



# Graph-structure constraint and Schatten $p$ -norm-based unsupervised domain adaptation for image classification

Heyou Chang<sup>1,2</sup> · Fanlong Zhang<sup>3</sup> · Guangwei Gao<sup>4</sup> · Hao Zheng<sup>1</sup>

Received: 20 March 2020 / Accepted: 17 July 2020 / Published online: 8 August 2020  
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## Abstract

Unsupervised domain adaptation, which aims to classify a target domain correctly only using a labeled source domain, has achieved promising performance yet remains a challenging problem. Most traditional methods focus on exploiting either geometric or statistical characteristics to reduce domain shifts. To take advantage of both sides, in this paper, we propose a unified framework incorporating both the geometric and statistical characteristics by adopting the non-convex Schatten  $p$ -norm and graph Laplacian constraints to preserve global and local structure information and constructing marginal and conditional distribution minimization terms to reduce the distribution shifts. Moreover, a classification error term on the source domain is embedded into the objective function to increase the discriminability. The proposed method has been evaluated on six datasets and the experimental results demonstrate the superiority of the proposed method over several state-of-the-art methods. The MATLAB code of our method will be publicly available at <https://github.com/HeyouChang/unsupervised-domain-adaptation>.

**Keywords** Unsupervised domain adaptation · Image classification · Distribution shift · Geometric structure · Statistical characteristic

## 1 Introduction

As an important research field of computer vision, image classification has been widely studied in the past few years, and many methods have been proposed (Lan et al. 2019; Wright et al. 2009; Yang et al. 2013). In these methods, it is

common to assume that samples from the training set and testing set have a similar distribution. However, it is difficult to assure that both the two sets follow the same distribution in many practical applications due to various factors (i.e. resolution, viewpoint and illumination). A model obtained from a training set usually fails in the testing set if their distributions are different. To contend with the scenario, transfer learning attracts lots of attention, which aims to effectively apply the knowledge learned from a training (source) domain to a testing (target) domain. Recently, numerous works on transfer learning have been proposed and have achieved exciting performance in many applications, such as image classification (Kobylarz et al. 2020; Singh et al. 2019; Wang et al. 2017), motion segmentation (Wang et al. 2018b) and image retrieval (Xu et al. 2019).

As a special case of transfer learning, domain adaptation (DA) attracts lots of attention in recent years. In DA, it is a popular strategy to take the information from both domains into consideration to extract new domain-invariant features for the two domains. According to the availability of the target labels in the training process, DA can be divided into two categories: semi-supervised DA and unsupervised DA.

---

✉ Heyou Chang  
cv\_hychang@126.com

✉ Hao Zheng  
zh710@163.com

Fanlong Zhang  
csfzhang@nau.edu.cn

Guangwei Gao  
csggao@gmail.com

<sup>1</sup> School of Information and Engineering, Nanjing Xiaozhuang University, Nanjing, China

<sup>2</sup> Laboratory of Image Science and Technology, Southeast University, Nanjing, China

<sup>3</sup> School of Technology, Nanjing Audit University, Nanjing, China

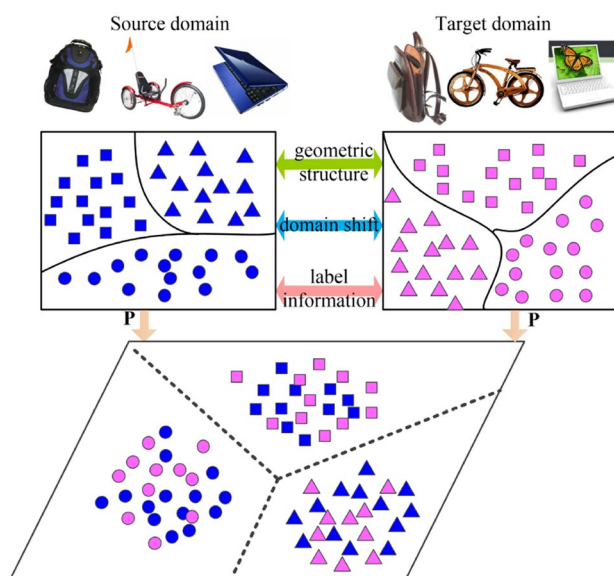
<sup>4</sup> Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

This paper focuses on the unsupervised DA, which is more difficult and challenging.

Finding a common latent subspace to reduce the distribution shift problem between the two domains is critical in unsupervised DA (Si et al. 2010; Singh and Nigam 2019). For example, Si et al. (2010) constructed a regularization to minimize the Bregman divergence between the two domains in a selected subspace. Gong et al. (2012) utilized geodesic flow kernel (GFK) to find a geodesic from the source domain to the target domain. Long et al. (2013b) proposed to project data of both domains into a new subspace by jointly embedding both marginal distribution and conditional distribution into a principled dimensionality reduction procedure. Ghifary et al. (2017) proposed scatter component analysis (SCA) for DA and domain generalization, in which data scatter is used as a geometrical measure to evaluate the separability of classes, the mismatch between domains and the separability of data. To reduce the distribution divergence and evaluate the importance of the marginal and conditional distributions, Wang et al. (2018a) proposed to learn a domain-invariant classifier in a Grassmann manifold with structural risk minimization. Although good performance has been reported, these methods seldom exploit the structural information of the data.

Many studies (Xu et al. 2016; Shao et al. 2014) have verified that importing structure constraints (such as low-rank constraint and sparse constraint) into the transfer learning process could effectively improve the DA performance. For instance, Shao et al. (2014) proposed low-rank transfer subspace learning (LTSL) for face recognition and object recognition, where each sample from target domain is reconstructed by the samples from source domain in a generalized subspace with a low-rank constraint. Xiao et al. (2019) proposed structure preservation and distribution alignment (SPDA) for unsupervised DA. However, these works ignore the statistical characteristics of the data. Moreover, all these methods use the convex nuclear norm to approximate the non-convex rank function, which may make the solution deviate considerably from the original solution (Nie et al. 2012).

To better approximate the original low rank, mitigate the distribution shift problem and further exploit the structure properties, we propose a unified framework that incorporates structure information and statistical distribution for unsupervised DA of image classification, as shown in Fig. 1. Specifically, the proposed framework retains both global and local structural information by constructing a graph-structure constraint and Schatten  $p$ -norm; and reduces the distribution shifts by aligning both marginal and conditional distributions between the source and target domains. Moreover, the label information of the source domain is also exploited effectively by  $\epsilon$ -dragging technique. An effective optimization procedure is proposed. Experiments on six datasets have



**Fig. 1** Illustration of our method. Blue: source samples. Pink: target samples. Circles, squares, and triangles indicate three different categories. By considering structural information, marginal and conditional distribution shifts, as well as label information of the source domain, the hyper-plane learned from the two domains can perfectly classify the target samples

been done and the results demonstrate the superiorities of the proposed approach.

The contributions of this work are summarized as follows. (1) A unified framework consists of global and local structure constraints, marginal and conditional distributions and classification error is proposed for unsupervised DA and the Schatten  $p$ -norm ( $0 < p < 1$ ) is introduced to capture the data structure more accurately. (2) Better performance has been achieved on six benchmark datasets than several representative unsupervised DA methods.

The rest of this paper is structured as follows: Sect. 2 introduces the notation used in this paper. Section 3 briefly reviews some related works. Section 4 shows the description and optimization procedure about the proposed method in detail. Section 5 discusses the experiments and results on the six datasets, and the last section concludes the paper.

## 2 Preliminaries

In this work, we denote  $X_s \in R^{d \times n_s}$ ,  $X_t \in R^{d \times n_t}$ ,  $Z \in R^{n_s \times n_t}$ ,  $P \in R^{d \times m}$  and  $E \in R^{m \times n_t}$  as the source data, target data, reconstruction matrix of  $X_t$ , projection matrix and noise matrix, respectively, where  $d$  is the dimensionality of each sample,  $n_s$  and  $n_t$  are the number of samples in  $X_s$  and  $X_t$ , respectively. Matrices (vectors) are denoted by boldface uppercase (lowercase) letters.  $A_{i,j}$  is the  $i, j$ th entry of  $A$ , and  $A_i$  represents the  $i$ th column of  $A$ .

