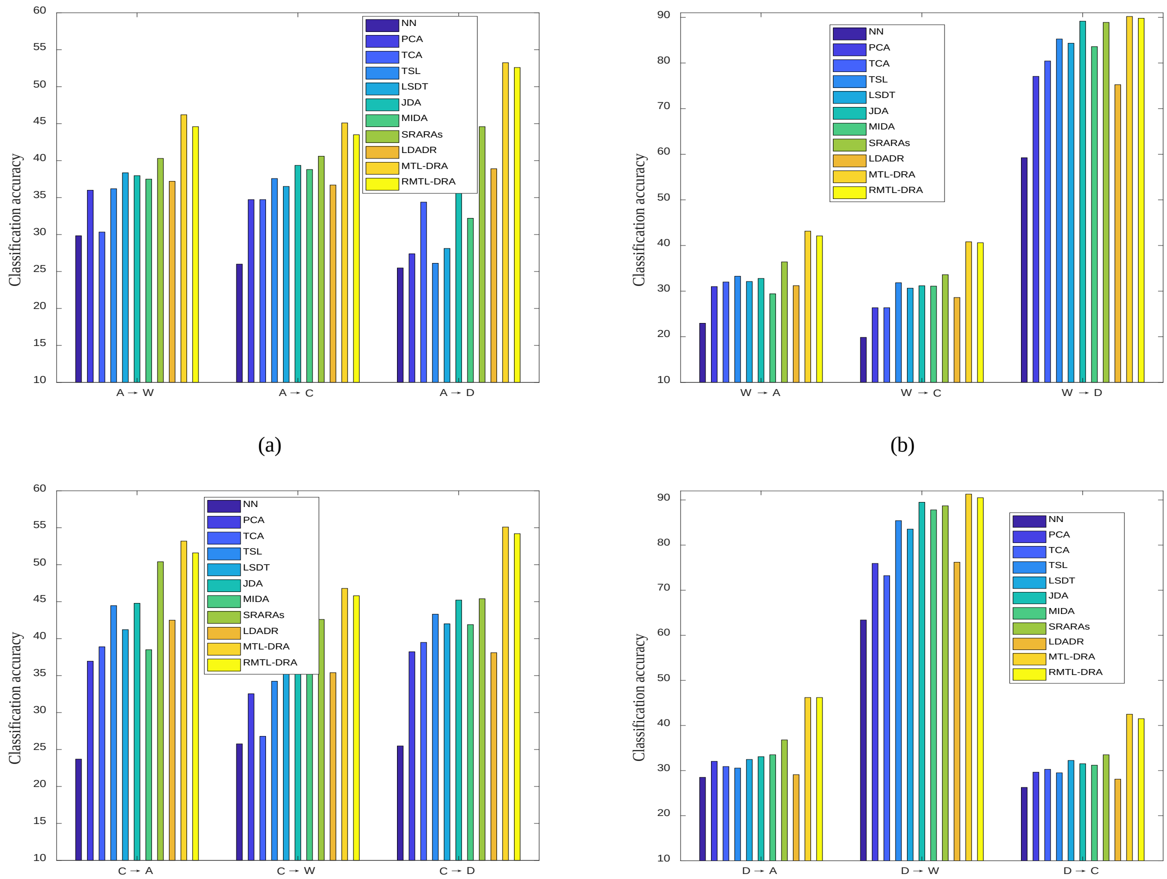
In the five pose subsets, two different subsets were randomly selected as the source data and the target data, respectively. Thus, there were 20 cross-domain datasets: PIE1 (P1) vs PIE2 (P2), PIE1 (P1) vs PIE3 (P3), PIE1 (P1) vs PIE4 (P4), PIE1 (P1) vs PIE5 (P5), $\cdots$, and PIE5 (P5) vs PIE4 (P4). In this way, the distributions of each cross-domain were significantly different. The experimental results with different cross domains are reported in Fig. 4.

From Fig. 4, we can see that MTL-DRA achieves the best adaptation performance on the CMU PIE database, which shows that transfer discriminant information and geometry structure information are important for transfer learning.

