Int J Comput Vis (2014) 109:94–109
DOI 10.1007/s11263-013-0693-1

Domain Adaptation for Face Recognition: Targetize Source Domain Bridged by Common Subspace

Meina Kan · Junting Wu · Shiguang Shan · Xilin Chen

Received: 15 March 2013 / Accepted: 7 December 2013 / Published online: 31 December 2013
© Springer Science+Business Media New York 2013

Abstract In many applications, a face recognition model learned on a source domain but applied to a novel target domain degenerates even significantly due to the mismatch between the two domains. Aiming at learning a better face recognition model for the target domain, this paper proposes a simple but effective domain adaptation approach that transfers the supervision knowledge from a labeled source domain to the unlabeled target domain. Our basic idea is to convert the source domain images to target domain (termed as *targetize the source domain* hereinafter), and at the same time keep its supervision information. For this purpose, each source domain image is simply represented as a linear combination of sparse target domain neighbors in the *image space*, with the combination coefficients however learnt in a *common subspace*. The principle behind this strategy is that, the common knowledge is only favorable for accurate cross-domain reconstruction, but for the classification in the target domain, the specific knowledge of the target domain is also essential and thus should be mostly preserved (through targetization in the image space in this work). To discover the common knowledge, specifically, a common subspace is learnt, in which the structures of both domains are preserved and meanwhile the disparity of source and target domains is reduced. The proposed method is extensively evaluated under three face recognition scenarios, i.e., domain adaptation across view angle, domain adaptation across ethnicity and domain adaptation across imaging condition. The experimental results illustrate the superiority of our method over those competitive ones.

Keywords Face recognition · Domain adaptation · Common subspace learning · Targetize the sourece domain

M. Kan · J. Wu · S. Shan (✉) · X. Chen
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology (ICT), CAS, Beijing 100190, China
e-mail: sgshan@ict.ac.cn

M. Kan
e-mail: kanmeina@ict.ac.cn

J. Wu
e-mail: junting.wu@vipl.ict.ac.cn

X. Chen
e-mail: xlchen@ict.ac.cn

1 Introduction

Machine learning has been widely used for various vision tasks, such as image classification, multimedia retrieval, and object recognition, etc. Typically, machine learning algorithms first learn a model from the training data, and then apply it to the test data. The learnt models generally work well when the training data and the test data are sampled from identical distribution, since in this case the training error is an optimistic estimate of the test error.

Unfortunately, in many real-world applications, it is difficult to collect training data that have the same distribution as testing data. Therefore, when the test and training data come from mildly or sometimes even wildly different distributions, most of the machine learning methods will seriously degenerate because the training error is no longer an optimistic estimate of the test error. For instance, the face recognition model trained from Mongolian can hardly be generalized to recognize Caucasian accurately. And if so, as seen from our empirical investigation shown in Fig. 1, the accuracy dramatically degenerates to 59 from 96 %, when we try to recognize Caucasian by using the model learnt from Mongolian, compared with the model from Caucasian.

To address this problem, one straightforward solution is collecting sufficient labeled data that can well describe the

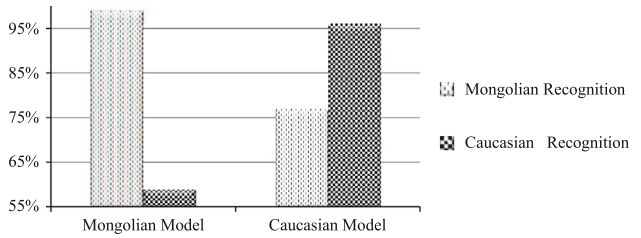


Fig. 1 An illustration of performance degeneration when applying the model learnt from one domain to a quite different domain. The Mongolian and Caucasian models are learnt by using Mongolian and Caucasian images respectively. The accuracy is 99 % when Mongolian model is used to recognize the Mongolian, while the accuracy is only 59 % when it is used to recognize Caucasian. Similarly, the accuracy is 96 % when we use Caucasian model to recognize Caucasian, while the accuracy is only 77 % when we use it to recognize the Mongolian

distribution of the test data, and then training a new model on these data. However, collecting and annotating sufficient data are very tedious and labor-intensive in practice.

To avoid the aforementioned tedious re-collection of new data, we can learn from the human vision system. As human beings, we indeed also suffer from cross-domain recognizing problem. For example, a Mongolian who never meet any Caucasian before may have difficulty in distinguishing the Caucasians. However, he/she can adapt to recognize them after facing limited number of Caucasians. What makes this possible is that much knowledge of distinguishing different people is common for Mongolian and Caucasian, and the human beings can adaptively transfer the recognition knowledge learnt from Mongolian to Caucasian.

A technique to analog the above procedure is domain adaptation (Blitzer et al. 2006; Uribe 2010), which tries to learn a better model for the target scenario/application, by *on one hand* borrowing some common knowledge from the source domain and *on the other hand* exploiting the particular information from the target domain but with limited supervision (e.g., unlabeled data or a limited number of labeled data). Domain adaptation is also more generally known as transfer learning (Pan and Yang 2010), which has been applied in various fields, e.g., natural language processing (Blitzer et al. 2006; Uribe 2010), computer vision (Uribe 2010; Gopalan et al. 2011; Jhuo et al. 2012; Gao et al. 2011; Hal 2009; Xue et al. 2007; Gong et al. 2012; Duan et al. 2012), etc.

In principle, domain adaptation attempts to transfer the rich knowledge in a supervised domain (termed as *source domain*) to another different but related domain with only limited information (termed as *target domain*) to induce a better model. Generally, the source domain and target domain share the same task but follow different distributions. According to whether or not the class labels are available for the training data of the target domain, domain adaptation can be categorized into two settings (Gopalan et al. 2011; Shi and Sha 2012), *supervised domain adaptation* and *unsupervised domain adaptation*. In scenario of supervised domain adaptation, labeled data (but generally in limited number) is available in the target domain, while in scenario of unsupervised domain adaptation only unlabeled data (but generally in large-scale) is available in the target domain. This work mainly concentrates on the unsupervised domain adaptation problem. Thus, in the following we only review the existing works focusing on the unsupervised domain adaptation.

In a simple scenario, the source domain shares the same support[1] as the target domain, but comprises a bias distribution due to the sampling strategy. Thus certain part of the source domain samples can be reused. An intuitive strategy is re-weighting the source domain samples to reduce the discrepancy between the sample distribution of the source domain and that of the target domain, e.g., sample selection bias (Zadrozny 2004; Huang et al. 2006), and particularly covariant shift (Shimodaira 2000; Sugiyamai et al. 2007; Sugiyama et al. 2008; Gretton et al. 2009; Bickel et al. 2009). Some previous works (Zadrozny 2004; Dudık et al. 2005) demonstrate how to assign the weights via estimating the probability densities of source and target domains. However, the probability density is difficult to estimate, especially in the high dimensional feature space. Hence, some of the subsequent works attempt to achieve the weights or importance without probability density estimation (Sugiyama et al. 2008).

Further, if the source and target domains only share part of the support or knowledge, it is conceivable to share part of model parameters or priors between both domains so as to shift the source domain model to the target domain (Bruzzone and Marconcini 2010; Duan et al. 2009; Chen et al. 2003; Xue et al. 2007). For instance, Chen et al. (2003) develop a progressive transductive support vector machine model to iteratively label and modify the unlabeled target domain, by means of pairwise labeling and dynamic adjusting to achieve a wider margin. The work (Wang et al. 2008) proposes to optimize a low-dimensional embedding space for the unlabeled target domain, with the information from both source and target domains embedded in the scatter matrices and the Laplacian graph. In the work (Bruzzone and Marconcini 2010), the discriminant function is adjusted step by step to the target domain by iteratively deleting the source domain samples and adding the target domain samples with estimated label until the final classification function is determined only based on target domain samples. Most of these methods exploit an iteration scheme to gradually adapt the knowledge of the source domain to the target domain. Therefore, the success

[1] In probability theory, the support of a probability distribution can be loosely thought of as the closure of the set of possible values of a random variable having that distribution. Here it can be regarded as the closure of the set of all possible instances.

of the adaptation heavily depends on whether the subsequent iteration can achieve a better model than the previous one.

When it is difficult to share parameters or priors between the source and target domains, an alternative scheme is exploring only the commonality of both domains, e.g., common feature representation or a common subspace for both domains, which can diminish the disparity between domains (Ben-David et al. 2007; Blitzer et al. 2006; Pan et al. 2011, 2009, 2008; Shi and Sha 2012; Qiu et al. 2012; Mehrotra et al. 2012; Si et al. 2010; Geng et al. 2011; Si et al. 2011; Raina et al. 2007; Shao et al. 2012). As a result, the distribution of the source domain and that of the target domain are in agreement in the common feature space, and thus the model trained on the labeled source domain can usually be used for target domain directly.

The key point of these approaches is how to determine the commonality of two domains. Some methods attempt to achieve it by explicitly reducing the difference between domains, while some other methods adopt a utility module for two domains, e.g., a unified dictionary.

To reduce the difference between domains and thus reach a common space, a few criterions are employed to measure the domain discrepancy, such as empirical maximum mean, mutual information, low-rank constrains and so on. In the work (Si et al. 2010), a transfer subspace learning framework is proposed to optimize a common subspace by minimizing the Bregman divergence between the distributions of two domains. In the work (Geng et al. 2011), a novel metric learning method is presented for domain adaptation by introducing a data-dependent regularizer that can minimize the empirical maximum mean discrepancy between domains in the reproducing kernel Hilbert space. In the work (Si et al. 2011), the authors calibrate the source and target distributions by minimizing the geodesic distance between two distributions that are represented as symmetric positive definite matrices in Riemannian symmetric space. In the work (Pan et al. 2011), a latent common space is obtained by finding a transform that can minimize the discrepancy between the marginal distributions of the source and target domains, and meanwhile preserve the data structure of the original space. In the work (Shi and Sha 2012), information theory is employed to obtain a common space, by maximizing the mutual information between the instances and the class labels and simultaneously minimizing the mutual information between the instances and the domain labels. In the work (Shao et al. 2012), a common subspace is achieved via a low-rank representation constraint, which attempts to ensure that each datum in source domain can be linearly represented by the samples in target domain.