The Schatten  $p$ -norm ( $0 < p < \infty$ ) of a matrix  $A \in R^{m \times n}$  is defined as  $\|A\|_{s_p} = (\sum_{i=1}^{\min\{m,n\}} \sigma_i^p)^{1/p} = (tr(A^T A)^{p/2})^{1/p}$ , where  $tr(\cdot)$  represents the trace operator. When  $p = 1$ , the Schatten  $p$ -norm becomes the nuclear norm ( $\|\cdot\|_*$ ). Although the Schatten  $p$ -norm is only a quasi-norm when  $p < 1$ , for convenience, we still call it the Schatten  $p$ -norm.

### 3 Related work

#### 3.1 Domain adaptation

DA aims at addressing the problem where the task of training and testing domains are the same while their data distributions are different. In DA, it is a popular strategy to learn a common feature subspace where the distributions of both domains are well aligned. Following this, some subspace learning methods have been proposed. In Si et al. (2010), the authors reduced the distribution shift by minimizing the Bregman divergence. Long et al. (2013b) made use of the pseudo-labels of the target domain to calculate conditional distribution shift and proposed joint distribution analysis (JDA). By adaptively weighting the marginal and conditional distributions, Wang et al. (2017) proposed balanced distribution adaptation (BDA) to improve the performance. The methods mentioned above mainly concentrate on minimizing the domain distribution shifts and overlook the geometric information among data.

To preserve specific properties (such as, low-rank and sparsity) in DA, various kinds of regularizers are exploited. For example, Xu et al. (2016) introduced both low-rank and sparse constraints on the coefficient matrix to preserve the structural information of data. Different from (Xu et al. 2016), Shao et al. (2014) constructed a generalized subspace term to preserve the local structure. These methods aim to exploit the geometric characteristics among samples while seldom leverage the statistical characteristics.

To take full advantage of the structure information and the statistical distribution, we incorporate global and local structure preservation, marginal and conditional distributions and classification error into a unified framework. Different from (Xiao et al. 2019), the proposed method (1) utilizes the non-convex Schatten  $p$ -norm to capture the global structure in the data more precisely, and (2) considers that the marginal distribution and conditional distribution have different importance to further minimize the domain shifts.

#### 3.2 Low-rank representation

Given a data matrix  $X \in R^{d \times m}$  with  $m$  samples and a dictionary  $D \in R^{d \times k}$  with  $k$  atoms, low-rank representation (LRR)

concentrates on seeking a representation matrix  $Z \in R^{k \times m}$ , which not only has the lowest rank but also can reconstruct the samples with the dictionary atoms through linear combination. LRR can be formulated as follows:

$$\min_{Z,E} \text{rank}(Z) + \lambda \|E\|_1, \text{ s.t. } X = DZ + E \tag{1}$$

where  $E$  represents noise matrix and  $\|\cdot\|_1$  is the sparsity constraint. Since the rank minimization problem is NP-hard, it is popular to use nuclear norm to approximate the rank function (Shao et al. 2014; Liu et al. 2013; Zhu et al. 2018; Yang et al. 2018).

However, the nuclear norm relaxation may deviate the outcome away from the real solution. The nuclear norm of a matrix is equal to the  $L_1$  norm of the matrix's singular vector. According to the definition of the Schatten  $p$ -norm, when  $p = 1$ , it is equal to the nuclear norm. When  $p \rightarrow 0$ , the Schatten  $p$ -norm becomes rank function under  $0^0 = 0$ . We can see that the Schatten  $p$ -norm is more approximate to the rank function than the nuclear norm when  $0 < p < 1$ . Therefore, we adopt the Schatten  $p$ -norm ( $0 < p < 1$ ) to obtain a closer solution of problem (1).

### 4 The proposed method

#### 4.1 Problem formulation

Based on the assumption that there is a common subspace shared by the samples from both source and target domains and the samples from the target domain can be linearly represented by the samples from the source domain in the common subspace, we can construct a general formulation for unsupervised DA:

$$\min_{P,Z,E} \text{rank}(Z) + \alpha \|Z\|_1 + \beta \|E\|_1 + f(P, X_s, Y_s) \tag{2}$$

$$\text{ s.t. } P^T X_t = P^T X_s Z + E$$

where  $f(P, X_s, Y_s)$  is a function for learning a discriminative projection matrix  $P$ , and  $Y_s$  is the binary label matrix of  $X_s$ .

Although the convex nuclear norm has been widely used to approximate  $\text{rank}(Z)$ , the relaxation may deviate the outcome from the real solution (Nie et al. 2012). Recently, some approaches (Chang et al. 2016, 2019; Wang et al. 2019) have verified that the non-convex Schatten  $p$ -norm minimization performs better than the nuclear norm minimization in image denoising when  $p$  is close to 0. To obtain a closer solution of problem (2), we adopt  $\|\cdot\|_{s_p}^p$  ( $0 < p < 1$ ) to approximate  $\text{rank}(Z)$  by:

$$\min_{P,Z,E} \|Z\|_{s_p}^p + \alpha \|Z\|_1 + \beta \|E\|_1 + f(P, X_s, Y_s) \tag{3}$$

$$\text{ s.t. } P^T X_t = P^T X_s Z + E$$

To leverage the local geometric structure information in the original feature space, we construct a graph-constraint term as

$$\sum_{i,j=1}^n (\mathbf{P}^T \mathbf{X}_i - \mathbf{P}^T \mathbf{X}_j)^2 W_{ij} = \text{tr}(\mathbf{P}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{P}) \tag{4}$$

where  $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t]$ ,  $n = n_s + n_t$ .  $W_{ij} = 1$ , if  $\mathbf{X}_i \in k\text{NN}(\mathbf{X}_j) \vee \mathbf{X}_j \in k\text{NN}(\mathbf{X}_i)$  and  $W_{ij} = 0$ , otherwise.  $\mathbf{L} = \mathbf{D} - \mathbf{W}$  is the normalized graph Laplacian matrix, and  $\mathbf{D}$  is a diagonal matrix with diagonal entries  $D_{i,i} = \sum_{j=1}^n W_{i,j}$ .

In addition to exploiting the structural information, reducing the distribution distance between the two domains is also significant for unsupervised DA. Therefore, the maximum mean discrepancy (MMD) is adopted to measure the marginal distribution difference between the two domains by

$$\|1/n_s \sum_{i=1}^{n_s} \mathbf{P}^T \mathbf{X}_i - 1/n_t \sum_{j=1}^{n_t} \mathbf{P}^T \mathbf{X}_j\|_F^2 = \text{tr}(\mathbf{P}^T \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{P}) \tag{5}$$

where  $\mathbf{B}$  is the marginal distribution MMD matrix, and  $B_{i,j}$  is computed as

$$B_{i,j} = \begin{cases} 1/n_s^2, & \mathbf{X}_i, \mathbf{X}_j \in X_s \\ 1/n_t^2, & \mathbf{X}_i, \mathbf{X}_j \in X_t \\ -1/(n_s n_t), & \text{otherwise} \end{cases}$$

Since the labels of target samples are not available, the conditional distribution difference between the two domains cannot be directly calculated. As in Wang et al. (2017), Long et al. (2013a), we calculate pseudo-labels of the target samples by applying some base classifiers (e.g., NN, SVM), where the classifiers are trained on the labeled source data. Then, the conditional distribution distance can be formulated as

$$\begin{aligned} & \sum_{c=1}^C \|1/n_s^c \sum_{i=1}^{n_s^c} \mathbf{P}^T \mathbf{X}_i - 1/n_t^c \sum_{j=1}^{n_t^c} \mathbf{P}^T \mathbf{X}_j\|_F^2 \\ &= \text{tr}(\mathbf{P}^T \mathbf{X} \sum_{c=1}^C \mathcal{A}_c \mathbf{X}^T \mathbf{P}) \end{aligned} \tag{6}$$

where  $C$  is the class number,  $\mathcal{A}_c$  is the conditional distribution matrix of  $c$ th class.  $\mathcal{A}_c$  is calculated as

$$(\mathcal{A}_c)_{i,j} = \begin{cases} 1/n_s^{c2}, & \mathbf{X}_i, \mathbf{X}_j \in X_s^c \\ 1/n_t^{c2}, & \mathbf{X}_i, \mathbf{X}_j \in X_t^c \\ -1/(n_s^c n_t^c), & (\mathbf{X}_i \in X_s^c, \mathbf{X}_j \in X_t^c) \vee (\mathbf{X}_j \in X_s^c, \mathbf{X}_i \in X_t^c) \\ 0, & \text{otherwise} \end{cases}$$

where  $X_s^c$  ( $X_t^c$ ) represents the set of samples belonging to  $c$ -th class in the source (target) domain, which includes  $n_s^c$  ( $n_t^c$ ) samples.