C. Experiments on MNIST and USPS Databases

In this section, we used the MNIST and USPS handwritten digit databases to conduct experiments to verify the performance of the proposed methods. There are 60,000 training images and 10,000 test images in the MNIST database [38], and the size of all images is 28×28. The USPS database [39] contains 9,298 labeled image, and the size of all images is 16×16. There are 10 digit categories in both databases. We used each digit class of the two databases to construct two recognition tasks. For the first task, the USPS database was used as the source domain and the MNIST database as the target domain, i.e., $U\rightarrow M$. Similarly, for the second task, the MNIST database was used as the source domain and the USPS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia



(c) (d)

Fig. 6. The classification accuracies of NN, PCA, TCA, TSL, LSDT, JDA, MIDA, SRARAs, LDADR, MTL-DRA, and RMTL-DRA on the Office and Caltech-256 databases. (a) Amazon is the source domain and the Webcam, Caltech-256, and DSLR are the target domain, respectively, (b) the Webcam is the source domain and the Amazon, Caltech-256, and DSLR are the target domain, respectively, (c) Caltech-256 is the source domain and the Amazon, Webcam, and DSLR are the target domain, respectively, (d) DSLR is the source domain and the Amazon, Webcam, and Caltech-256 are the target domain, respectively.

database as the target domain, i.e., $M \to U$. In the two recognition tasks, we used the original images and images with 10% pixel corruption to carry out experiments. That is, data of the source and target domains are original samples or data from the source and target domain are all with 10% pixel corruption. Fig. 5 shows the experimental results of all the compared methods on the two databases.

D. Experiments on Office and Caltech-256 Databases

As a visual domain adaptation benchmark dataset, Office includes three different object categories, i.e., Amazon, DSLR, and Webcam. There are 4,652 images from 31 object categories in the database. The images of each category in the Amazon domain is 90 images on average and the average image number in DSLR or the Webcam domain is 30. There are 30,607 object images in Caltech-256 database.

In our experiments, we selected four domains to conduct experiments, i.e., A (Amazon), W (Webcam), C (Caltech-256), and D (DSLR). The experimental design is as follows. Two different databases were randomly selected from the four databases as the source and target domains. Thus, we have 12 cross domain, i.e., $A \to W$, $A \to C$, $A \to D$,…, $C \to D$. The experimental results are reported in Fig. 6.

From Fig. 6, we can see that our proposed methods obtain the best classification accuracy of the compared methods, which shows that our methods can migrate more useful information to the target domain for image classification.

E. Experiments on the COIL100 Database

In this section, we use the Columbia Object Image Library 100 (COIL 100) [37] to test the performance of the proposed methods on object classification. There are 100 objects in the COIL 100 database, and each object has 72 images. In our experiments, we partitioned the database into two subsets COIL1 and COIL2. COIL1 includes all images taken in the poses of $[0^\circ, 85^\circ] \cup [180^\circ, 265^\circ]$, hence the number of all images is 3,600. COIL2 is constructed from all images with directions of $[90^\circ, 175^\circ] \cup [270^\circ, 355^\circ]$, hence the number of all images is again 3,600. The two subsets have relatively different distributions, and were used iteratively as the source and target data, i.e., COIL1 (source) vs COIL2 (target) ($C1 \to C2$) and COIL2 (source) vs COIL1 (target) ($C2 \to C1$). To verify the robustness of the proposed methods, we added black blocks to the images of the target data as noise to conduct experiments. The size of the added block was 10×10 . Fig. 7 shows some image examples with the block occlusion. Tables II and III show the experimental results on the COIL100 database. From the two tables, we can see that our methods outperform other compared methods.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