To establish a utility module for two domains, several techniques are proposed, such as the pivot feature, or a unified dictionary. In the work (Blitzer et al. 2006), it is suggested to align the corresponding features of different domains by their relevance with the predefined pivot features. In the works (Raina et al. 2007; Mehrotra et al. 2012), it is considered that a unified dictionary learnt from the data of all domains can capture the commonality between domains, and the corresponding sparse representation is used as the common feature representation across domain. In the work (Gopalan et al. 2011), a common intermediate feature representation is achieved by projecting the data to a serial of subspaces sampled along the path between the source domain and target domain on the Grassmann manifold.

In most of the above common feature/subspace based methods, they only exploit the common characteristics of the source and target domains, while ignore the particular knowledge of the target domain that should be beneficial for the task in target domain. Therefore, this paper makes an attempt to achieve a better recognition model for the target domain, by exploiting not only the common knowledge shared between the source and target domains, but also the particular knowledge about the target domain. For this purpose, we propose to convert the labeled source domain samples to the target domain in the image space (termed as *Targetize the Source domain*), bridged by a common subspace. Hereinafter, we abbreviate our method as TSD.

Specifically, the image targetization is achieved in the image space by representing each source domain image as a linear combination of sparse target domain neighbors, while the combination coefficients are determined in a common subspace. After the targetization, the class label of each targetized sample is the same as that of the original source domain sample. Therefore, we actually have generated a labeled "virtual" target domain via the targetization, which makes it possible to learn a supervised classifier for the target domain. In addition, for the common knowledge learning, we aim at learning a common subspace, where the structure of each domain is preserved and meanwhile the source and target data are enforced to interlace sufficiently.

The reminder of this paper is organized as follows. Section 2 presents our TSD method, followed by a detailed description about the optimization in Section 3. In Section 4, we evaluate the proposed method on three domain adaptation face recognition scenarios, i.e., domain adaptation across view angle, domain adaptation across ethnicity and domain adaptation across imaging condition. Finally, a conclusion is given in the last section.

2 Targetize Source Domain Bridged by Common Subspace

In this section, we first describe the proposed TSD method in detail, and then give some discussions about the differences from the related works.

2.1 Notations

For clarity, we first formally define some notations. In the whole text, upper-case and lower-case characters represent the matrices and vectors respectively.

The data matrix of the training images from source domain is denoted as $\mathbf{X}_s = [\mathbf{x}_1^s, \mathbf{x}_2^s, \cdots, \mathbf{x}_{n_s}^s] \in \mathbb{R}^{d_s \times n_s}$, with their class labels $\mathbf{y}_s = [y_1^s, y_2^s, \cdots, y_{n_s}^s], y_i^s \in \{1, 2, \cdots, c_s\}$, where $\mathbf{x}_i^s|_{i=1}^{n_s} \in \mathbb{R}^{d_s \times 1}$ is the feature representation of the i-th source domain sample of d_s dimension, n_s is the number of training samples in the source domain, y_i^s is the class label of the i-th source domain sample, and c_s is the number of classes in the source domain.

Similarly, the data matrix of the training data from target domain is denoted as $\mathbf{X}_t = [\mathbf{x}_1^t, \mathbf{x}_2^t, \cdots, \mathbf{x}_{n_t}^t] \in \mathbb{R}^{d_t \times n_t}$ without class label, where $\mathbf{x}_i^t|_{i=1}^{n_t} \in \mathbb{R}^{d_t \times 1}$ is the feature representation of the i-th target domain sample of d_t dimension, and n_t is the number of samples in the target domain.

Unless otherwise specified, the symbols s and t used in the superscript or subscript mean the source domain and target domain respectively.

2.2 Domain Adaptation via TSD

Given a labeled training set of source domain, i.e., $(\mathbf{X}_s, \mathbf{y}_s)$, it can be readily used to learn a discriminant model favorable for the task in the source domain. However, what we desire is a discriminant model for the target domain where only the unlabeled training data $\mathbf{X}_t$ is available. As mentioned before, although the model learnt from source domain is not applicable for the target domain, the supervision knowledge of source domain must be beneficial for the model learning in the target domain (Pan and Yang 2010; Blitzer et al. 2006; Uribe 2010).

Following this idea, we attempt to transfer the knowledge of source domain to the target domain by directly converting the source domain samples to the target domain, along with their class labels. As a result, the labeled targetized source domain has similar distribution as the target domain, and can be naturally employed for any further supervised model learning, e.g., Fisher's Linear Discriminant Analysis.

In TSD, each source domain sample is targetized to the target domain by being reconstructed via sparse neighbors from target domain. However the source domain and target domain might be distant from each other, which makes the sparse reconstruction in the original sample space very difficult (if not impossible). Therefore, we propose to first project the samples from both domains to a common subspace, and then reconstruct each source domain sample by using only a limited number of neighbors from the target domain in the common subspace. After having the sparse reconstruction coefficients, we finally apply them in the original sample space to convert the source domain samples to the target domain, so as to keep the particular information of the target domain. An overall schema of the proposed framework is shown in Fig. 2.

This subsection is organized as follows: we first formulate how to learn the common subspace and the sparse coeffi-

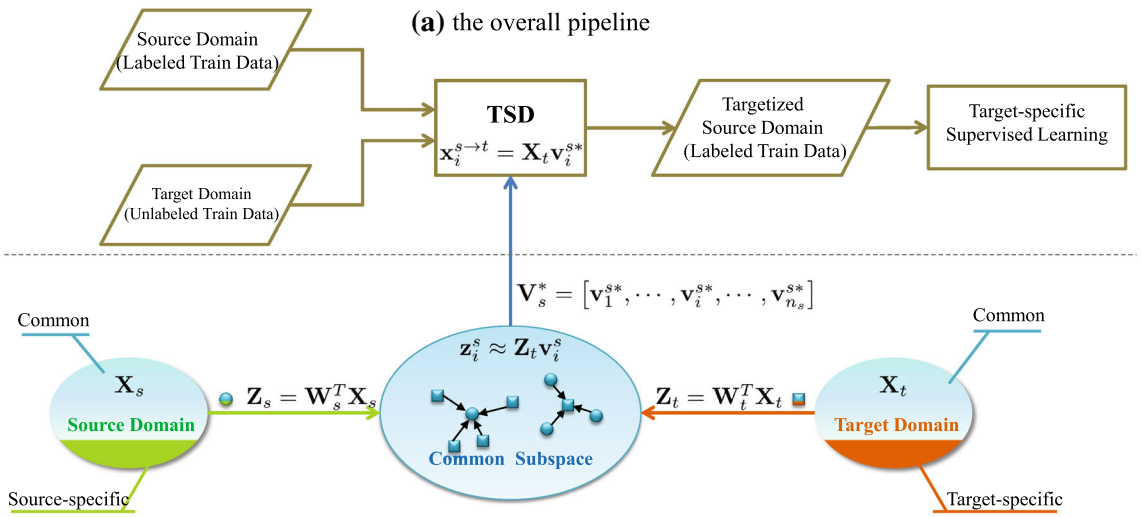


Fig. 2 The overall schema of the proposed method: (**a**) the overall pipeline of our domain adaptation method: first targetize each source domain sample to the target domain through a linear combination of sparse target domain neighbors, and then conduct the supervised learning on the targetized source domain; (**b**) the targetization coefficients used in (**a**) are learnt in a common subspace rather than in the original sample space

cients, followed by the targetization based on the learnt sparse coefficients. Then, we present how the targetized source domain is used to learn a supervised target model and how to do classification in the target domain.

2.2.1 Common Subspace Learning

One of the key components of our TSD is the common subspace, where the source and target domains are sufficiently interlaced. Since two different domains may have large discrepancy or even in different space, we propose to learn two coupled projections, denoted as $\mathbf{W}_s$ and $\mathbf{W}_t$, to respectively project the source and target domain into a common subspace. The projected source and target domain samples in the common subspace are denoted as $\mathbf{Z}_s = [\mathbf{z}_1^s, \mathbf{z}_2^s, \cdots, \mathbf{z}_{n_s}^s]$ and $\mathbf{Z}_t = [\mathbf{z}_1^t, \mathbf{z}_2^t, \cdots, \mathbf{z}_{n_t}^t]$ respectively:

$$\mathbf{z}_i^s = \mathbf{W}_s^T \mathbf{x}_i^s, i = 1, 2, \cdots, n_s, \text{ or } \mathbf{Z}_s = \mathbf{W}_s^T \mathbf{X}_s, \quad (1)$$

$$\mathbf{z}_i^t = \mathbf{W}_t^T \mathbf{x}_i^t, i = 1, 2, \cdots, n_t, \text{ or } \mathbf{Z}_t = \mathbf{W}_t^T \mathbf{X}_t. \quad (2)$$

Briefly summarizing the previous discussions, the common subspace should satisfy the following requirements:

(1) In the common subspace, the samples from both domains should be well interlaced, so as to reduce the discrepancy of the source and target domains.
(2) In the common subspace, the structures of both domains should be well preserved, in order to keep enough discriminant information for further model learning.

To meet the above two requirements, two terms, *Sparse Reconstruction* and *Max-Variance*, are respectively proposed and described as bellow, followed by their combination to form the overall objective function.