To maximize the class separation distance and improve classification accuracy, the label information  $\mathbf{Y}_s$  of the source samples should be considered. Following Xu et al. (2016), we design a non-negative label relaxation matrix  $\mathbf{M}$  to increase inter-class separation in the source domain as much as possible and perform the label regularization as follows:

$$\|\mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})\|_F^2 \tag{7}$$

where  $B_{ij} = 1$ , if  $Y_s(i,j) = 1$  and  $B_{ij} = -1$ , otherwise.  $\odot$  is the hadamard product operator.

By jointly taking geometric regularization, statistical regularization, and label regularization into consideration, the final objective function is formulated as follows

$$\begin{aligned} & \min_{\mathbf{P}, \mathbf{Z}, \mathbf{E}, \mathbf{M}} \|\mathbf{Z}\|_{s_p}^p + \alpha \|\mathbf{Z}\|_1 + \beta \|\mathbf{E}\|_1 \\ & + \|\mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})\|_F^2 \\ & + \text{tr}(\mathbf{P}^T \mathbf{X} (\gamma \mathbf{L} + \eta \mathbf{B} + \nu \sum_{c=1}^C \mathcal{A}_c) \mathbf{X}^T \mathbf{P}) \end{aligned} \tag{8}$$

$$\text{s.t. } \mathbf{P}^T \mathbf{X}_t = \mathbf{P}^T \mathbf{X}_s \mathbf{Z} + \mathbf{E}$$

where  $\alpha, \beta, \gamma, \eta$  and  $\nu$  are the trade-off parameters, and  $p \in (0, 1)$ . In the experiment,  $p$  is set to 1/2 for convenience. Compared with  $\|\mathbf{Z}\|_{s^*}$ ,  $\|\mathbf{Z}\|_{s_{1/2}}^{1/2}$  is much closer to  $\text{rank}(\mathbf{Z})$ .

### 4.2 Optimization

To effectively solve problem (8), two auxiliary variables  $\mathbf{J}$  and  $\mathbf{R}$  are first introduced to make the problem separable. Then, (8) can be rewritten as:

$$\begin{aligned} & \min_{\mathbf{P}, \mathbf{Z}, \mathbf{E}, \mathbf{M}, \mathbf{J}, \mathbf{R}} \|\mathbf{J}\|_{s_{1/2}}^{1/2} + \alpha \|\mathbf{R}\|_1 + \beta \|\mathbf{E}\|_1 \\ & + \|\mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})\|_F^2 \\ & + \text{tr}(\mathbf{P}^T \mathbf{X} (\gamma \mathbf{L} + \eta \mathbf{B} + \nu \sum_{c=1}^C \mathcal{A}_c) \mathbf{X}^T \mathbf{P}) \\ & \text{s.t. } \mathbf{P}^T \mathbf{X}_t = \mathbf{P}^T \mathbf{X}_s \mathbf{Z} + \mathbf{E}, \mathbf{Z} = \mathbf{J}, \mathbf{Z} = \mathbf{R} \end{aligned} \tag{9}$$

Then, augmented Lagrangian multiplier (ALM) is applied to solve (9). The augmented Lagrangian function  $H$  of (9) is:

$$\begin{aligned} & H(\mathbf{P}, \mathbf{Z}, \mathbf{E}, \mathbf{M}, \mathbf{J}, \mathbf{R}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mu) = \|\mathbf{J}\|_{s_{1/2}}^{1/2} \\ & + \alpha \|\mathbf{R}\|_1 + \beta \|\mathbf{E}\|_1 + \|\mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})\|_F^2 \\ & + \text{tr}(\mathbf{P}^T \mathbf{X} (\gamma \mathbf{L} + \eta \mathbf{B} + \nu \sum_{c=1}^C \mathcal{A}_c) \mathbf{X}^T \mathbf{P}) \\ & + \langle \mathbf{Y}_1, \mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E} \rangle + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J} \rangle + \langle \mathbf{Y}_3, \mathbf{Z} - \mathbf{R} \rangle \\ & + \mu/2 (\|\mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2 + \|\mathbf{Z} - \mathbf{R}\|_F^2) \end{aligned} \tag{10}$$

where  $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{A}^T \mathbf{B})$ ,  $\mathbf{Y}_1, \mathbf{Y}_2$  and  $\mathbf{Y}_3$  are Lagrange multipliers and  $\mu > 0$  is a penalty parameter. At each iteration step, only one variable is updated while fixing the other variables. The updating process is as follows.

**Step 1. Optimizing  $\mathbf{J}$**

Keeping other variables constant, the problem in Eq. (10) can be simplified as follows:

$$\text{argmin}_{\mathbf{J}} 1/\mu \|\mathbf{J}\|_{s_1/2}^{1/2} + 1/2 \|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}_2/\mu)\|_F^2 \tag{11}$$

Equation (11) can be solved according (Chang et al. 2016), which gives details for the solution.

**Step 2. Optimizing  $\mathbf{P}$**

Keeping other variables constant, Eq. (10) can be rewritten as

$$\begin{aligned} \text{argmin}_{\mathbf{P}} & \text{tr} \left( \mathbf{P}^T \mathbf{X} (\gamma \mathbf{L} + \eta \mathbf{B} + \nu \sum_{c=1}^C \mathcal{A}_c) \mathbf{X}^T \mathbf{P} \right) \\ & + \|\mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})\|_F^2 \\ & + \mu/2 \|\mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E} + \mathbf{Y}_1/\mu\|_F^2 \end{aligned} \tag{12}$$

$\mathbf{P}$  can be updated by taking the stationary point of (12) as

$$\begin{aligned} \mathbf{P}^* = & (\mathbf{X}_s \mathbf{X}_s^T + \mathbf{X} \mathcal{K}_1 \mathbf{X}^T + \mu \mathcal{K}_2 \mathcal{K}_2^T + \lambda \mathbf{I})^{-1} (\mathbf{X}_s (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M})^T \\ & + \mu \mathcal{K}_2 (\mathbf{E} - \mathbf{Y}_1/\mu)^T) \end{aligned} \tag{13}$$

where  $\mathcal{K}_1 = \gamma \mathbf{L} + \eta \mathbf{B} + \nu \sum_{c=1}^C \mathcal{A}_c$ ,  $\mathcal{K}_2 = \mathbf{X}_t - \mathbf{X}_s \mathbf{Z}$ ,  $\mathbf{I}$  is an identity matrix, and  $\lambda$  is a small positive constant.