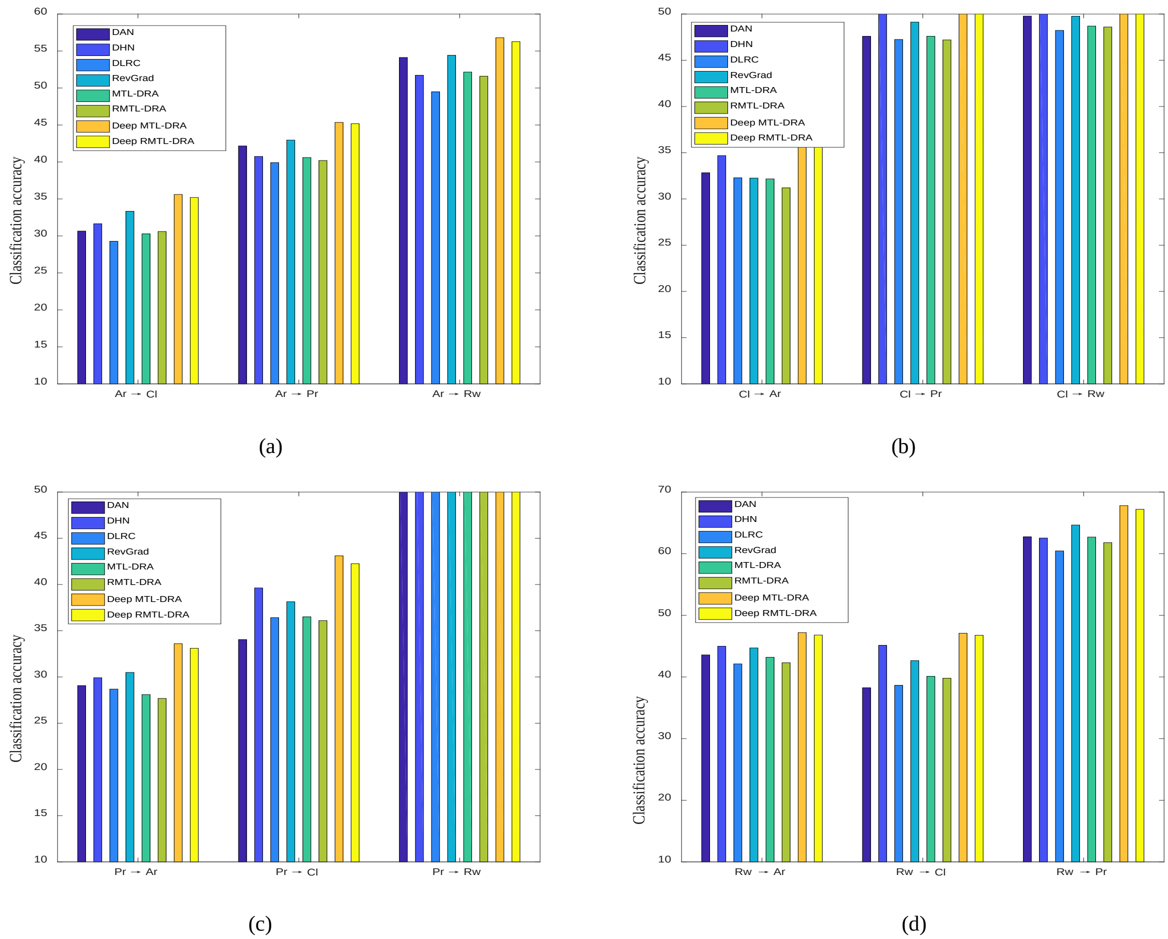


Fig. 8. The classification accuracies of DAN, DHN, DLRC, RevGrad, MTL-DRA, RMTL-DRA, Deep MTL-DRA, and Deep RMTL-DRA on the Office and Home databases. (a) Art is the source domain and Clipart, Product, and Real-World are the target domain, respectively, (b) Clipart is the source domain and Art, Product, and Real-World are the target domain, respectively, (c) Product is the source domain and Art, Clipart, and Real-World are the target domain, respectively, (d) Real-World is the source domain and Art, Clipart, and Product are the target domain, respectively.

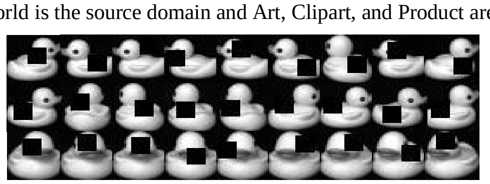


Fig. 7. Image examples with the block occlusion from the COIL100 database.