Sparse Reconstruction: Sparse representation or sparse coding (Wright et al. 2009; Huang and Aviyente 2007) has been widely used for reconstruction and classification. It attempts to reconstruct a signal or a sample $\mathbf{x}$ by using only a limited number of samples from a basis set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ as follows:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \|\mathbf{x} - \mathbf{X}\mathbf{v}\|^2, \quad s.t. \|\mathbf{v}\|_0 \leq \tau, \quad (3)$$

where $\|\mathbf{v}\|_0$ is the number of non-zero entries in $\mathbf{v}$, also the number of samples selected from $\mathbf{X}$ for the reconstruction. Usually, $\|\mathbf{v}\|_0$ is required to be a small number, so the selected samples corresponding to the non-zero entries are most likely to be the neighbors of $\mathbf{x}$ to ensure a small reconstruct error. Therefore, if each sample from one domain is reconstructed by using only sparse adjcent but distinct samples from the other domain and vice versa, the samples from both domains are well interlaced with each other. Specifically, in the common subspace each source domain sample should be sparsely reconstructed by using only a few adjacent samples of target domain as bellow:

$$\begin{aligned}\left[\mathbf{V}_s^*, \mathbf{W}_s^*, \mathbf{W}_t^*\right] &= \arg\min_{\mathbf{V}_s, \mathbf{W}_s, \mathbf{W}_t} \|\mathbf{Z}_s - \mathbf{Z}_t \mathbf{V}_s\|_F^2 \\ &= \arg\min_{\mathbf{V}_s, \mathbf{W}_s, \mathbf{W}_t} \|\mathbf{W}_s^T \mathbf{X}_s - \mathbf{W}_t^T \mathbf{X}_t \mathbf{V}_s\|_F^2, \\ &s.t. \|\mathbf{v}_i^s\|_0 \leq \tau, i = 1, 2, \cdots, n_s, \end{aligned} \quad (4)$$

with $\mathbf{V}_s = [\mathbf{v}_1^s, \mathbf{v}_2^s, \cdots, \mathbf{v}_{n_s}^s]$. Similarly, in the common subspace each target domain sample should be sparsely reconstructed via only several neighbors from source domain:

$$\begin{aligned}\left[\mathbf{V}_t^*, \mathbf{W}_s^*, \mathbf{W}_t^*\right] &= \arg\min_{\mathbf{V}_t, \mathbf{W}_s, \mathbf{W}_t} \|\mathbf{Z}_t - \mathbf{Z}_s \mathbf{V}_t\|_F^2 \\ &= \arg\min_{\mathbf{V}_t, \mathbf{W}_s, \mathbf{W}_t} \|\mathbf{W}_t^T \mathbf{X}_t - \mathbf{W}_s^T \mathbf{X}_s \mathbf{V}_t\|_F^2, \\ &s.t. \|\mathbf{v}_i^t\|_0 \leq \tau, i = 1, 2, \cdots, n_t, \end{aligned} \quad (5)$$

with $\mathbf{V}_t = [\mathbf{v}_1^t, \mathbf{v}_2^t, \cdots, \mathbf{v}_{n_t}^t]$. In both Eqs. (4) and (5), τ is the parameter to control the sparsity in terms of the number of samples used for the reconstruction.

Max-Variance: Besides reducing the discrepancy of the source and target domains, in the common subspace the structures of both domains are also expected to be preserved, so as to keep as much information as possible for the discrimination. At the first thought, some locality or neighborhood preserving methods, e.g., LLE (Roweis and Saul 2000), LPP (He and Niyogi 2004), would be a good choice. However, for our task, we actually need to balance the global and local structures, since both are important for building a discriminative recognition model. Methods like LLE or LPP cannot guarantee to keep the global structure, which might hurt the classification. Therefore, following the works (Pan et al. 2011, 2008), the structure is preserved just by simply maximizing the variance of each domain as follows:

$$\mathbf{W}_s^* = \arg\max_{\mathbf{W}_s} \mathrm{Tr}\left(\mathbf{W}_s^T \mathbf{X}_s \mathbf{X}_s^T \mathbf{W}_s\right), \quad (6)$$

$$\mathbf{W}_t^* = \arg\max_{\mathbf{W}_t} \mathrm{Tr}\left(\mathbf{W}_t^T \mathbf{X}_t \mathbf{X}_t^T \mathbf{W}_t\right), \quad (7)$$

where $\mathrm{Tr}(\cdot)$ means the trace of a matrix, $\mathbf{X}_s$ and $\mathbf{X}_t$ are assumed to have zero mean. In addition, the above max-variance terms are also necessary to ensure the sparse reconstruction feasible. Without them, all the samples from both domains may converge together, thus degenerating to a meaningless solution.

Overall Objective Function: Our goal is to obtain a common subspace, in which the above two requirements are simultaneously satisfied. For this purpose, we combine them together by formulating it as a Fisher criterion-like objective function as bellow:

$$\left[\mathbf{W}_s^*, \mathbf{W}_t^*, \mathbf{V}_s^*, \mathbf{V}_t^*\right] = \arg\max_{\mathbf{W}_s, \mathbf{W}_t, \mathbf{V}_s, \mathbf{V}_t} \frac{\mathrm{Tr}\left(\frac{1}{n_s}\mathbf{W}_s^T\mathbf{X}_s\mathbf{X}_s^T\mathbf{W}_s + \frac{1}{n_t}\mathbf{W}_t^T\mathbf{X}_t\mathbf{X}_t^T\mathbf{W}_t\right)}{\frac{1}{n_s}\|\mathbf{W}_s^T\mathbf{X}_s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{V}_s\|_F^2 + \frac{1}{n_t}\|\mathbf{W}_t^T\mathbf{X}_t - \mathbf{W}_s^T\mathbf{X}_s\mathbf{V}_t\|_F^2} \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, \|\mathbf{v}_j^t\|_0 \le \tau, i = 1, \cdots, n_s, j = 1, \cdots, n_t. \tag{8}$$

From this formulation, it is clear that we actually jointly optimize the common subspace and the sparse reconstruction coefficients. How to optimize the objective in Eq. (8) will be presented in detail in Section 3.

2.2.2 Targetizing the Source Domain in Image Space

After optimizing Eq. (8), we can achieve the common subspace in terms of $\left[\mathbf{W}_s^*, \mathbf{W}_t^*\right]$, and the sparse reconstruction relationship in terms of $\left[\mathbf{V}_s^*, \mathbf{V}_t^*\right]$. As mentioned, existing works generally then learn the target model directly in the common subspace, since the source and target domains have similar distribution in it. In this way, however, the particular knowledge of the target domain is not exploited, which will degrade the performance in target domain. So, we instead perform the targetization in the original sample space so as to keep the particular knowledge of the target domain.

Formally, the targetized source domain, denoted as $\mathbf{X}_{s\to t} = \left[\mathbf{x}_1^{s\to t}, \mathbf{x}_2^{s\to t}, \cdots, \mathbf{x}_{n_s}^{s\to t}\right] \in \mathbb{R}^{d_t \times n_s}$, is achieved by applying the sparse coefficients $\mathbf{V}_s^* = \left[\mathbf{v}_1^{s*}, \mathbf{v}_2^{s*}, \cdots, \mathbf{v}_{n_s}^{s*}\right]$ obtained from Eq. (8) in the original sample space:

$$\mathbf{x}_i^{s\to t} = \mathbf{X}_t\mathbf{v}_i^{s*}, i = 1, 2, \cdots, n_s, \tag{9}$$

which can be equivalently written in matrix form as:

$$\mathbf{X}_{s\to t} = \mathbf{X}_t\mathbf{V}_s^*. \tag{10}$$

Meanwhile, the class label of each source domain sample y_s is kept for the transformed sample $\mathbf{x}_{s\to t}$. So, in theory, any supervised method can be used to learn a recognition model based on $(\mathbf{X}_{s\to t}, \mathbf{y}_s)$.

2.3 Supervised Model Learning and Testing

2.3.1 Supervised Model Learning in the Target Domain

Fisher's Linear Discriminant (FLD) analysis (Belhumeur et al. 1997) is a widely used approach for discriminative feature extraction, and face recognition (Liu and Wechsler 2002; Su et al. 2009). In this work, we employ FLD for supervised feature extraction:

$$\mathbf{W}_{fld}^* = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T\mathbf{S}_b\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_t\mathbf{W}|}, \tag{11}$$

where $|\cdot|$ means the determinant of a matrix. The total scatter matrix $\mathbf{S}_t$ is calculated from the target domain while the between-class scatter matrix $\mathbf{S}_b$ is calculated from the targetized source domain as:

$$\mathbf{S}_t = \sum\nolimits_{i=1}^{n_t}\left(\mathbf{x}_i^t - \boldsymbol{\mu}^t\right)\left(\mathbf{x}_i^t - \boldsymbol{\mu}^t\right)^T, \tag{12}$$

$$\mathbf{S}_b = \sum\nolimits_{j=1}^{c_s} n_{s_j}\left(\boldsymbol{\mu}_j^{s\to t} - \boldsymbol{\mu}^{s\to t}\right)\left(\boldsymbol{\mu}_j^{s\to t} - \boldsymbol{\mu}^{s\to t}\right)^T, \tag{13}$$

where $\boldsymbol{\mu}^t = \sum_{i=1}^{n_t}\frac{1}{n_t}\mathbf{x}_i^t$ is the mean of $\mathbf{X}_t$, $\boldsymbol{\mu}_j^{s\to t} = \frac{1}{n_{s_j}}\sum_{y_k^s=j}\mathbf{x}_k^{s\to t}$ is the mean of the j-th class of the targetized source domain, n_{s_j} is the number of samples in the j-th class, $\boldsymbol{\mu}^{s\to t} = \sum_{j=1}^{n_s}\frac{1}{n_s}\mathbf{x}_j^{s\to t}$ is the mean of $\mathbf{X}_{s\to t}$, and c_s is the number of classes of the source domain. With Eqs. (12) and (13), Eq. (11) can be analytically solved by using generalized eigenvalue decomposition.

Here, please kindly note that no class label is needed for the computation of total scatter matrix, so it is intuitively more plausible to directly use the unsupervised target data for computing $\mathbf{S}_t$. Nevertheless, it is also feasible to use the targetized source domain, which is actually a re-sampled target domain data. Our experiments show only very trivial difference for these two alternatives.

2.3.2 Classification in Target Domain

Given a supervised source domain and an unsupervised target domain, our TSD can output a discriminant feature extractor $\mathbf{W}_{fld}^*$ in Eq. (11) for the target domain. Then, to perform face recognition in target domain, a testing set composed of a gallery and a probe set is needed. The gallery contains the enrolled face images with known identities, while the probe set includes face images to be identified.