**Step 3. Optimizing  $\mathbf{Z}$**

Keeping other variables constant,  $\mathbf{Z}$  can be updated by solving

$$\begin{aligned} \text{argmin}_{\mathbf{Z}} & \|\mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E} + \mathbf{Y}_1/\mu\|_F^2 \\ & + \|\mathbf{Z} - \mathbf{J} + \mathbf{Y}_2/\mu\|_F^2 + \|\mathbf{Z} - \mathbf{R} + \mathbf{Y}_3/\mu\|_F^2 \end{aligned} \tag{14}$$

$\mathbf{Z}$  can be calculated by

$$\begin{aligned} \mathbf{Z}^* = & (\mathbf{X}_s^T \mathbf{P} \mathbf{P}^T \mathbf{X}_s + 2\mathbf{I})^{-1} (\mathbf{X}_s^T \mathbf{P} (\mathbf{P}^T \mathbf{X}_t - \mathbf{E} \\ & + \mathbf{Y}_1/\mu)^T + \mathbf{J} - \mathbf{Y}_2/\mu + \mathbf{R} - \mathbf{Y}_3/\mu) \end{aligned} \tag{15}$$

**Step 4. Optimizing  $\mathbf{R}$**

Keeping other variables constant,  $\mathbf{R}$  can be solved as

$$\begin{aligned} \mathbf{R}^* = & \text{argmin}_{\mathbf{R}} \alpha/\mu \|\mathbf{R}\|_1 + 1/2 \|\mathbf{Z} - \mathbf{R} + \mathbf{Y}_3/\mu\|_F^2 \\ = & \Phi_{\alpha/\mu} [\mathbf{Z} + \mathbf{Y}_3/\mu] \end{aligned} \tag{16}$$

where  $\Phi_w[\cdot]$  is the soft thresholding (shrinkage) operator, and  $\Phi_w[x] = \text{signmax}(|x| - w, 0)$

**Step 5. Optimizing  $\mathbf{E}$**

Keeping other variables constant,  $\mathbf{E}$  is updated by

$$\begin{aligned} \mathbf{E}^* = & \text{argmin}_{\mathbf{E}} \beta/\mu \|\mathbf{E}\|_1 \\ & + 1/2 \left\| \mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E} + \mathbf{Y}_1/\mu \right\|_F^2 \\ = & \Phi_{\beta} [\mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} + \mathbf{Y}_1/\mu] \end{aligned} \tag{17}$$

**Step 6. Optimizing  $\mathbf{M}$**

Keeping other variables constant,  $\mathbf{M}$  can be solved by

$$\mathbf{M}^* = \text{argmin}_{\mathbf{M}} \left\| \mathbf{P}^T \mathbf{X}_s - (\mathbf{Y}_s + \mathbf{B} \odot \mathbf{M}) \right\|_F^2 \tag{18}$$

$\mathbf{M}$  can be calculated by

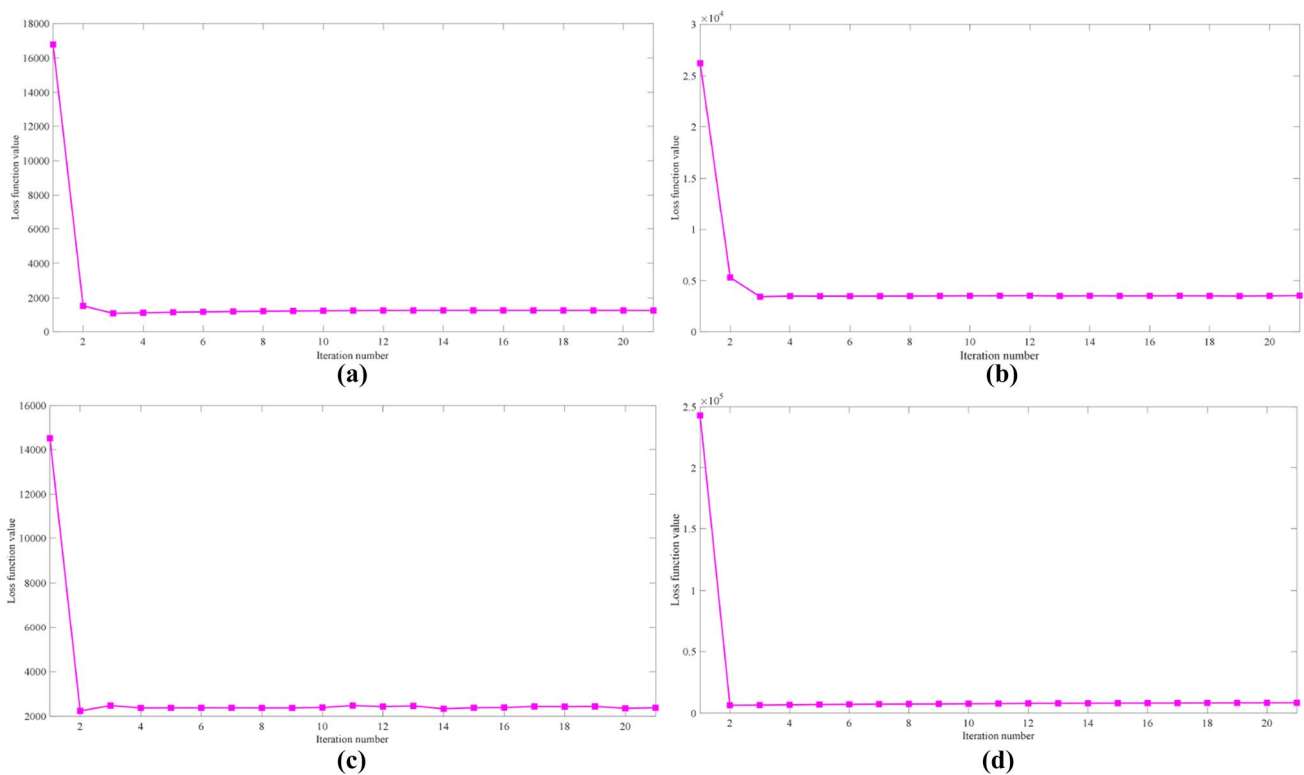
$$\mathbf{M} = \max((\mathbf{P}^T \mathbf{X}_s - \mathbf{Y}_s) \odot \mathbf{B}, 0) \tag{19}$$

The whole optimization of Eq. (8) is summarized in Algorithm 1. A projective matrix  $\mathbf{P}$  is output by the Algorithm 1. When classification, all samples of both domains are first transferred to a new subspace by multiplying  $\mathbf{P}^T$ . Then the label of one target sample is the label of its nearest source sample in the subspace by carrying out a one-nearest-neighbor algorithm.

**Algorithm 1** Solving algorithm of the proposed method

- 1: **Input:** source data  $\mathbf{X}_s$ , target data  $\mathbf{X}_t$ , source data label  $\mathbf{Y}_s$ ,  $\mathbf{B}$ ,  $\alpha, \beta, \gamma, \eta$  and  $\nu$
- 2: **Initialization:**  $\mathbf{M} = \mathbf{I}, \mathbf{Z} = \mathbf{J} = \mathbf{R} = \mathbf{0}, \mathbf{E} = \mathbf{0}, \mathbf{Y}_1 = \mathbf{Y}_2 = \mathbf{Y}_3 = \mathbf{0}, \mathbf{P}, \mathbf{L}, \mathbf{B}$  and  $\mathcal{A}_i, i=1, \dots, C$
- 3: **While not converged do**
- 4:     Update  $\mathbf{J}$  by Eq.(11) with fixing others
- 5:     Update  $\mathbf{P}$  by Eq.(13) with fixing others
- 6:     Update  $\mathbf{Z}$  by Eq.(15) with fixing others
- 7:     Update  $\mathbf{R}$  by Eq.(16) with fixing others
- 8:     Update  $\mathbf{E}$  by Eq.(17) with fixing others
- 9:     Update  $\mathbf{M}$  by Eq.(19) with fixing others
- 10:    Update  $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$  and  $\mu$
- 11:    Check the convergence conditions:  

$$\|\mathbf{P}^T \mathbf{X}_t - \mathbf{P}^T \mathbf{X}_s \mathbf{Z} - \mathbf{E}\|_\alpha < \epsilon, \|\mathbf{Z} - \mathbf{J}\|_\alpha < \epsilon, \|\mathbf{Z} - \mathbf{R}\|_\alpha < \epsilon$$
- 12: **End While**
- 13: **Output:**  $\mathbf{P}$



**Fig. 2** The objective function value versus the iteration number on different datasets. **a** COIL1→COIL2, **b** MNIST→USPS, **c** Caltech256→Amazon, **d** PIE1→PIE2