F. Comparison of Deep Learning

In this sub-section, we extend our methods to a deep case and conduct experiments to verify their performance. By using the features learned by convolutional neural network (CNN) [49], we given the deep extension of our methods, that are deep MTL-DRA and deep RMTL-DRA. To verify the performance of the deep extension of our methods, we conduct experiments on the Office + Home [50] benchmark. As a most recent cross-domain benchmark, Office + Home has been widely used in domain adaptation. There are four domains contain 15500 images from 65 objects. Specifically, four domains are Art (artistic drawing objects), Clipart (images collected from www.clipart.com), Product (samples similar to Amazon almost with clean back- ground), and Real-World (object images taken with regular cameras).

Some deep transfer methods were used as compared methods, including deep adaptation network (DAN) [51], deep hashing network (DHN) [50], deep low-rank coding (DLRC) [52], and reverse gradient (RevGrad) [53]. Fig. 8 is the experiments on the Office + Home benchmark. We can find that deep MTL-DRA and deep RMTL-DRA outperform better than other methods, which show that MTL-DRA and RMTL-DRA are effective for deep learning.

G. Results and Discussion

From the theory and experimental results, we can make the following observations.

1) From Figs. 4-6 and Tables II and III, we can see that NN and PCA have lower recognition rates than other compared methods. This is mainly because the other compared transfer learning methods can learn more information than NN and PCA.

2) Among all the transfer learning methods, we find that the recognition rate of TCA is lower than others. This is because TCA has the limitation that the difference in the conditional distributions is not explicitly reduced [46].

3) JDA performs better than TCA, TSL, and LSDT, mainly because JDA can match the conditional distributions by exploring sufficient statistics.

4) In most cases, our method MTL-DRA has the best recognition rate. This is because MTL-DRA not only migrates the discrimination information from the source domain to the

target domain, but also enhances the robustness to noise.

5) When the data are noisy, the classification performance of RMTL-DRA is the best. This verifies that encoding the nuclear norm as a constraint of the transform matrix can avoid negative transfers for image classification.

V. Conclusion

What to transfer and how to transfer are two key problems for transfer learning. Especially, how to effectively transfer information from the source domain to the target domain is important. In this paper, to transfer the local geometry structure information and ensure the transform matrix is robust or sparse, we propose a novel transfer learning method, named manifold transfer learning via discriminant regression analysis (MTL-DRA) for image classification. In MTL-DRA, we encode discriminant information of the source domain to the target domain by introducing between- and within-class graphs. Furthermore, with different norms as constraints, MTL-DRA overcomes the disturbance of noise and avoids negative transfers. To further enhance the robustness of MTL-DRA, we encode a nuclear norm instead of the L_{21} norm as a constraint, and propose robust MTL-DRA (RMTL-DRA). The optimization, solution, convergence and complexity analyses of the two proposed methods have been described in detail. Extensive experiments conducted on five public benchmarks verify the performance of the proposed methods. The experimental results show the effectiveness of the proposed methods for transfer learning.

Appendix

Let us assume that MTL-DRA reaches a stationary point. The KKT conditions for (16) are derived as follows.

$$\begin{aligned}
&W^T X_t - W^T X_s Z - E = 0, Z - Z_1 = 0, Z_1 - AD = 0,\\
&\frac{\partial L}{\partial W} = X_s(X_s^T W - U^T) + (2\beta R^T + \mu QQ^T)W - \mu QQ_1^T + GW = 0,\\
&\frac{\partial L}{\partial Z} = X_s^T W W^T X_s Z - X_s^T W Q_2 + Z - Z_1 + \frac{M_2}{\mu} = 0,\\
&\frac{\partial L}{\partial Z_1} = 2Z_1 - Z - \frac{M_2}{\mu} - AD + \frac{M_3}{\mu} = 0,\\
&\frac{\partial L}{\partial A} = A + \mu ADD^T - (\mu Z_1 + M_3)D^T = 0,\\
&\frac{\partial L}{\partial D} = D + \mu A^T AD - A^T(\mu Z_1 + M_3) = 0,\\
&M_1 \in \alpha \partial_E \| E \|_1,
\end{aligned} \tag{56}$$