To do testing, the discriminative feature of each gallery image $\mathbf{x}_i^g$ is firstly extracted as $\mathbf{W}_{fld}^{*T}\mathbf{x}_i^g$. Similarly, for each probe image $\mathbf{x}^{test}$, its discriminative feature is extracted as $\mathbf{W}_{fld}^{*T}\mathbf{x}^{test}$. Next, its similarity to each gallery face image is calculated via cosine function as $sim(\mathbf{x}^{test}, \mathbf{x}_i^g) = cosine(\mathbf{W}_{fld}^{*T}\mathbf{x}^{test}, \mathbf{W}_{fld}^{*T}\mathbf{x}_i^g)$. Finally, the identity of $\mathbf{x}^{test}$ is determined by using the Nearest Neighbor classifier as $Label(\mathbf{x}^{test}) = Label(\mathbf{x}_k^g), k = \arg\max_i sim(\mathbf{x}^{test}, \mathbf{x}_i^g)$.

2.4 Discussion

2.4.1 Targetization in the Original Image Space

As mentioned before, the targetized source domain $\mathbf{X}_{s\to t}$ in Eq. (10) is expected to contain not only the common characteristics of both domains, but also the particular knowledge of target domain. To further illustrate this point, we formally denote the commonality of $\mathbf{X}_s$ and $\mathbf{X}_t$ in the original image subspace as $\mathbf{X}_s^{(c)} = \mathbf{W}_s^*\mathbf{W}_s^{*T}\mathbf{X}_s$ and $\mathbf{X}_t^{(c)} = \mathbf{W}_t^*\mathbf{W}_t^{*T}\mathbf{X}_t$ respectively. Note that $\mathbf{X}_s^{(c)}$ and $\mathbf{X}_t^{(c)}$ only capture the common part of the source and target domains.

For the target samples, their common part shared with the source domain are removed, the remaining part, i.e.,

$\left(\mathbf{X}_t - \mathbf{X}_t^{(c)}\right)\mathbf{V}_s^*$ contains the particular information of the target domain, which can be utilized to build a better discriminant model for the target domain. The following rewriting of Eq. (10) can further reveal this point:

$$\mathbf{X}_{s\to t} = \mathbf{X}_t\mathbf{V}_s^* = \underbrace{\mathbf{X}_t^{(c)}\mathbf{V}_s^*}_{common} + \underbrace{\left(\mathbf{X}_t - \mathbf{X}_t^{(c)}\right)\mathbf{V}_s^*}_{paticular} \tag{14}$$

The commonality of the source and target domains in Eq. (14) bridges the gap between the two domains to determine the reconstruction relationship for them, and after that the particular information of the target domain is recalled during the reconstruction, which is equivalent to the reconstruction in the original image space.

2.4.2 Difference from the Existing Works

Difference with LRR (Jhuo et al. 2012)

To our best knowledge, the LRR is the most similar one to ours. However, our TSD is quite different from LRR in two-fold: (1) In LRR, only the source domain samples are projected to an intermediate subspace, which can pull the source domain closer but not enough to the target domain. In contrast, by removing the particular parts of both domains (not just that of the source domain), our TSD projects both of them into a common subspace, which can make the source domain and target domain closer enough to each other. (2) In LRR, the source domain samples lose the individuality when being transformed to the target domain due to the low-rank constraint, while in our TSD, each source domain sample is transformed independently which can preserve the diversity of different subjects.

Difference with LTSL (Shao et al. 2012)

In LTSL, a unified transform is exploited to project both source and target domains into a common subspace with a low-rank constraint, and the classification is completed in the common subspace. On the contrary, two transforms are exploited to respectively project the source domain and target domain to a common subspace which can make the domain discrepancy smaller. Another advantage of TSD is that the targetization and the classification are both conducted in the original sample space, which can preserve more information about the target domain leading to a better recognition model for the target domain.

Difference from Other Existing Works

Our method shares some common idea with the works (Mehrotra et al. 2012; Si et al. 2010; Geng et al. 2011; Si et al. 2011; Raina et al. 2007) in using common space. However, these methods exploit single transform to capture the commonality between domains, in terms of unified dictionary, distance metric, or projection. Specifically, our TSD method is different from them in the following aspects: (1) Our TSD method employs two transforms to respectively project the source and target domains into a common subspace, which can make the source and target domains closer to each other. (2) Besides using the commonality, the TSD method further exploits the particular information of the target domain, leading to a better recognition model for the target domain. In other perspective, the common subspace in our TSD is only used as a bridge to link the domains, while the target model is learnt in the original sample space, which is quite different from the above methods. (3) The measurement of the discrepancy between the source and target domains is different. In the works (Si et al. 2010; Geng et al. 2011), the Bregman divergence and the empirical maximum mean discrepancy are used. They are all parameter-based measurements, while the sparse reconstruction used in our TSD are more flexible as a non-parameter measurement.

3 Optimization

As the optimization problem of Eq. (8) is not convex for all variables, we exploit an alternation method to iteratively solve the projection matrices ($\mathbf{W}_s$, $\mathbf{W}_t$) and the sparse reconstruction coefficients ($\mathbf{V}_s$, $\mathbf{V}_t$).

Step 1: given $\mathbf{W}_s$ and $\mathbf{W}_t$, optimize $\mathbf{V}_s$ and $\mathbf{V}_t$.

As seen from Eq. (8), $\mathbf{V}_s$ and $\mathbf{V}_t$ are independent of each other given $\mathbf{W}_s$ and $\mathbf{W}_t$, so they can be optimized independently. Namely, $\mathbf{V}_s$ is optimized as:

$$\mathbf{V}_s^* = \arg\min_{\mathbf{V}_s} \|\mathbf{W}_s^T\mathbf{X}_s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{V}_s\|_F^2, \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, i = 1, 2, \cdots, n_s. \tag{15}$$

Furthermore, $\mathbf{v}_1^s, \mathbf{v}_2^s, \cdots, \mathbf{v}_{n_s}^s$ in Eq. (15) are also independent of each other, which means each $\mathbf{v}_i^s$ can be separately solved as a lasso problem:

$$\mathbf{v}_i^{s*} = \arg\min_{\mathbf{v}_i^s} \|\mathbf{W}_s^T\mathbf{x}_i^s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{v}_i^s\|^2, \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, i = 1, 2, \cdots, n_s. \tag{16}$$

However, if each of the above lasso problems is solved independently, some source domain samples may make use of the same target samples for the sparse reconstruction. For example, in the experiments, we found that the samples of the same subject in one domain may select the same single sample from the other domain, which means the targetized source samples of the same class may lose its intra-personal variations. Therefore, to preserve the diversity of the source domain and guarantee that the source domain and target domain are interlaced sufficiently, an additional penalty term is added to Eq. (16) to enhance the preference of varied samples for the reconstruction as bellow:

$$\mathbf{v}_i^{s*} = \arg\min_{\mathbf{v}_i^s} \|\mathbf{W}_s^T\mathbf{x}_i^s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{v}_i^s\|^2 + \lambda\|1 - \mathbf{h}_t^T\mathbf{v}_i^s\|^2, \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, i = 1, 2, \cdots, n_s. \tag{17}$$

Input: the source domain samples $\mathbf{X}_s$ and the target domain samples $\mathbf{X}_t$.
Output: the projection matrices $\mathbf{W}_s^*$, $\mathbf{W}_t^*$; the sparse reconstruction coefficients $\mathbf{V}_s^*$, $\mathbf{V}_t^*$; and the targetized source domain $\mathbf{X}_{s\to t}$.
1. Initialize $\mathbf{W}_s$, $\mathbf{W}_t$ as random matrices with the norm of each column as 1;
2. **While** the changes of variables is larger than ε and not reach a maximum number of iterations **do**
  2.1 Optimize the sparse reconstruction coefficients $\mathbf{V}_s$ and $\mathbf{V}_t$ according to Eq. (18) and Eq. (21):
    2.1.1 for i=1:n_s
      $\mathbf{v}_i^{s*} = \arg\min_{\mathbf{v}_i^s} \|\mathbf{W}_s^T\mathbf{x}_i^s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{v}_i^s\|^2 + \lambda\|1 - \mathbf{h}_t^T\mathbf{v}_i^s\|^2, s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau.$
      $\mathbf{h}_t \Leftarrow \mathbf{h}_t - \frac{0.5}{\max(|\mathbf{v}_i^{s*}|)}|\mathbf{v}_i^{s*}|.$
    end
    2.1.2 for i=1:n_t
      $\mathbf{v}_i^{t*} = \arg\min_{\mathbf{v}_i^t} \|\mathbf{W}_t^T\mathbf{x}_i^t - \mathbf{W}_s^T\mathbf{X}_s\mathbf{v}_i^t\|^2 + \lambda\|1 - \mathbf{h}_s^T\mathbf{v}_i^t\|^2, s.t.\ \|\mathbf{v}_i^t\|_0 \le \tau.$
      $\mathbf{h}_s \Leftarrow \mathbf{h}_s - \frac{0.5}{\max(|\mathbf{v}_i^{t*}|)}|\mathbf{v}_i^{t*}|.$
    end
  2.2 Optimize the two projection matrices $\mathbf{W}_s$ and $\mathbf{W}_t$ according to Eq. (27):
    $\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathrm{Tr}\left(\frac{\mathbf{W}^T\boldsymbol{\Sigma}_b\mathbf{W}}{\mathbf{W}^T\boldsymbol{\Sigma}_w\mathbf{W}}\right), where \mathbf{W} = [\mathbf{W}_s^T\ \mathbf{W}_t^T]^T.$
  end
3. Calculate the targetized source domain: $\mathbf{X}_{s\to t} = \mathbf{X}_t\mathbf{V}_s^*$;
4. Apply FLD on the targetized $(\mathbf{X}_{s\to t}, \mathbf{Y}_s)$ according to Eq. (11).

Algorithm 1: Targetize source domain bridged by common subspace learning.

Here, the indicator $\mathbf{h}_t \in \mathbb{R}^{n_t\times 1}$ assigns a weight to each target sample, indicating how many times the sample has been used. So, the penalty term will punish those target samples that are over-used to reconstruct the source samples via a small weight. As a result, it enforces different source domain samples to select distinct target domain samples, which can facilitate the preserving of the intra-personal variations for the targetized source domain. In our implementation, $\mathbf{h}_t$ is initialized as an all-1 vector, meaning that all samples have equal availability in the subsequent reconstruction.