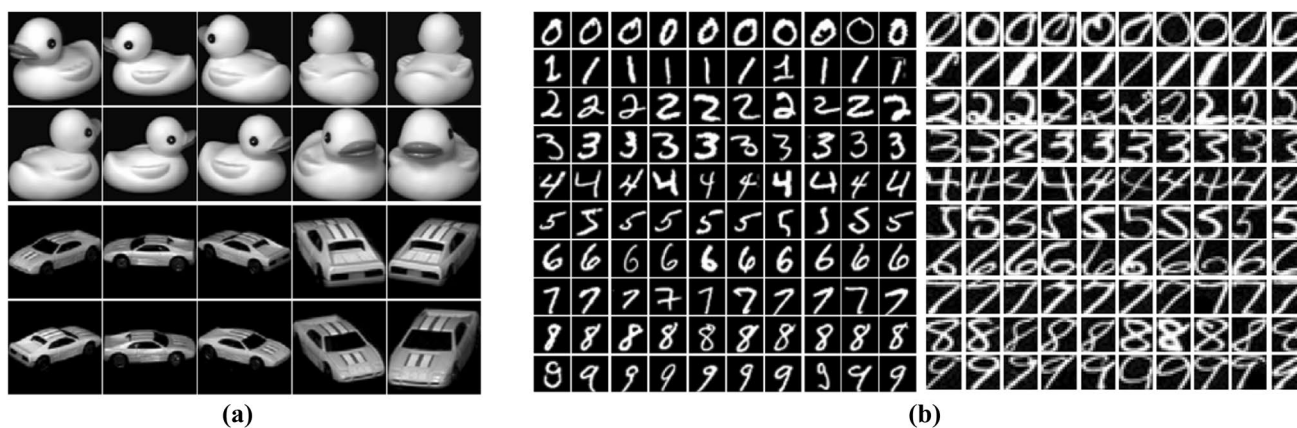
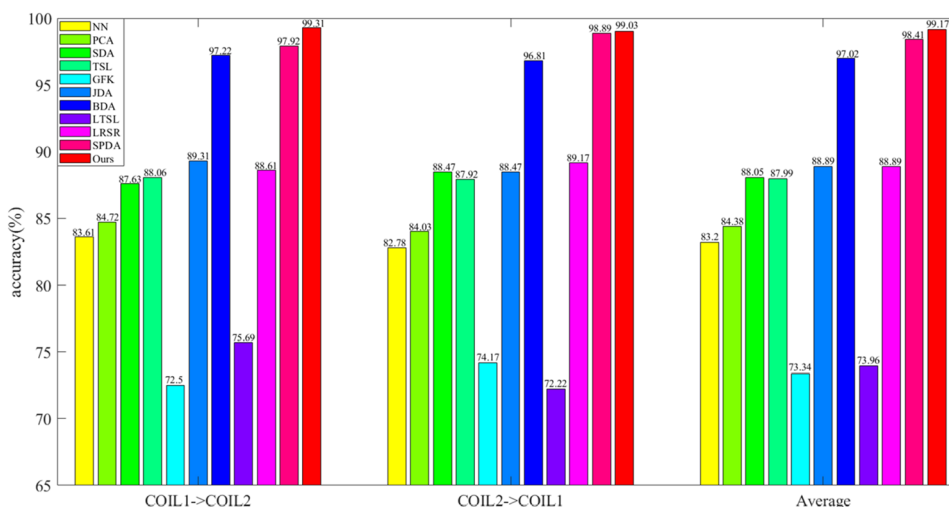


Fig. 3 a Samples of the COIL20 dataset. b Some examples of the MNIST dataset (left) and USPS dataset (right)

Fig. 4 Classification accuracies (%) on the COIL20 dataset



### 4.3 Complexity and convergence analysis

In Algorithm 1, the great mass of run time is consumed in optimizing  $\mathbf{J}$ ,  $\mathbf{P}$  and  $\mathbf{Z}$ , which require matrix inversion and singular value decomposition (SVD). The complexities of optimizing  $\mathbf{J}$ ,  $\mathbf{P}$  and  $\mathbf{Z}$  are  $O(n_s n_t^2)$  (we assume that  $n_t \leq n_s$ ),  $O(d^3 + 2d^2(n_s + n_t))$  and  $O(dn_s m + n_s^3)$ , respectively. Then, the total time complexity of the proposed algorithm is  $O(N(n_s n_t^2 + d^3 + 2d^2(n_s + n_t) + dn_s m + n_s^3))$ , where  $N$  is the maximum iteration. In our experiments,  $N$  is set 15.

Since problem (8) is not smooth (the Schatten 1/2 norm is non-convex) and there are more than two blocks in the proposed algorithm, it is difficult to give a convergent proof of Algorithm 1 in theory. Figure 2 shows the the value of (8) with respect to the number of iterations on four datasets. From Fig. 2, we can see that the objective function value decreases as the number of iterations increases.

### 5 Experiments

The proposed method is evaluated on six datasets which are widely used for unsupervised transfer learning: COIL20, Office, Caltech-256, USPS, MNIST, and CMU PIE. The first three datasets are used for object classification, the middle two datasets are used for digit classification and the last dataset is used for face classification. The proposed method is compared with the latest ten unsupervised transfer learning methods, i.e., SDA (Sun and Saenko 2015), TSL (Si et al. 2010), GFK (Gong et al. 2012), JDA (Long et al. 2013b), BDA (Wang et al. 2017), SCA (Ghifary et al. 2017), LTSL (Shao et al. 2014), LRSR (Xu et al. 2016), LRDRM (Razzaghi et al. 2019) and SPDA (Xiao et al. 2019). Two other standard machine learning methods (i.e., NN and PCA) are also include. The results of all the competing methods are either referenced from the original papers or from widely published results to ensure a fair comparison.



Fig. 5 Some images from the Office and Caltech-256 datasets

**Table 1** Classification accuracies (%) on the Office and Caltech256 datasets (The number in bold indicates the highest accuracy and the number in underline represents the second-highest accuracy)

	Standard learning		Domain adaptation										
	NN	PCA	SDA	TSL	GFK	JDA	BDA	SCA	LTSL	LRSR	LRDRM	SPDA	Ours
C→A	23.70	36.95	49.69	44.47	41.02	44.78	44.89	43.74	50.57	51.25	<u>53.91</u>	52.82	<b>54.59</b>
C→W	25.76	32.54	38.98	34.24	40.68	41.69	38.64	33.56	<u>47.15</u>	38.64	40.00	40.68	<b>48.14</b>
C→D	25.48	38.22	40.13	43.31	38.85	45.22	47.77	39.49	48.40	47.13	49.68	<u>51.59</u>	<b>53.50</b>
A→C	26.00	34.73	39.54	37.58	40.25	39.36	40.78	38.29	37.65	43.37	<u>44.07</u>	43.37	<b>44.88</b>
A→W	29.83	35.59	30.85	33.90	38.98	37.97	39.32	33.90	39.03	36.61	37.28	<b>43.39</b>	<u>39.66</u>
A→D	25.48	27.39	33.76	26.11	36.31	39.49	43.31	34.21	38.89	38.85	42.03	<b>46.50</b>	<u>45.86</u>
W→C	19.86	26.36	34.73	29.83	30.72	31.17	28.94	30.63	<b>35.45</b>	29.83	<u>35.12</u>	31.97	32.17
W→A	22.96	31.00	<u>39.25</u>	30.27	29.75	32.78	32.99	30.48	<b>45.17</b>	34.13	35.07	37.97	37.68
W→D	59.24	77.07	75.80	87.26	80.89	89.17	<u>91.72</u>	<b>92.36</b>	72.61	82.80	84.71	89.81	89.12
D→C	26.27	29.65	<u>35.89</u>	28.50	30.28	31.52	32.50	32.32	35.08	31.61	<b>36.42</b>	33.84	34.02
D→A	28.50	32.05	<b>38.73</b>	27.56	32.05	33.09	33.09	33.72	38.00	33.19	35.07	38.20	<u>38.31</u>
D→W	63.39	75.93	76.95	<u>85.42</u>	75.59	89.49	<b>91.86</b>	88.81	73.75	77.29	76.94	82.37	84.07
Average	31.37	39.79	44.52	42.37	42.95	46.31	47.15	44.29	46.83	45.39	47.45	<u>49.32</u>	<b>50.17</b>