From the seventh relationship in (56), we can obtain that:

$$\begin{aligned}
&W^T X_t - W^T X_s Z + \frac{M_1}{\mu} \in W^T X_t - W^T X_s Z\\
&+\alpha \frac{\partial_E \| W^T X_t - W^T X_s Z \|_1}{\mu} \triangleq \vartheta_{\frac{\alpha}{\mu}}(W^T X_t - W^T X_s Z)
\end{aligned}, \tag{57}$$

where $\vartheta_{\alpha/\mu}(t) \triangleq t - (\mu/\alpha)\partial|t|$ is applied element-wise to $W^T X_t - W^T X_s Z$.

Then we can obtain the following relationship [31]:

$$E = \vartheta^{-1}_{\frac{\alpha}{\mu}}(W^T X_t - W^T X_s Z + \frac{M_1}{\mu}) \equiv S(W^T X_t - W^T X_s Z + \frac{M_1}{\mu}, \frac{\alpha}{\mu}), \tag{58}$$

where $S(x,\tau) = sign(x)\max(|x| - \tau, 0)$.

Therefore, the KKT conditions are as follows:

$$\begin{aligned}
&W^T X_t - W^T X_s Z - E = 0, Z - Z_1 = 0, Z_1 - AD = 0,\\
&\frac{\partial L}{\partial W} = X_s(X_s^T W - U^T) + (2\beta R^T + \mu QQ^T)W\\
&- \mu QQ_1^T + GW = 0,\\
&\frac{\partial L}{\partial Z} = X_s^T W W^T X_s Z - X_s^T W Q_2 + Z - Z_1 + \frac{M_2}{\mu} = 0,\\
&\frac{\partial L}{\partial Z_1} = 2Z_1 - Z - \frac{M_2}{\mu} - AD + \frac{M_3}{\mu} = 0,\\
&\frac{\partial L}{\partial A} = A + \mu ADD^T - (\mu Z_1 + M_3)D^T = 0,\\
&\frac{\partial L}{\partial D} = D + \mu A^T AD - A^T(\mu Z_1 + M_3) = 0,\\
&E = S(W^T X_t - W^T X_s Z + \frac{M_1}{\mu}, \frac{\alpha}{\mu}),
\end{aligned}, \tag{59}$$

Next, we prove the convergence of MTL-DRA to a point that satisfies the KKT conditions.

Theorem 1: Let $\theta \triangleq (W, Z, Z_1, A, D, E, M_1, M_2, M_3, \mu)$ and $\{\theta^j\}_j^\infty$ be generated by MTL-DRA. Assume that $\{\theta^j\}_j^\infty$ is bounded, and $\lim_{j\to\infty}\{\theta^{j+1} - \theta^j\} = 0$. Then, any accumulation point of $\{\theta^j\}_j^\infty$ satisfies the KKT conditions. Specifically, whenever $\{\theta^j\}_j^\infty$ converges, it converges to a KKT point.

Proof: We first obtain the Lagrange multipliers M_1, M_2, and M_3 from Algorithm 1:

$$\begin{aligned}
&M_1^+ = M_1 + \mu(W^T X_t - W^T X_s Z - E)\\
&M_2^+ = M_2 + \mu(Z - Z_1)\\
&M_3^+ = M_3 + \mu(Z_1 - AD)
\end{aligned}, \tag{60}$$

where M_i^+ $(i = 1,2,3)$ is a next point of M_i in a sequence $\{M_i^j\}_{j=1}^\infty$. If sequences of variables $\{M_1^j\}_{j=1}^\infty$, $\{M_2^j\}_{j=1}^\infty$, and $\{M_3^j\}_{j=1}^\infty$ converge to a stationary point, that is, $(M_1^+ - M_1) \to 0$, $(M_2^+ - M_2) \to 0$, and $(M_3^+ - M_3) \to 0$, then $(W^T X_t - W^T X_s Z - E) \to 0$, $(Z - Z_1) \to 0$, and $(Z_1 - AD) \to 0$. Thus, the first three of the KKT conditions are satisfied.