The problem in Eq. (17) can be easily solved by using a forward stepwise regression like algorithm, i.e., the Least angle regression (Efron et al. 2004; Donoho 2006), by being reformulated as the following form:

$$\mathbf{v}_i^{s*} = \arg\min_{\mathbf{v}_i^s} \|\tilde{\mathbf{z}}_i^s - \tilde{\mathbf{Z}}_t\mathbf{v}_i^s\|^2, \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, i = 1, 2, \cdots, n_s, \tag{18}$$

with $\tilde{\mathbf{z}}_i^s = \begin{bmatrix}\mathbf{W}_s^T\mathbf{x}_i^s \\ \sqrt{\lambda}\end{bmatrix}$ and $\tilde{\mathbf{Z}}_t = \begin{bmatrix}\mathbf{W}_t^T\mathbf{X}_t \\ \sqrt{\lambda}\mathbf{h}_t^T\end{bmatrix}$.

After achieving the $\mathbf{v}_i^{s*}$, those samples selected for the reconstruction are punished by being decreased the weights as follows:

$$\mathbf{h}_t \Leftarrow \mathbf{h}_t - \frac{0.5}{\max(|\mathbf{v}_i^{s*}|)}|\mathbf{v}_i^{s*}|. \tag{19}$$

$|\mathbf{v}_i^{s*}|$ means the absolute operation on each element of the vector $\mathbf{v}_i^{s*}$. The sparsity parameter τ in Eq. (18) can ensure that only a limited number of entries in $\mathbf{v}_i^{s*}$ are non-zero. As a result, Eq. (19) decreases the availability of those samples selected for reconstruction in this turn, while keeps the availability of the rest ones unchanged.

Similarly, each $\mathbf{v}_i^t$ in $\mathbf{V}_t$ is also independent of each other and can be separately optimized as:

$$\mathbf{v}_i^{t*} = \arg\min_{\mathbf{v}_i^t} \|\mathbf{W}_t^T\mathbf{x}_i^t - \mathbf{W}_s^T\mathbf{X}_s\mathbf{v}_i^t\|^2 + \lambda\|1 - \mathbf{h}_s^T\mathbf{v}_i^t\|^2, \quad s.t.\ \|\mathbf{v}_i^t\|_0 \le \tau, i = 1, 2, \cdots, n_t, \tag{20}$$

which can be easily solved by using the Least angle regression by further being reformulated as:

$$\mathbf{v}_i^{t*} = \arg\min_{\mathbf{v}_i^t} \|\tilde{\mathbf{z}}_i^t - \tilde{\mathbf{Z}}_s\mathbf{v}_i^t\|^2, \quad s.t.\ \|\mathbf{v}_i^s\|_0 \le \tau, i = 1, 2, \cdots, n_t, \tag{21}$$

with $\tilde{\mathbf{z}}_i^t = \begin{bmatrix}\mathbf{W}_t^T\mathbf{x}_i^t \\ \sqrt{\lambda}\end{bmatrix}$ and $\tilde{\mathbf{Z}}_s = \begin{bmatrix}\mathbf{W}_s^T\mathbf{X}_s \\ \sqrt{\lambda}\mathbf{h}_s^T\end{bmatrix}$.

The indicator $\mathbf{h}_s \in \mathbb{R}^{n_s\times 1}$ is employed to reflect the availability of source domain samples for reconstructing target domain samples. $\mathbf{h}_s$ is also initialized as an all-1 vector and is updated as follows:

$$\mathbf{h}_s \Leftarrow \mathbf{h}_s - \frac{0.5}{\max(|\mathbf{v}_i^{t*}|)}|\mathbf{v}_i^{t*}|. \tag{22}$$

Step 2: given $\mathbf{V}_s$ and $\mathbf{V}_t$, optimize $\mathbf{W}_s$ and $\mathbf{W}_t$.

Fixing $\mathbf{V}_s$ and $\mathbf{V}_t$, Eq. (8) can be re-formulated to:

$$[\mathbf{W}_s^*, \mathbf{W}_t^*] = \arg\max_{\mathbf{W}_s,\mathbf{W}_t} \frac{\mathrm{Tr}\left(\frac{1}{n_s}\mathbf{W}_s^T\mathbf{X}_s\mathbf{X}_s^T\mathbf{W}_s + \frac{1}{n_t}\mathbf{W}_t^T\mathbf{X}_t\mathbf{X}_t\mathbf{W}_t^T\right)}{\frac{1}{n_s}\|\mathbf{W}_s^T\mathbf{X}_s - \mathbf{W}_t^T\mathbf{X}_t\mathbf{V}_s\|_F^2 + \frac{1}{n_t}\|\mathbf{W}_t^T\mathbf{X}_t - \mathbf{W}_s^T\mathbf{X}_s\mathbf{V}_t\|_F^2} \tag{23}$$

By concatenating $\mathbf{W}_s$ and $\mathbf{W}_t$ as one matrix $\mathbf{W} = [\mathbf{W}_s^T\ \mathbf{W}_t^T]^T$, Eq. (23) can be further re-formulated as the following problem with a norm-1 constraint:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}\left(\mathbf{W}^T\boldsymbol{\Sigma}_b\mathbf{W}\right)}{\mathrm{Tr}\left(\mathbf{W}^T\boldsymbol{\Sigma}_w\mathbf{W}\right)}, s.t.\ \|\mathbf{w}_i\|^2 = 1, \tag{24}$$

where $\mathbf{w}_i$ is the i-th column of $\mathbf{W}$. $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_w$ are defined as:

$$\mathbf{\Sigma}_b = \begin{bmatrix} \frac{1}{n_s}\mathbf{X}_s\mathbf{X}_s^T & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_t}\mathbf{X}_t\mathbf{X}_t^T \end{bmatrix}, \tag{25}$$

$$\mathbf{\Sigma}_w = \begin{bmatrix} \mathbf{X}_s\left(\frac{\mathbf{I}}{n_s}+\frac{\mathbf{V}_t\mathbf{V}_t^T}{n_t}\right)\mathbf{X}_s^T, & -\mathbf{X}_s\left(\frac{\mathbf{V}_s^T}{n_s}+\frac{\mathbf{V}_t}{n_t}\right)\mathbf{X}_t^T \\ -\mathbf{X}_t\left(\frac{\mathbf{V}_t^T}{n_t}+\frac{\mathbf{V}_s}{n_s}\right)\mathbf{X}_s^T, & \mathbf{X}_t\left(\frac{\mathbf{I}}{n_t}+\frac{\mathbf{V}_s\mathbf{V}_s^T}{n_s}\right)\mathbf{X}_t^T \end{bmatrix}. \tag{26}$$

According to the works (Wang et al. 2007; Jia et al. 2009), Eq. (24) is in the trace ratio form, for which the closed form solution does not exist. We therefore relax Eq. (24) into a more tractable ratio trace form:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \mathrm{Tr}\left(\frac{\mathbf{W}^T\mathbf{\Sigma}_b\mathbf{W}}{\mathbf{W}^T\mathbf{\Sigma}_w\mathbf{W}}\right), s.t. \|\mathbf{w}_i\|^2 = 1. \tag{27}$$

Eq. (27) can be easily solved by using the generalized eigenvalue decomposition.

Step 3: Repeat Step 1 and 2 until $\mathbf{W}_s$, $\mathbf{W}_t$, $\mathbf{V}_s$ and $\mathbf{V}_t$ converge or exceeds a maximum number of iterations. For clarity, the overall algorithm is summarized in Algorithm 1.

4 Experiments

In this section, after describing the experimental settings and the partition of each dataset, we validate the proposed method from several different perspectives. Firstly, we investigate the insensitivity of our TSD method to the parameters. Secondly, we validate the necessity of each component in TSD method. Thirdly, we empirically check the convergence of the objective function. Finally, we evaluate the TSD method by comparing with existing methods on three face recognition scenarios: (1) *domain adaptation across view angle*, where the source and target domains correspond to different view angles on MultiPIE dataset (Gross et al. 2007); (2) *domain adaptation across ethnicity*, where the source and target domains contain subjects from two different ethnicities: a Mongolian dataset OFD (XianJiaotong 2006) and a Caucasian dataset XM2VTS (Messer et al. 1999); (3) *domain adaptation across imaging condition*, where the images in the source and target domains are captured under two different imaging conditions, which is simulated by the constrained XM2VTS (Messer et al. 1999) and the unconstrained FRGC (Phillips et al. 2005).

4.1 Basic Experimental Settings

As this work mainly focuses on the unsupervised domain adaptation problem, in all experiments, the source domain training data is labeled, while the target domain training data is unlabeled. Besides, a gallery and a probe set from the target domain are also needed to evaluate the performance of the learnt target model. The measurement for performance is classification accuracy, i.e., rank-1 recognition rate.

In all experiments, the face images are aligned according to the manually labeled eye locations, and then normalized to 80 × 64 pixels on OFD, XM2VTS and FRGC datasets, but to 40 × 32 pixels on MultiPIE dataset for quick evaluation of multiple settings. After this normalization, each image is represented as a column vector by stacking its raw pixels, of which the dimensionality is further reduced through principal component analysis (PCA) (Turk and Pentland 1991) with 98 % energy kept, and then used as the input feature (i.e., to form the $\mathbf{X}_s$ and $\mathbf{X}_t$). Please note that PCA is applied separately for the source and target domains.

The parameters of all involved methods are tuned to report the best results unless otherwise specified. For our TSD method, the dimension of the common subspace, i.e., the column size of $\mathbf{W}_s$ and $\mathbf{W}_t$, is empirically set to 300 in all the experiments. For fairness, all the reported results of our TSD method are the mean accuracy of 10 trials with randomly initialized $\mathbf{W}_s$ and $\mathbf{W}_t$.

4.2 Partition of Datasets for the Evaluation

4.2.1 Datasets for Domain Adaptation Across View Angle

In this work, we define domain adaptation across view angle as that: one view is used as the source domain, e.g. frontal view, and another view, e.g. the profile, as the target domain. For this evaluation, the MultiPIE dataset (Gross et al. 2007) is exploited. It contains more than 750,000 images of 337 subjects under various poses, illuminations and expressions. Specifically, a subset including about 14,450 images from all subjects from four collecting sessions at seven poses (−45°, −30°, −15°, 0°, 15°, 30°, 45°), with three expressions (Neutral, Smile, Disgust) and no flush illumination is selected as the evaluation dataset. This evaluation dataset is further divided into seven subsets according to view angle. For each view angle, the images of 200 subjects with about seven randomly selected images per subject are used for training, and the images of the remaining 137 subjects are used for testing. Among the testing images, 1 and 4 images per subject are randomly selected as the gallery and probe images respectively. Some exemplar images are shown in Fig. 3a.