## 5.1 COIL20 dataset

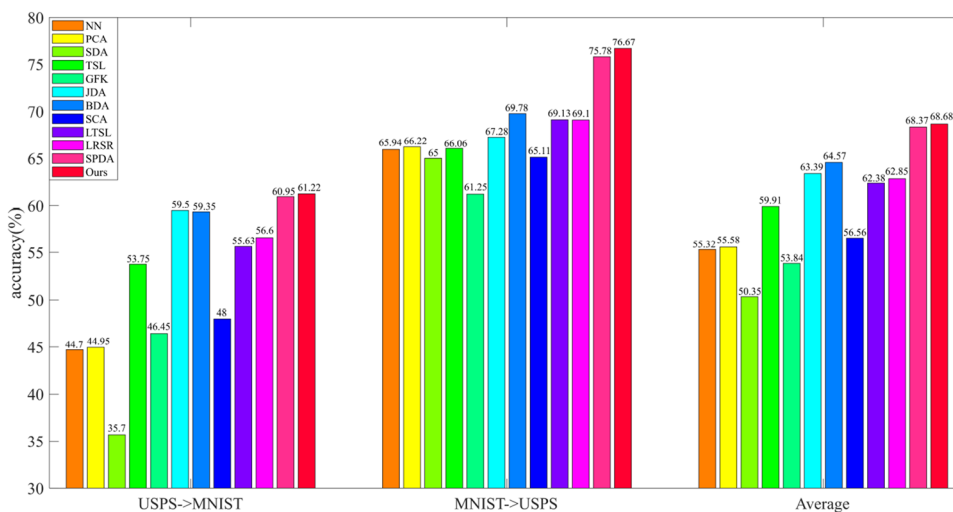
This dataset consists of 20 objects with 1440  $32 \times 32$  gray images. The images were taken every  $5^\circ$  as the objects were rotated on a motorized turntable against a black background. In the experiment, the dataset is divided into two subsets according to the rotation angle. COIL1 contains the images of all objects taken at angles of  $[0^\circ, 85^\circ]$  and  $[180^\circ, 265^\circ]$ ; and COIL2 contains the images of all objects taken in the directions of  $[90^\circ, 175^\circ]$  and  $[270^\circ, 355^\circ]$ . Each subset contains 720 images. Some samples are shown in Fig. 3a. The

images in COIL1 and COIL2 follow different distributions. In the experiments, one subset is selected as the source set and the other subset as the target set. Then, two pairs of domain adaptation are constructed.

Figure 4 shows that the proposed method is superior to other competing methods. In particular, the proposed method achieves the highest accuracies of **99.31%** and **99.03%** on COIL1→COIL2 and COIL2→COIL1, respectively. Compared with standard learning methods, the proposed method achieves improvements of approximately **14.0%** on both experiments. Compared with other DA methods, there are



**Fig. 6** Classification accuracies (%) on the USPS and MNIST datasets



at least **1.39%** and **0.14%** improvement in the two experiments, respectively, which demonstrate the effectiveness of the proposed unified framework incorporating non-convex Schatten  $p$ -norm, graph-structure constraint, statistical distributions, and classification error.

### 5.2 Office and Caltech-256 datasets

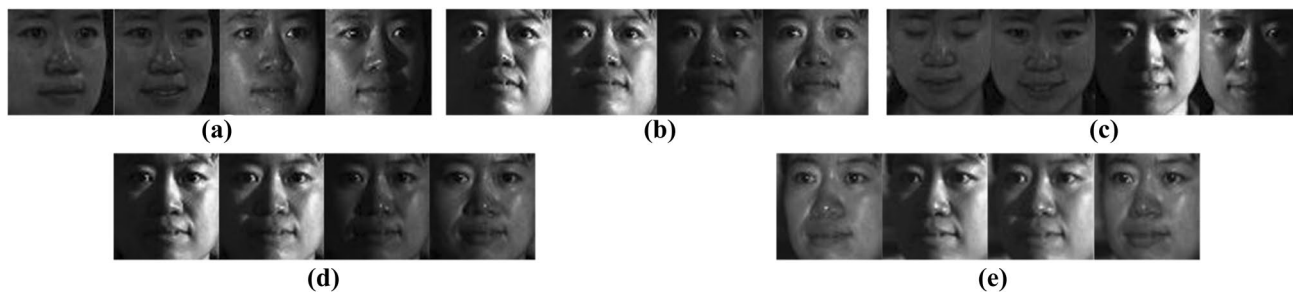
The Office dataset consists of 4652 images in 32 categories. These images are divided into three image groups: Amazon, Webcam and DSLR (denoted by A, W and D, respectively). The images in the three groups are obtained from the online merchant, a web camera with low-resolution, and a digital SLR camera with high-resolution, respectively. Different from the Office dataset, which is a standard DA benchmark, the Caltech-256 dataset (denoted by C) is a benchmark for image classification, which consists of more than 30,000 images of 256 categories. The two datasets share 10 classes, namely, backpack, bike, calculator, headphones, keyboard, laptop computer, monitor, mouse, mug and projector, as shown in Fig. 5. It can be seen that the variances of the images in the datasets are quite large. In the experiments, two different parts are randomly selected from the four parts

(i.e., A, W, D, and C) as the source and target domains, and in all 12 DA experiments are constructed, i.e.,  $C \rightarrow A$ ,  $C \rightarrow W$ , ...,  $D \rightarrow W$ .

Table 1 lists the performance of all competing methods. The proposed method achieves the highest average accuracy over the 12 cross-domain experiments and has at least **10.38%** (**0.85%**) improvement over the standard learning methods (DA methods). In the 12 cross-domain experiments, the proposed method achieves four best performances and three second-best performances.

### 5.3 USPS and MNIST datasets

The USPS dataset consists of 9298 images with a size of  $16 \times 16$  images in all, where 7291 images are for training and 2007 images for testing. The MNIST dataset includes 70,000 images with a size of  $28 \times 28$ , where 60,000 images are for training and 10,000 images for testing. The two datasets share 10 classes of digits but follow very different distributions. Some examples are shown in Fig. 3b. Following the settings of (Long et al. 2013b), two subsets are formed by randomly selecting 1800 and 2000 images from the USPS and MNIST, respectively. By selecting one subset as the



**Fig. 7** Some examples of the CMU PIE dataset. **a** Left pose, **b** upward pose, **c** downward pose, **d** front pose, **e** right pose

**Table 2** Classification accuracies (%) on the CMU PIE dataset

	GFK	TSL	TCA	LTSL	JDA	LRSR	LRDRM	Ours
1→2	26.15	44.08	26.70	22.96	58.81	65.87	62.67	<b>70.04</b>
1→3	27.27	47.49	27.08	20.65	54.23	64.09	67.21	<b>70.10</b>
1→4	31.15	62.78	31.06	31.81	84.50	82.03	83.08	<b>95.97</b>
1→5	17.59	36.15	18.08	12.07	49.75	54.90	58.39	<b>61.83</b>
2→1	25.24	46.28	26.14	18.25	57.62	45.04	55.61	<b>72.12</b>
2→3	47.37	57.60	47.98	16.05	62.93	53.49	55.88	<b>68.32</b>
2→4	54.25	71.43	54.73	45.15	75.82	71.43	77.56	<b>87.50</b>
2→5	27.08	35.66	28.06	17.52	39.89	47.97	47.67	<b>58.27</b>
3→1	21.82	36.94	21.91	22.36	50.96	52.49	60.05	<b>78.15</b>
3→2	43.16	47.02	43.65	20.26	57.95	55.56	46.83	<b>75.51</b>
3→4	46.41	59.45	47.67	57.34	68.45	77.50	76.60	<b>91.83</b>
3→5	26.78	36.34	27.57	24.57	39.95	54.11	56.86	<b>65.99</b>
4→1	34.24	63.66	33.82	51.20	80.58	81.54	85.47	<b>93.75</b>
4→2	62.92	72.68	64.52	70.10	82.63	85.39	84.89	<b>93.31</b>
4→3	73.35	83.52	74.08	72.00	87.25	82.23	84.00	<b>92.40</b>
4→5	37.38	44.79	38.91	48.28	54.66	72.61	73.77	<b>77.57</b>
5→1	20.35	33.28	20.35	13.06	46.46	52.19	59.63	<b>60.08</b>
5→2	24.62	34.13	24.98	21.61	42.05	49.41	52.73	<b>66.30</b>
5→3	28.49	36.58	28.86	17.03	53.31	58.45	53.79	<b>66.97</b>
5→4	31.33	38.75	31.36	29.59	57.01	64.31	71.76	<b>77.98</b>
Average	35.35	49.43	35.88	31.59	60.24	63.53	65.72	<b>76.20</b>

source set and the other subset as the target set, two DA experiments are constructed: USPS→MNIST and MNIST→USPS.