Next, we have the following equation from Algorithm 1 for the fourth KKT condition:

$$\begin{aligned}
&(X_s X_s^T + 2\beta R^T + \mu QQ^T + G)(W^+ - W)\\
&= X_s U^T + \mu QQ_1^T - (X_s X_s^T + 2\beta R^T + \mu QQ^T + G)W
\end{aligned}. \tag{61}$$

From the fourth condition in (59), we can derive $X_s U^T + \mu QQ_1^T - (X_s X_s^T + 2\beta R^T + \mu QQ^T + G)W \to 0$, when $(W^+ - W) \to 0$.

Similar to the procedure used to verify the fourth condition, the fifth KKT condition in (59) can also be obtained:

$$\begin{aligned}&(X_s^T WW^T X_s + I)(Z^+ - Z)\\ &= X_s^T WQ_2 + Z_1 - \frac{M_2}{\mu} - (X_s^T WW^T X_s + I)Z\end{aligned} . \quad (62)$$

We can infer that $X_s^T WQ_2 + Z_1 - \frac{M_2}{\mu} - (X_s^T WW^T X_s + I)Z \to 0$, when $(Z^+ - Z) \to 0$.

The sixth KKT condition is

$$Z_1^+ - Z_1 = \frac{1}{2}(Z + AD + \frac{M_2 - M_3}{\mu}) - Z_1 . \quad (63)$$

We have $\frac{1}{2}(Z + AD + \frac{M_2 - M_3}{\mu}) - Z_1 \to 0$, when $(Z_1^+ - Z_1) \to 0$.

Similarly, the seventh KKT condition is

$$(A^+ - A)(I + \mu DD^T) = (\mu Z_1 + M_3)D^T - A - \mu ADD^T \quad (64)$$

we have $(\mu Z_1 + M_3)D^T - A - \mu ADD^T \to 0$ as $(A^+ - A) \to 0$.

The eighth KKT condition is

$$(I + \mu A^T A)(D^+ - D) = A^T(\mu Z_1 + M_3) - D - \mu A^T AD \quad (65)$$

When $(D^+ - D) \to 0$, we have $A^T(\mu Z_1 + M_3) - D - \mu A^T AD \to 0$.

Last, we obtain the following equation:

$$E^+ - E = S(W^T X_t - W^T X_s Z + \frac{M_1}{\mu}, \frac{\alpha}{\mu}) - E . \quad (66)$$

When $(E^+ - E) \to 0$, we obtain the last KKT condition.

Since $\{\theta^j\}_j^\infty$ is bounded by assumption, $\{A^+ A\}_{j=1}^\infty$ and $\{D^+ D\}_{j=1}^\infty$ in (64) and (65) are bounded as well. Hence, $\lim_{j\to\infty}(\theta^{j+1} - \theta^j) = 0$ implies that both sides of equations (60)-(66) tend to zero as $j \to \infty$. Therefore, the sequence $\{\theta^j\}_j^\infty$ asymptotically satisfies the KKT conditions for (16). □

REFERENCES

[1] Q. Fu, Y. Luo, Y. Wen, D. Tao, Y. Li, and L. Duan, "Toward Intelligent Product Retrieval for TV-to-Online (T2O) Application: A Transfer Metric Learning Approach," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2114-2125, Aug. 2018.

[2] L. Shao, F. Zhu, and X. Li, "Transfer Learning for Visual Categorizations: A Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019-1034, May. 2015.

[3] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative Transfer Subspace Learning via Low-Rank and Sparse Representation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850-863, Feb. 2016.

[4] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring Knowledge Fragments for Learning Distance Metric from a Heterogeneous Domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 1013-1026, Apr. 2019.

[5] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting Image Memorability Through Adaptive Transfer Learning From External Sources," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775-788, Apr. 2016.

[6] Z. Peng, W. Zhang, N. Han, X. Fang, P. Kang, and L. Teng, "Active Transfer Learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1022-1036, Apr. 2020.