In summary, for each view angle on MultiPIE, 1,383 images from 200 subjects are used as the training data, 137 images from the rest 137 subjects are used as gallery, and 553 images from 137 subjects are used as probes.

In order to simulate the domain adaptation across view angle, two view subsets, e.g., view −45° subset and view 45° subset, are selected as source and target domain respectively. If one view is used as source domain, the class label of its training data will be given; otherwise, the class label will not

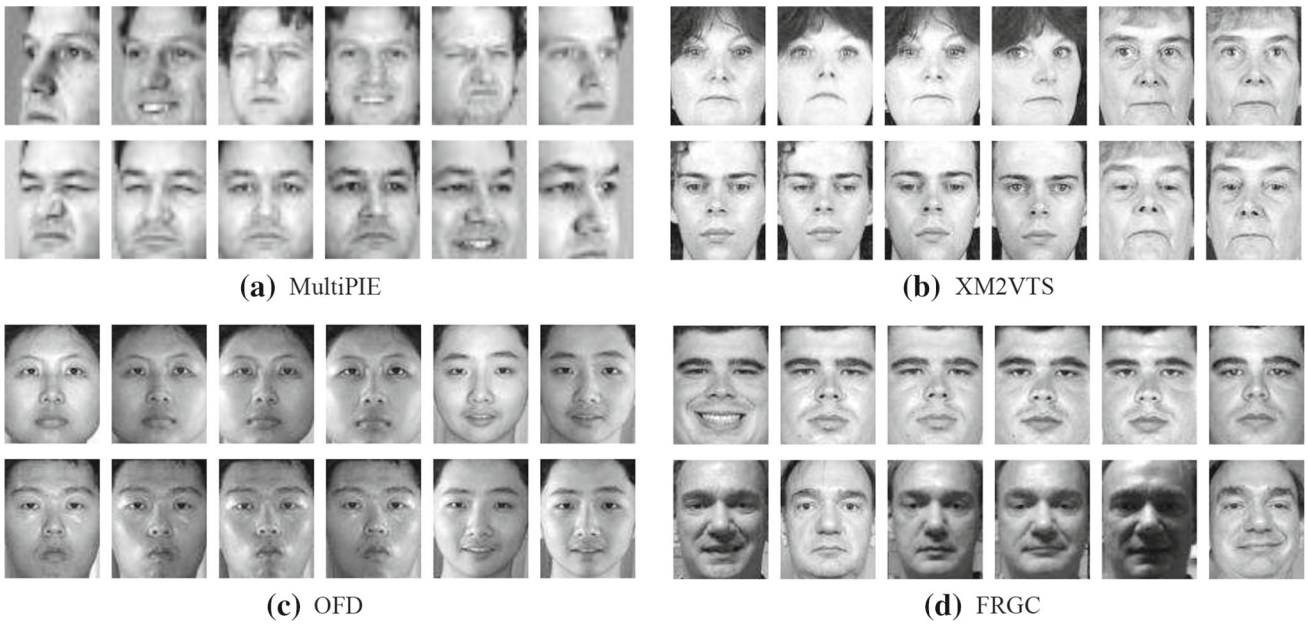


Fig. 3 The exemplar images from **(a)** MultiPIE dataset, **(b)** OFD dataset, **(c)** XM2VTS dataset, and **(d)** FRGC dataset.

Table 1 An exemplar of evaluation settings for MutiPIE dataset.

domain	datasets		#subjects	#images
-45° (source domain)	labeled training data		200	1383 (about 7 images per subject)
45° (target domain)	unlabeled training data		200	1383 (about 7 images per subject)
	test data	gallery	137	137 (1 image per subject)
		probe set	137	553 (about 4 images per subject)

be given for the target training data. An illustration of this evaluation setting is shown in Table 1.

4.2.2 Datasets for Domain Adaptation Across Ethnicity

For domain adaptation across ethnicity, the XM2VTS dataset (Messer et al. 1999) consisting of mainly Caucasian and the Oriented Face Dataset (OFD) (XianJiaotong 2006) consisting of mainly Mongolian are used. The XM2VTS dataset contains 3,440 images of 295 subjects taken over a period of 4 months with different pose and illumination variations. Eight images per subject with slight pose variations are randomly selected for evaluation. Specifically, for each subject, 4 of the 8 images are randomly selected to form the training set. Then, the remaining images form the testing set: for each subject, 1 image is enrolled into gallery and the left 3 images are used as probes. Some exemplar images are shown in Fig. 3b.

To match the scale of XM2VTS, we use only a subset of OFD dataset, containing 800 subjects with 4 images per subject under slight lighting variations. For OFD database, images of the first 400 subjects are used as training data, and images of the rest 400 subjects are used for testing. Specifically, 1 image per subject is randomly selected to form the gallery, and the rest 3 images of each subject are used as the probes. Some exemplar images are shown in Fig. 3c.

In summary, for XM2VTS, 1,180 images from 295 subjects are used as training data, 295 images from 295 subjects are used as gallery, and 885 images from 295 subjects are used as probe. For OFD, 1,600 images from 400 subjects are used as training data, 400 images from 400 subjects (one per subject) are used as gallery, and 1,200 images from 400 subjects are used as probes. For both datasets, if one is used as source domain, the training data of this dataset is used along with class label; if the dataset is used as the target domain, the training set is used without class label.

4.2.3 Datasets for Domain Adaptation Across Imaging Condition

For face recognition, another important factor that can cause the distribution different is the imaging condition, e.g., the images collected in constrained environment usually look different from the images collected in unconstrained environment. To simulate this scenario, we employ the XM2VTS dataset (Messer et al. 1999) collected in constrained environment and FRGC datasets (Phillips et al. 2005) collected in

unconstrained environment for evaluation. The XM2VTS is organized the same as previously mentioned.

The FRGC dataset consists of about 50,000 recordings. It has six experiments, but here, we follow the protocol for Experiment 4 which is designed for unconstrained face verification. Differently, we perform face recognition in this work. For the evaluation, we randomly select 10 images per subject from the standard training set to form a new training set, then randomly select 1 image per subject from the standard target set as the gallery, and select $2 \sim 6$ images per subject from the standard query set as the probe. In total, 2,220 images are included in the training set, 2,520 in probe set, and 466 images in gallery set, respectively. Some exemplar images are shown in Fig. 3d.

Similarly, for both XM2VTS and FRGC, if one is used as source domain, the training set of this dataset is used along with class label; if the dataset is used as target domain, the training set is used without class label.

4.3 Influence of the Parameters in TSD

Our TSD method has two parameters, τ and λ in Eq. (17), which are used to control respectively the sparsity and the disperse selection of samples for reconstruction. To test their influence on the recognition performance, we conduct an experiment on the MultiPIE dataset with $-45°$ as source domain and $45°$ as target domain. The results are shown in Fig. 4. We can observe that the performances roughly increase first and then decrease with the increase of λ, independent of the parameter τ. From the results, how the performances change with the parameter τ is not so clear for different lambda. However, it seems clear that the best choice for τ is 1. The possible principle behind is that the smaller the number of neighbors used for the cross-domain reconstruction, the closer both domains are. Please note that, although the performances of our TSD change with respect to the parameters τ and λ, the fluctuation is trivial as seen in Fig. 4, which demonstrates that TSD is insensitive to τ and λ.

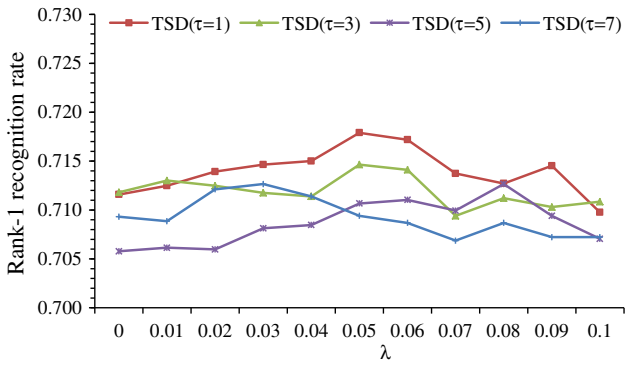


Fig. 4 The performance of TSD w.r.t different λ and τ on MultiPIE dataset with $-45°$ as source domain, and $45°$ as target domain.

4.4 Necessity of each Component in TSD

As emphasized, the key feature of our TSD is targetizing the source domain in the original sample space but learning the reconstruction coefficients in the common subspace. So, it is desirable to validate the essentiality of each component in our TSD method, by investigating the follow strategies:

(1) *Sparse Reconstruction (SR_Ori) only*. In this setting, the coefficients for the targetization is directly achieved in the **original sample space** rather than common subspace.
(2) *Common Subspace (CS) only*. In this setting, the whole process is conducted in the common subspace, i.e., the supervised model is learnt on the source domain, and is directly applied for the target domain in the common subspace (without returning to the original sample space). This strategy discards the particular knowledge of the target domain.
(3) *TSD without the Max Variance term and the Indicator* $\mathbf{h}_t/\mathbf{h}_s$ (TSD w/o MV&I). This setting evaluates the necessity of max variance term and the indicators. Please kindly note that the indicators $\mathbf{h}_s$ and $\mathbf{h}_t$ are designed to explicitly enforce different source/target samples to be sparsely constructed by different target/source samples, so as to guarantee the diversity of targetized source domain.
(4) *TSD without the max variance term, but with the indicator* $\mathbf{h}_t/\mathbf{h}_s$ (TSD w/o MV). This setting evaluates the performance when without the max variance term.

The above settings are evaluated on three tasks, i.e., domain adaptation across view angle, ethnicity and imaging condition respectively, as shown in Tables 2 and 3.