Figure 6 lists the classification accuracies of all methods. The proposed method achieves the best performance, with improvements of at least **0.27%** and **0.89%** on the two experiments. It should be noted that LTSL (Shao et al. 2014) and LRSR (Xu et al. 2016) perform better than most of the other DA methods, except BDA (Wang et al. 2017) and JDA (Long et al. 2013b). The main reason is that the background and structure of the digit image are simple. In this case, the materiality of the geometric structure is greater than that of the distribution difference. By taking both structure information and distribution shifts into consideration, SPDA and the proposed method achieve better performance than other competing methods. Compared with SPDA, the proposed method utilizes the Schatten  $p$ -norm to exploit the geometric structure more precisely and weigh the marginal distribution and conditional distribution differently to further minimize the domain shifts, which results in better performance.

#### 5.4 CMU PIE dataset

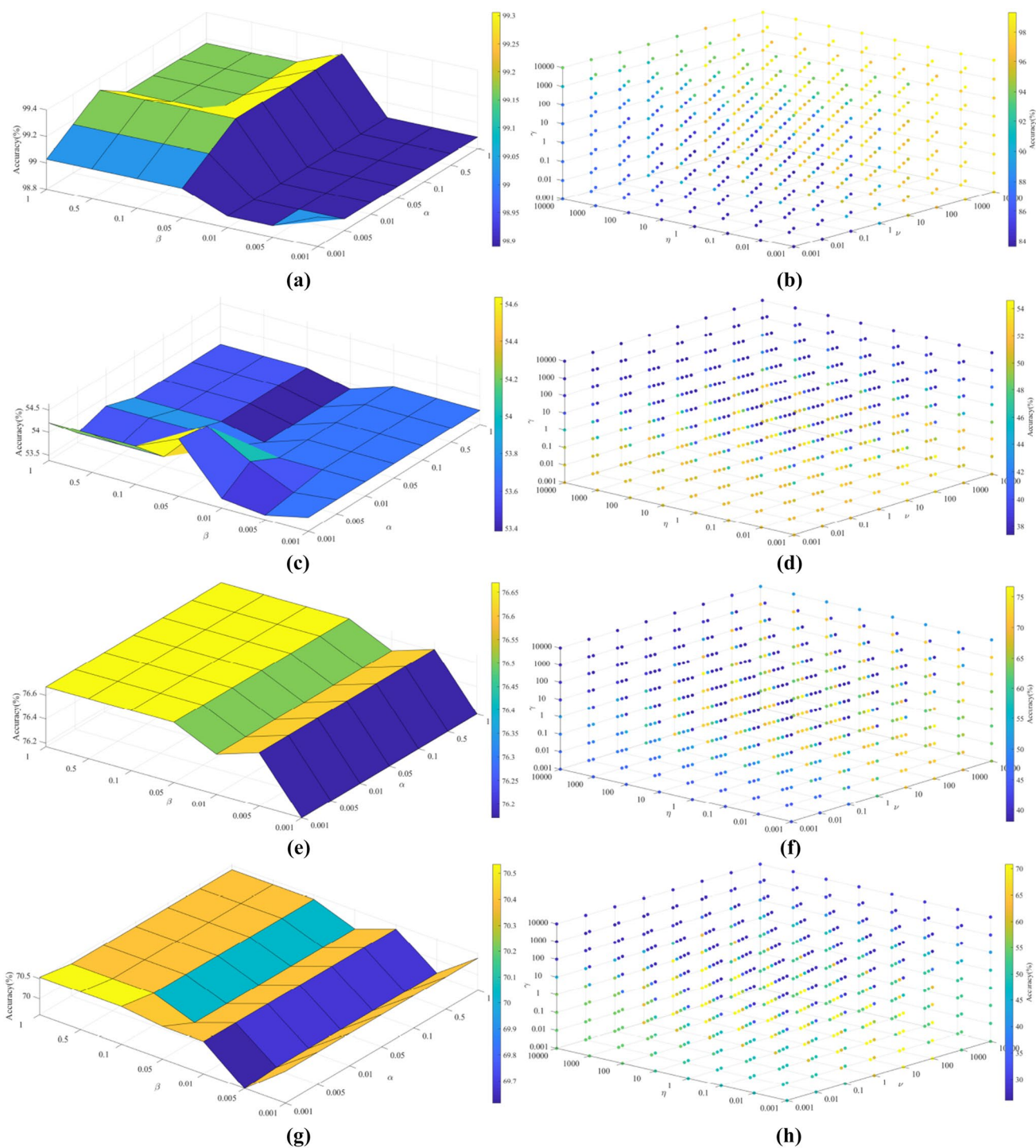
There are 41,368 face images with a size of  $32 \times 32$  in the CMU PIE dataset. The images are taken from 68 people under different illumination conditions, poses and expressions. In this experiment, five subsets of PIE are used to test different methods. Each subset corresponds to a distinct pose

with illumination and expression variations: PIE1 (C05, left pose, 3332 images, 49 images for each person), PIE2 (C07, upward pose, 1632 images, 24 images for each person), PIE3 (C09, downward pose, 1632 images, 24 images for each person), PIE4 (C27, front pose, 3332 images, 49 images for each person), and PIE5 (C29, right pose, 1632 images, 24 images for each person), as seen as in Fig. 7. Following the experimental settings in Xu et al. (2016), two different subsets are randomly chosen to be the source set and target set, which results in 20 different pairs of DA experiments, i.e.,  $1 \rightarrow 2, 1 \rightarrow 3, \dots, 5 \rightarrow 4$ .

Table 2 shows the classification accuracies of different methods. The proposed method outperforms other competing methods in all 20 cross-domain datasets and achieves **10.84%** improvement on the average accuracy. The main reason is that the proposed method uses both statistical distribution and geometric structure information, which reduces the distribution shifts between the source domain and the target domain further than other methods.

#### 5.5 Discussion of parameters

There are five parameters in the objective function (8):  $\alpha$  and  $\beta$  are  $l_1$  regularization on the representation and noise matrix, respectively.  $\gamma$ ,  $\eta$  and  $\nu$  are used to balance the graph regularization, marginal distribution and conditional distribution, respectively. To verify the impacts of the parameters, we calculate the results of the proposed method with



**Fig. 8** Classification accuracies (%) of the proposed method with different parameters on different datasets. **a**  $\alpha$  and  $\beta$  for COIL1→COIL2, **b**  $\gamma$ ,  $\eta$  and  $\nu$  for COIL1→COIL2, **c**  $\alpha$  and  $\beta$  for Caltech256→Amazon, **d**  $\gamma$ ,  $\eta$  and  $\nu$  for Caltech256→Amazon, **e**  $\alpha$  and  $\beta$  for MNIST→USPS, **f**  $\gamma$ ,  $\eta$  and  $\nu$  for MNIST→USPS, **g**  $\alpha$  and  $\beta$  for CMU PIE1→CMU PIE2, **h**  $\gamma$ ,  $\eta$  and  $\nu$  for CMU PIE1→CMU PIE2

different combinations of values. The values of each parameter are selected from a small set. Specifically, the parameters  $\alpha$  and  $\beta$  are searched in  $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$ ,

**d**  $\gamma$ ,  $\eta$  and  $\nu$  for Caltech256→Amazon, **e**  $\alpha$  and  $\beta$  for MNIST→USPS, **f**  $\gamma$ ,  $\eta$  and  $\nu$  for MNIST→USPS, **g**  $\alpha$  and  $\beta$  for CMU PIE1→CMU PIE2, **h**  $\gamma$ ,  $\eta$  and  $\nu$  for CMU PIE1→CMU PIE2

and the search ranges for the parameters  $\gamma, \eta$  and  $\nu$  are  $\{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$ .