[7] D. Wang, C. Lu, J. Wu, H. Liu, W. Zhang, F. Zhuang, and H. Zhang, "Softly Associative Transfer Learning for Cross-Domain Classification," *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2019.2891577.

[8] Y. Chen, S. Song, S. Li, L. Yang, and C. Wu, "Domain Space Transfer Extreme Learning Machine for Domain Adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1909-1922, May 2019.

[9] B. Gholami, O. Rudovic, and V. Pavlovic, "PUnDA: Probabilistic Unsupervised Domain Adaptation for Knowledge Transfer Across Visual Categories," *in Proc. IEEE Int. Conf. Comput. Vis.*, 22-29, Oct. 2017.

[10] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C. Chen, "Guide Subspace Learning for Unsupervised Domain Adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, DOI: 10.1109/TNNLS.2019.2944455.

[11] S. Wang, L. Zhang, W. Zuo, and B. Zhang, "Class-specific Reconstruction Transfer Learning for Visual Recognition Across Domains," *IEEE Trans. Image Process.*, vol. 29, pp. 2424-2438, 2020.

[12] L. Zhang, S. Wang, G. Huang, W. Zuo, J. Yang, and D. Zhang, "Manifold Criterion Guided Transfer Learning via Intermediate Domain Generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3759-3773, Dec. 2019.

[13] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting Self-Supervised Learning via Knowledge Transfer," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 18-23, Jun. 2018.

[14] Y. Li, J. Zhang, K. Huang, and J. Zhang, "Mixed Supervised Object Detection with Robust Objectness Transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 639-653, Mar. 2019.

[15] Y. Han, F. Wu, Y. Zhuang, and X. He, "Multi-Label Transfer Learning with Sparse Representation," *IEEE Trans. Circuits Syst. Video Technol*, vol. 20, no. 8, pp. 1110-1121, Aug. 2010.

[16] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting Image Memorability Through Adaptive Transfer Learning From External Source," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1050-1062, May. 2017.

[17] L. Zhang, S. Jha, T. Liu, and G. Pei, "Discriminative Kernel Transfer Learning via L21-Norm Minimization," *in Proc. Int. Joint Conf. Neural Netw.*, 24-29, Jul. 2016.

[18] Y. Chen, Z. Lai, W. Wong, L. Shen, and Q. Hu, "Low-Rank Linear Embedding for Image Recognition," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3212-3222, Dec. 2018.

[19] Y. Lu, Z. Lai, X. Li, W. Wong, C. Yuan, and D. Zhang, "Low-Rank 2-D Neighborhood Preserving Projection for Enhanced Robust Image Representation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1859-1872, May 2019.

[20] F. Nie, Z. Hu, and X. Li, "Matrix Completion Based on Non-Convex Low-Rank Approximation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2378-2388, May 2019.

[21] P. Jing, Y. Su, L. Nie, X. Bai, J. Liu, and M. Wang, "Low-Rank Multi-View Embedding Learning for Micro-Video Popularity Prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1519-1532, Aug. 2018.

[22] Y. Su, D. Hong, Y. Li, and P. Jing, "Low-Rank Regularized Deep Collaborative Matrix Factorization for Micro-video Multi-label Classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 740-744, 2020.

[23] Y. Lu, C. Yuan, Z. Lai, X. Li, W. Wong, and D. Zhang, "Nuclear Norm-Based 2DLPP for Image Classification," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2391-2403, Nov. 2017.

[24] P. Jing, S. Ye, L, Nie, J. Liu, and Y. Su, "Low-Rank Regularized Multi-Representation Learning for Fashion Compatibility Prediction," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1555-1566, Jun. 2020.

[25] P. Li, J. Yu, M Wang, L. Zhang, D. Cai, and X. Li, "Constrained Low-Rank Learning Using Least Squares-Based Regularization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4250-4262, Dec. 2017.

[26] 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.

[27] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Jan. 2010.

[28] S. M. Xiang, F. P. Nie, G. F. Meng, C. H. Pan, and C. S. Zhang, "Discriminative least squares regressions for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol, 23, no. 11, pp. 1738-1754, Nov. 2012.