As seen, the SR_Ori strategy performs much worse than TSD. We conjecture that it is because the two domains depart too far away from each other, which makes it too difficult (if not impossible) to reconstruct the source domain sample with the target domain samples. As a result, the coefficients

Table 2 Evaluation of each component in TSD in terms of domain adaptation across view angle on MultiPIE dataset

	View angles					
Source	$-45°$	$45°$	$0°$	$0°$	$-30°$	$30°$
Target	$45°$	$-45°$	$-45°$	$45°$	$30°$	$-30°$
SR_Ori	0.436	0.488	0.591	0.564	0.515	0.523
CS	0.629	0.600	0.620	0.707	0.653	0.671
TSD w/o MV&I	0.707	0.702	0.694	0.713	0.768	0.752
TSD w/o MV	0.707	0.712	0.716	0.714	0.782	0.771
TSD	0.718	0.709	0.725	0.731	0.791	0.773

Table 3 Evaluation of each component in TSD in terms of domain adaptation across ethnicity and imaging condition

	Ethnicities		Imaging conditions	
Source	OFD	XM2VTS	FRGC	XM2VTS
Target	XM2VTS	OFD	XM2VTS	FRGC
SR_Ori	0.815	0.932	0.779	0.075
CS	0.811	0.966	0.769	0.168
TSD w/o MV&I	0.813	0.980	0.825	0.228
TSD w/o MV	0.816	0.981	0.827	0.229
TSD	0.858	0.974	0.855	0.232

for the sparse reconstruction obtained in the original sample space are not reliable as those obtained in the common space, which implies the importance of bridging the two domains via a common subspace. This conclusion is further proved by the better performance of CS than SR_Ori.

It can be also seen that the TSD outperforms the CS strategy significantly. This gain can be safely attributed to the use of the particular knowledge of the target domain, which are not preserved in the common subspace.

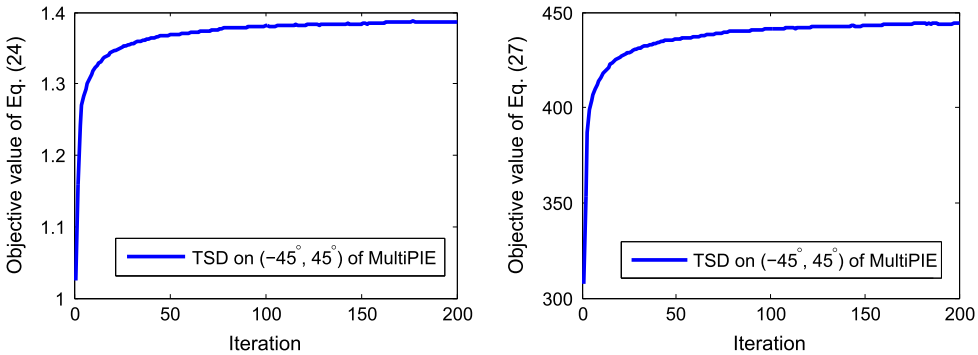
Another observation from Tables 2 and 3 is that TSD degenerates if the max-variance term and/or the indicator are removed. The reason behind is that, without these two terms many source domain samples may be reconstructed by the same target domain samples, which might lead to smaller reconstruction error, but cannot promise the closeness between domains with enough structural information.

The above observations and investigations demonstrate that the common subspace is fundamental for knowledge transfer, while making use of more information about target domain can further benefit the adaptation. This argument forms the principle of our basic idea that converting the source samples to target domain in the original sample space, with the converting parameters however learnt in the common space, where the two domains preserve their own structures but are pulled as close as possible.

4.5 Convergency of the Optimization for TSD

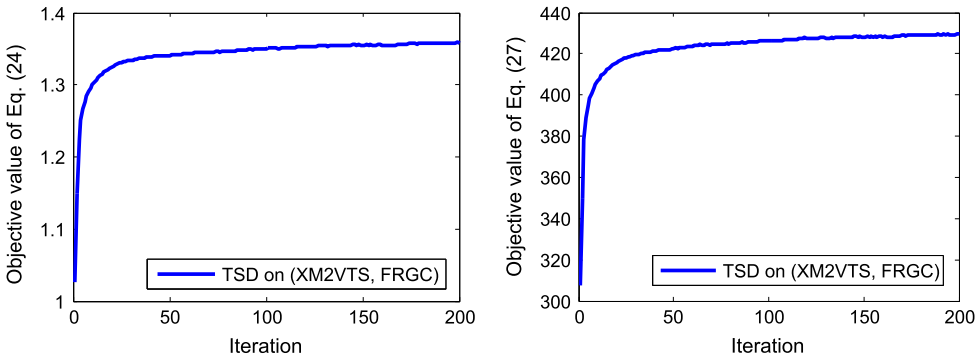
Since the objective in Eq. (8) is optimized in an alternation way, we would like to explore the convergency of the algorithm. While it is hard to prove its convergence theoretically, we can anyway show some empirical evaluation. Here, we show the objective values of Eqs. (24) and (27) on MultiPIE with $-45°$ as source domain and $45°$ as target domain. The results are shown in Figs. 5 and 6. As seen, the algorithm steadily converges after a number of iterations.

Besides, TSD is generally not time consuming. For example, the training time on MultiPIE only takes about 6 minutes on a desktop PC with 3.10 GHz CPU and 8 GB memory.



(a) objective value of Eq. (24) **(b)** objective value of Eq. (27)

Fig. 5 The objective values of Eqs. (24) and (27) in scenario of domain adaptation across pose, with -45° and 45° as the source and target domains respectively



(a) objective value of Eq. (24) **(b)** objective value of Eq. (27)

Fig. 6 The objective values of Eqs. (24) and (27) in scenario of domain adaptation across imaging condition, with XM2VTS and FRGC as source and target domains respectively

4.6 Comparison with the Existing Works

To further validate our method, we compare it with several existing methods under three scenarios, i.e., domain adaptation across view angle, domain adaptation across ethnicity and domain adaptation across imaging condition. Several competitive approaches are briefly described as bellow.

PCA (Turk and Pentland 1991): PCA is a typical unsupervised method, taken as the baseline. PCA is directly conducted on target domain, and the dimension is determined by keeping 98 % energy.

FLD (Belhumeur et al. 1997): Fisher's Linear Discriminant analysis is a widely-used supervised method for feature extraction. The FLD model learnt from the supervised source domain without any adaptation is also tested as a baseline by being directly applied to the target domain.

TDR (Wang et al. 2008): Transferred Dimension Reduction method is an effective method to directly combine the information from both domains into one model, i.e., assemble the scatter of source domain and the estimated scatter of target domain into one FLD-like model.

ITL (Shi and Sha 2012): Information Theoretical Learning method aims at obtaining a common subspace through information theory. ITL endeavors to reduce the disparity of two domains measured by the mutual information of samples and domain labels, and keeps the discrimination on the target domain measured by the entropy of predicted class label at the mean time.

SGF (Gopalan et al. 2011): In Sampling Geodesic Flow approach, a series of intermediate common representations are created by projecting the data onto the sampled intermediate subspaces. These subspaces are sampled along the path from the source domain to target domain on the Grassmann manifold. As suggested in the work (Gopalan et al. 2011), partial least squares (PLS) method is employed for final classification.

Among these methods, PCA only exploits the unlabeled training data of target domain and FLD only exploits the labeled training data of source domain, while the rest domain adaptation methods, i.e., TDR, ITL, SGF and our TSD, make use of both labeled training data of source domain and unlabeled training data of target domain. For clarity, in all tables, we use "Source Model" representing the models learnt by using only the training data of source domain, and "Target Model" representing the models learnt by using only the training data of target domain.

4.6.1 Domain Adaptation Across View Angle

Domain adaptation across view angle attempts to adapt the knowledge from one view to another. This problem is practical and interesting, because we can generally have sufficient labeled training images in frontal view, but lack of enough labeled data in side view. The MultiPIE datasets are employed to simulate this scenario: with one view as the source domain and another view as the target domain.

The evaluation results are shown in Table 4. As seen, FLD performs only slightly better than PCA since the two domains are different from each other and thus it is difficult to achieve a good performance by directly applying the source domain model to the target domain without any adaptation. Among the existing domain adaptation methods, SGF and ITL achieve much better performance than PCA, as they exploit the source domain knowledge via common subspace or domain-invariant feature representation. TDR outperforms both of them as it combines all the information from both domains in the original sample space, not only the commonality. However, TDR is worse than our TSD, especially when the difference of the source and target domains is relatively large, e.g. from −45° to 45°. The improvement of our TSD benefits from *on one hand* targetizing the source domain in the original samples space, where particular knowledge of the target domain is preserved, and *on the other hand* achieving the reconstruction coefficient for the targetization in a common subspace, which can ensure a more stable cross-domain reconstruction relationship.

Table 4 Rank-1 face recognition rates of domain adaptation across view angle on MultiPIE dataset

Source	Target	Target model	Source model	Domain adaptation			
		PCA	FLD	TDR	SGF	ITL	TSD
−45°	45°	0.533	0.644	0.679	0.631	0.664	0.718
45°	−45°	0.535	0.637	0.669	0.608	0.640	0.709
0°	−45°	0.535	0.586	0.685	0.609	0.620	0.725
0°	45°	0.535	0.467	0.697	0.631	0.671	0.731
−45°	30°	0.534	0.691	0.748	0.664	0.658	0.785
45°	−30°	0.543	0.671	0.741	0.667	0.696	0.764
−30°	30°	0.588	0.692	0.789	0.635	0.666	0.791
30°	−30°	0.584	0.696	0.752	0.676	0.698	0.773

Table 5 Rank-1 face recognition rates of domain adaptation across ethnicity.