The results for COIL1→COIL2, Caltech256→Amazon, USPS→MNIST, PIE1→PIE2 with different parameters are shown in Fig. 8. From Fig. 8, we can see that the classification accuracies are roughly consistent. Specifically, the method is insensitive to  $\alpha$  and  $\beta$  for all of the datasets. The accuracy range is within 1.0%. The values of  $\gamma$ ,  $\eta$  and  $\nu$  are different since the roles of their corresponding regularizations are diverse for different datasets.

## 5.6 Discussion with deep learning methods

Deep learning is a popular technology and has been applied in various applications (Yu et al. 2019; Lu et al. 2018, 2019). In Ding and Fu (2018), developed a deep transfer low-rank coding (DTLC) for cross-domain learning. Benefitting from the convolution neural networks, DTLC could capture more representative and discriminative image features. In Long et al. (2017), proposed joint adaptation networks to learn a transfer network by introducing the joint maximum mean discrepancy criterion and adversarial training strategy.

The image features play an essential role in the classification task. Intuitively, the proposed method will perform better if deep features are used, because deep features have a greater advantage in representation and discrimination than the handcrafted features. In the experiments, simply handcrafted features are used (SURF feature for the Office and Caltech256 datasets, grayscale pixel values for the other datasets). For practical applications, the deep features of images can be first extracted through trained deep neural networks, then used for unsupervised DA via (8).

## 6 Conclusions

This paper presented a novel unsupervised DA method for image classification. The method exploits both geometric and statistical characteristics of the samples by (1) constructing graph-structure constraint and the Schatten  $p$ -norm on the reconstruction matrix, (2) minimizing both marginal and conditional distributions, to reduce the distribution shift problem between the source and target domain. By projecting the samples of both domains into a common separable subspace, the samples from the target domain can be correctly classified. An iterative algorithm is proposed for effectively solving the proposed method. Extensive experiments on six datasets are done and the results verify the advantages of the proposed method.

**Acknowledgements** This work was partially supported by the NSFC under Grant Nos. 61806098, 61976118, 61972212, 61603192 and 61772568, the Natural Science Foundation of Jiangsu Province under Grant No. BK20190089 and BK20180142, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under

Grant Nos. 18KJB520029 and 17KJB520020, Nanjing Xiaozhuang University under Grant Nos. 2017NXY49.

## References

- Chang H, Luo L, Yang J, Yang M (2016) Schatten  $p$ -norm based principal component analysis. *Neurocomputing* 207:754–762
- Chang H, Zhang F, Gao G, Zheng H (2019) Structure-constrained discriminative dictionary learning based on Schatten  $p$ -norm for face recognition. *Digital Signal Process* 95
- Ding Z, Fu Y (2018) Deep transfer low-rank coding for cross-domain learning. *IEEE Trans Neural Netw Learn Syst* 30(6):1768–1779
- Ghifary M, Balduzzi D, Kleijn W, Zhang M (2017) Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Trans Pattern Anal Mach Intell* 39(7):1414–1430
- Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: *CVPR*, pp 2066–2073
- Kobylarz J, Bird JJ, Faria DR, Ribeiro EP, Ekárt A (2020) Thumbs up, thumbs down: non-verbal human-robot interaction through real-time emg classification via inductive and supervised transductive transfer learning. *J Ambient Intell Humaniz Comput* 1–11
- Lan R, Lu H, Zhou Y, Liu Z, Luo X (2019) An lbp encoding scheme jointly using quaternionic representation and angular information. *Neural Comput Appl* 1–7
- Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2013) Robust recovery of subspace structures by low-rank representation. *IEEE Trans Pattern Anal Mach Intell* 35(1):171–184
- Long M, Ding G, Wang J, Sun J, Guo Y, Yu PS (2013a) Transfer sparse coding for robust image representation. In: *CVPR*, pp 407–414
- Long M, Wang J, Ding G, Sun J, Yu PS (2013b) Transfer feature learning with joint distribution adaptation. In: *ICCV*, pp 2200–2207
- Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. In: *ICML*, pp 2208–2217
- Lu H, Li Y, Mu S, Wang D, Kim H, Serikawa S (2018) Motor anomaly detection for unmanned aerial vehicles using reinforcement learning. *IEEE Internet Things J* 5(4):2315–2322
- Lu H, Wang D, Li Y, Li J, Li X, Kim H, Serikawa S, Humar I (2019) Conet: a cognitive ocean network. *IEEE Wirel Commun* 26(3):90–96
- Nie F, Huang H, Ding C (2012) Low-rank matrix recovery via efficient Schatten  $p$ -norm minimization. In: *AAAI*, pp 655–661
- Razzaghi P, Razzaghi P, Abbasi K (2019) Transfer subspace learning via low-rank and discriminative reconstruction matrix. *Knowl Based Syst* 163:174–185
- Shao M, Kit D, Fu Y (2014) Generalized transfer subspace learning through low-rank constraint. *Int J Comput Vis* 109(1–2):74–93
- Si S, Tao D, Geng B (2010) Bregman divergence-based regularization for transfer subspace learning. *IEEE Trans Knowl Data Eng* 22(7):929–942
- Singh A, Nigam A (2019) Effect of identity mapping, transfer learning and domain knowledge on the robustness and generalization ability of a network: a biometric based case study. *J Ambient Intell Humaniz Comput* 11:1905–1922
- Singh R, Ahmed T, Singh R, Udmale SS, Singh SK (2019) Identifying tiny faces in thermal images using transfer learning. *J Ambient Intell Humaniz Comput* 11:1957–1966
- Sun B, Saenko K (2015) Subspace distribution alignment for unsupervised domain adaptation. In: *BMVC*, pp 24.1–10
- Wang J, Chen Y, Hao S, Feng W, Shen Z (2017) Balanced distribution adaptation for transfer learning. In: *ICDM*, pp 1129–1134
- Wang J, Feng W, Chen Y, Yu H, Huang M, Yu PS (2018a) Visual domain adaptation with manifold embedded distribution alignment. In: *ACM MM*, pp 402–410

- Wang J, Wang H, Gao G, Lu H, Zhang Z (2019) Single underwater image enhancement based on  $l_p$ -norm decomposition. *IEEE Access* 7:1145199–145213
- Wang L, Ding Z, Fu Y (2018b) Low-rank transfer human motion segmentation. *IEEE Trans Image Process* 28(3):1023–1034
- Wright J, Yang AY, Ganesh A, Ma Y (2009) Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell* 31(2):210–227
- Xiao T, Liu P, Zhao W, Liu H, Tang X (2019) Structure preservation and distribution alignment in discriminative transfer subspace learning. *Neurocomputing* 337:218–234
- Xu X, Lu H, Song J, Yang Y, Shen H, Li X (2019) Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Trans Cybern*
- Xu Y, Fang X, Wu J, Li X, Zhang D (2016) Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Trans Image Process* 25(2):850–863
- Yang J, Chu D, Zhang L, Xu Y, Yang J (2013) Sparse representation classifier steered discriminative projection with applications to face recognition. *IEEE Trans Neural Netw Learn Syst* 24(7):1023–1035
- Yang Z, Zhang H, Xu D, Zhang F, Yang G (2018) Double truncated nuclear norm-based matrix decomposition with application to background modeling. *J Ambient Intell Humaniz Comput* 1–10
- Yu Y, Tang S, Aizawa K, Aizawa A (2019) Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE Trans Neural Netw Learn Syst* 30(4):1250–1258
- Zhu D, Gao G, Gao H, Lu H (2018) Nuclear norm regularized structural orthogonal procrustes regression for face hallucination with pose. In: *International symposium on artificial intelligence and robotics*. Springer, pp 159–169

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.