[29] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Math. Program.*, 2010.

[30] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[31] E. Kim, M. Lee, and S. Oh "Elastic-net regularization of singular values for robust subspace learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 915–923.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3007340, IEEE Transactions on Multimedia

[32] F. Zhang, J. Yang, J. Qian, and Y. Xu, "Nuclear Norm-Based 2-DPCA for Extracting Features From Images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2247-2260, Oct. 2015.

[33] Y. Lu, C. Yuan, Z. Lai, X. Li, D. Zhang, and W. Wong, "Horizontal and Vertical Nuclear Norm-Based 2DLDA for Image Representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 941-955, Apr. 2019.

[34] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust Neighborhood Preserving Projection by Nuclear/L21-Norm Regularization for Image Feature Extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607-1622, Apr. 2017.

[35] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear Norm-Based Matrix Regression with Applications to Face Recognition with Occlusion and Illumination Changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[36] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 15, pp. 1615–1618, Dec. 2003.

[37] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Rep. CUCS-006-96, 1996.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[39] A. Frank and A. Asuncion. (2010). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[40] L. Zhang and D. Zhang, "Robust Visual Knowledge Transfer via Extreme Learning Machine-Based Domain Adaptation," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4959-4973, Oct. 2016.

[41] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *Int. J. Comput. Vis.*, vol. 109, no. 1, pp. 74–93, 2014.

[42] P. N. Belhumeur, J. P. Hepanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[43] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[44] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.

[45] L. Zhang, W. Zuo, and D. Zhang, "LSDT: Latent Sparse Domain Transfer Learning for Visual Adaption," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1177-1191, Mar. 2016.

[46] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[47] W. Jiang, H. Gao, W. Lu, W. Liu, F. Chung, and H. Huang, "Stacked Robust Adaptively Regularized Auto-Regressions for Domain Adaptation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 561-574, Mar. 2019.

[48] J. Tao, H. Xu, and J. Fu, "Low-Rank Constrained Latent Domain Adaptation Co-Regression for Robust Depression Recognition," *IEEE Access*, vol. 7, pp. 145406-145425, 2019.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hintom, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097-1105.

[50] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep Hashing Network for Unsupervised Domain Adaptation," in *Proc. CVPR*, 2017, pp. 5018-5027.

[51] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning Transferable Features with Deep Adaptation Networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 97-105.

[52] Z. Ding, M. Shao, and Y. Fu, "Deep Low-Rank Coding for Transfer Learning," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3453-3459.

[53] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1180-1189.

Yuwu Lu received the B.S. degree in mathematics from Xingtai University, Xingtai, China, in 2008, the M.S. degree in mathematics from the Inner Mongolia University of Technology, Hohhot, China, in 2011, and the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2015.

He has been a Postdoctoral Fellow with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. He is an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen. He has authored more than 15 scientific papers in pattern recognition and computer vision. His current research interests include pattern recognition and machine learning.

Wenjing Wang received the B.S. degree in software engineering from Henan University, Kai Feng, China, in 2018. She is currently working toward the M.S. degree in the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her current research interests include transfer learning and low rank learning

Chun Yuan is currently an Associate Professor in the Division of Information Science and Technology in Graduate school at Shenzhen, Tsinghua University. He received the M.S. and Ph.D. degrees from the Department of Computer Science and technology, Tsinghua University, Beijing, China, in 1999 and 2002, respectively. He once worked at the INRIA-Rocquencourt, Paris, France, as a Post-doc research fellow from 2003 to 2004. In 2002, he worked at Microsoft Research Asia, Beijing, China, as an intern. His research interests include computer vision, machine learning, video coding and processing, cryptography and digital rights management.

Xuelong Li (M'02-SM'07-F'12) is a full professor with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P.R. China.

Zhihui Lai received the B.S. degree in mathematics from South China Normal University, M.S. degree from Jinan University, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2002, 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. He has published over 60 scientific articles. Now he is an associate editor of International Journal of Machine Learning and Cybernetics. For more information including all papers and related codes, the readers are referred to the website (http://www.scholat.com/laizhihui).