Source	Target	Target model	Source model	Domain adaptation			
		PCA	FLD	TDR	SGF	ITL	TSD
OFD	XM2VTS	0.745	0.768	0.845	0.783	0.715	0.858
XM2VTS	OFD	0.268	0.593	0.971	0.653	0.110	0.974

Table 6 Rank-1 face recognition rates of domain adaptation across imaging condition

Source	Target	Target model	Source model	Domain adaptation			
		PCA	FLD	TDR	SGF	ITL	TSD
FRGC	XM2VTS	0.745	0.767	0.831	0.791	0.731	0.855
XM2VTS	FRGC	0.012	0.083	0.183	0.073	0.041	0.232

4.6.2 Domain Adaptation Across Ethnicity

For domain adaptation across ethnicity, the Mongolian OFD dataset (XianJiaotong 2006) and the Caucasian XM2VTS dataset (Messer et al. 1999) are exploited, with one as source domain and the other as target domain. The evaluation is shown in Table 5, from which similar conclusion can be drawn. The unsupervised PCA performs worst, followed by FLD and SGF. TDR and our TSD perform the best with comparable performance. This is because the source domain is not significantly different from target domain, considering that the images of both domains are frontal without large variations in expression and occlusion.[2]

4.6.3 Domain Adaptation Across Imaging Condition

For domain adaptation across imaging condition, the XM2 VTS dataset (Messer et al. 1999) collected in constrained condition and the FRGC dataset (Phillips et al. 2005) collected in unconstrained condition are employed, with one as source domain and the other as target domain. The evaluation results are shown in Table 6. As seen, all methods performs better on XM2VTS than on FRGC, since XM2VTS is collected under constrained environment, which means an easier task than that on FRGC. When FRGC is used as the target domain, the recognition rates of PCA, FLD, SGF and ITL are extremely low, even lower than 10 %. This might be attributed to the quite large variations from the uncontrolled condition which forms a challenging task. Compared with the above methods, TDR can achieve much better performance up to 18.3 %, while our method can impressively outperform TDR with the accuracy up to 23.2 %.

4.7 Discussion

As clearly seen from the above evaluation, our proposed TSD outperforms other domain adaptation approaches. Specially, from the comparison results in Tables 2 and 3, the superiority of our TSD is attributed to the exploiting not only the common knowledge of source and target domains, but also the particular knowledge of the target domain. The principle behind is twofold: firstly, it is more reasonable to determine the reconstruction coefficients in a common subspace, in which the particular parts of both domains are removed, because they are useless for determining the cross-domain relationship; secondly, the particular characteristics of target domain are well preserved by applying the reconstruction coefficients in the original sample space, i.e., targetization of the source domain. Such a strategy actually replaces the source-domain-particular components with target-domain-particular components, while keeping the common components of source and target domains.

A big difference of our TSD from most existing works, e.g., TDR and ITL, is that, the whole targetization is totally unsupervised, even without using any supervision information of the source domain. However, the class label naturally remains unchanged for the targetized source domain, which can be further used by any supervised method, e.g., Fisher's linear discriminant analysis.

Another intrinsic advantage of our method is that, it can still work even when the feature dimensions of the source and target domains are different, while many previous methods, e.g., the TDR, cannot survive in this scenario. This advantage of our method comes from that two different projections are exploited for source and target domains.

In case of domain adaption between the flipped view angles, e.g., $(-45°, 45°)$ and $(-30°, 30°)$, the optimal $\mathbf{W}_s$

[2] In this experiment the performance of ITL is even worse than PCA, however this does not mean the inferiority of ITL since the data distribution in this setting does not agree with the assumption of ITL: ITL assumes that the data in both source and target domains are tightly clustered, and clusters from both domains are aligned if they correspond to the same class. In this setting here, the source and target domains only have several samples in each class which are difficult to form a tight cluster, and even worse the samples from the source and target domains are from totally different classes.

and $\mathbf{W}_t$ should be the flip of each other and the best recognition model for target domain can be obtained from the flipped source images. In current work, we are unable to achieve this optimal solution. This indeed forms a very good future direction to deepen the proposed framework. However, we mildly think that this does not degenerate our method too much, as our TSD is a general method without assuming this kind of strong correspondence between view angles.

5 Conclusions and future works

In this work, we propose an unsupervised domain adaptation method via targetizing the source domain images bridged by the common subspace learning. Our method directly converts the source domain data to the target domain in the image space rather than in the common subspace, while the sparse reconstruction coefficients are learnt in the common subspace. For each source domain image, such a *Targetizing* strategy actually preserves the commonality between domains, while substitutes some target-particular components for its source-particular components.

The evaluations on three face recognition tasks demonstrate the superiority of our TSD to the existing methods. Besides, the investigations also imply that both the commonality between domains and the particularity of target domain are essential for domain adaptation: the commonality can bridge the gap between domains, while the particularity of the target domain can preserve its specific characteristics, which are beneficial for the task in target domain.

As discussed, only unsupervised information is considered in our targetization. But, the supervised information of the source domain does provide more discriminative information, which can be further exploited to guide the targetization of the source domain. Therefore, one of our future works will study on how to exploit the supervision information in the procedure of adaptation, especially during the common subspace learning.

Acknowledgments This work is partially supported by Natural Science Foundation of China under contracts nos. 61025010, 61173065, and 61222211. The authors would like to thank the guest editors and the reviewers for their valuable comments and suggestions. The authors also would like to thank the Edwin Zinan Zeng for his advices about the writing.

References

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *19*(7), 711–720.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems NIPS*, *19*, 137–144.

Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. *The Journal of Machine Learning Research (JMLR)*, *10*, 2137–2155.

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 120–128).

Bruzzone, L., & Marconcini, M. (2010). Domain adaptation problems: a dasvm classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *32*(5), 770–787.

Chen, Y., Wang, G., & Dong, S. (2003). Learning with progressive transductive support vector machine. *Pattern Recognition Letters (PRL)*, *24*(12), 1845–1855.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, *59*(6), 797–829.

Duan, L., Tsang, I. W., Xu, D., & Maybank, S. J. (2009). Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1375–1381).

Duan, L., Xu, D., Tsang, I., & Luo, J. (2012). Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *34*(9), 1667–1680.

Dudık, M., Schapire, R. E., & Phillips, S. J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems (NIPS)*, *17*, 323–330.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *39*(4), 407–499.

Gao, X., Wang, X., Li, X., & Tao, D. (2011). Transfer latent variable model based on divergence analysis. *Pattern Recognition (PR)*, *44*(10–11), 2358–2366.

Geng, B., Tao, D., & Xu, C. (2011). Daml: Domain adaptation metric learning. *IEEE Transactions on Image Processing (T-IP)*, *20*(10), 2980–2989.

Gong, B., Shi, Y., Sha, F., & Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *0*, 2066–2073.

Gopalan, R., Li, R., & Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 999–1006).

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning* (pp. 131–160). Cambridge: MIT Press.

Gross, R., Matthews, I., Cohn, J., kanada, T., & Baker, S. (2007). The cmu multi-pose, illumination, and expression (multi-pie) face database. Tech. rep., Carnegie Mellon University Robotics Institute. TR-07-08.

Hal, DI. (2009). Bayesian multitask learning with latent hierarchies. In *Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 135–142).

He, X., & Niyogi, P. (2004). Locality preserving projections. *Advances in Neural Information Processing Systems NIPS*, *16*, 153–160.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2006). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*.

Huang, K., & Aviyente, S. (2007). Sparse representation for signal classification. *Advances in Neural Information Processing Systems NIPS*, *19*, 609–616.

Jhuo, IH., Liu, D., Lee, D. T., & Chang, S. F. (2012). Robust visual domain adaptation with low-rank reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2168–2175).

Jia, Y., Nie, F., & Zhang, C. (2009). Trace ratio problem revisited. *IEEE Transactions on Neural Networks (T-NN)*, *20*(4), 729–735.

Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing (T-IP)*, *11*(4), 467–476.

Mehrotra, R., Agrawal, R., Haider, S. A. (2012). Dictionary based sparse representation for domain adaptation. In *ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 2395–2398).

Messer, K., Matas, M., Kittler, J., Lttin, J., & Maitre, G. (1999). Xm2vtsdb: The extended m2vts database. In *International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)* (pp. 72–77).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (T-KDE)*, *22*(10), 1345–1359.

Pan, S. J., Kwok, J. T., & Yang, Q. (2008) Transfer learning via dimensionality reduction. In *AAAI Conference on Artificial Intelligence (AAAI)* (pp. 677–682).

Pan, S. J., Tsang, I. W., Kwok, J. T., Yang, Q. (2009). Domain adaptation via transfer component analysis. In *International Joint Conferences on Artificial Intelligence (IJCAI)* (pp. 1187–1192).

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks (T-NN)*, *22*(2), 199–210.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., et al. (2005). Overview of the face recognition grand challenge. *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, *1*, 947–954.

Qiu, Q., Patel, V. M., Turaga, P., & Chellappa, R. (2012). Domain adaptive dictionary learning. In *European Conference on Computer Vision (ECCV)* (pp. 631–645).

Raina, R., Battle, A., Lee, H., Packer, B., Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *International Conference on Machine Learning (ICML)* (pp 759–766).

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Shao, M., Castillo, C., Gu, Z., Fu, Y. (2012). Low-rank transfer subspace learning. In *IEEE International Conference on Data Mining (ICDM)* (pp. 1104–1109).

Shi, Y., & Sha, F. (2012). Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*.

Shimodaira, Hidetoshi. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244.

Si, S., Tao, D., & Geng, B. (2010). Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering T-KDE*, *22*(7), 929–942.

Si, S., Liu, W., Tao, D., & Chan, K. P. (2011). Distribution calibration in riemannian symmetric space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *41*(4), 921–930.

Su, Y., Shan, S., Chen, X., & Gao, W. (2009). Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Transactions on Image Processing T-IP*, *18*(8), 1885–1896.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., & Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. *In: Advances in Neural Information Processing Systems NIPS*, *20*, 1433–1440.

Sugiyamai, M., Krauledat, M., & Müller, K. R. (2007). Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research (JMLR)*, *8*, 985–1005.

Turk, M. A., & Pentland, A. P. (1991). Face recognition using eigenfaces. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *591*, 586–591.

Uribe, D. (2010). Domain adaptation in sentiment classification. In *International Conference on Machine Learning and Applications (ICMLA)* (pp. 857–860).

Wang, H., Yan, S., Xu, D., Tang, X., & Huang, T. (2007). Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Wang, Z., Song, Y., Zhang, C. (2008). Transferred dimensionality reduction. In *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)* (pp. 550–565).

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *31*(2), 210–227.

XianJiaotong, U. (2006). http://www.aiar.xjtu.edu.cn/dfrlsjk5.htm.

Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research (JMLR)*, *8*, 35–63.

Zadrozny, & Bianca (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of International Conference on Machine Learning (ICML)* (p. 114).