Discriminative Kernel Transfer Learning via $l_{2,1}$-Norm Minimization

Lei Zhang
College of Communication Engineering,
Chongqing University
Chongqing 400044, China
leizhang@cqu.edu.cn

Sunil Kr. Jha
University of Information Science and Technology "St. Paul the Apostle",
6000 Ohrid, Macedonia
sdrsunil76@gmail.com

Tao Liu
College of Communication Engineering,
Chongqing University
Chongqing 400044, China
cquliutao@cqu.edu.cn

***Abstract*—In this paper, we propose a $l_{2,1}$-norm based discriminative robust transfer learning (DKTL) method for domain adaptation tasks. The key idea is to simultaneously learn discriminative subspaces by using the proposed *domain-class-consistency* (*DCC*) metric, and the representation based robust transfer model between source domain and target domain via l_{21}-norm minimization. The *DCC* metric includes two parts: *domain-consistency* used to measure the between-domain distribution discrepancy and *class-consistency* used to measure the within-domain class separability. The objective of transfer learning is to maximize the proposed metric, while for easily formulating this metric in model, we propose to minimize the domain-class-inconsistency, such that both domain distribution mismatch and class inseparability are well addressed. Two advantages of the proposed method are that on one hand the robust sparse coding selects a few valuable source data with noises (outliers) removed during knowledge transfer, and the proposed DCC metric can help to pursue discriminative subspaces of different domains for classification based transfer learning tasks. Extensive experiments demonstrate the superiority of the proposed method over other state-of-the-art domain adaptation methods.**

***Keywords*—Transfer learning; sparse coding; discriminative subspace; domain adaptation**

I. INTRODUCTION

A basic assumption of machine learning is that the training data and testing data should hold similar probability distribution, i.e. independent identical distribution (*i.i.d*) which shares the same subspace. However, in many real applications, machine learning faces with the dilemma of insufficient labeled data. For learning a robust classification model, researchers have to "borrow" more data from other domains for training. There is one problem of the borrowed data is that the domain mismatch between *source* domain and *target* domain violates the basic assumption of machine learning. Such domain mismatch often results from a variety of visual cues or abrupt feature changes, such as camera viewpoint, resolution (e.g. image sensor from webcam to DSLR), illumination conditions, color correction, poses (e.g. faces with different angles), and background, etc. Physically, such distribution mismatch or domain shift is common place in vision problems. With this violation, it has been shown that significant performance degradation is suffered in classification [1]. For a typical object recognition scenario in computer vision, users would like to recognize a few objects captured by a mobile phone via a learned model, while training by using the labeled training data from an existing object dataset, such as Caltech 256 [2] or web images that may be sampled under different visual cues from that of the users. Some example images of objects from different domains are shown in Figure 1, which explicitly shows the domain shifts.

Fig. 1. Examples of object images from 4 sources: Amazon (1st row), DSLR (2nd row), Webcam (3rd row) and Caltech (4th row).

For handling such domain mismatch issues, transfer learning and domain adaptation based methods have been emerged [3-10]. Generally, existing methods are classifier and feature oriented domain adaptation techniques. The classifier based methods advocate learning a transfer classifier on the source data by leveraging a few labeled data from the target domain. The borrowed target data is used for regularization, which can help improving the decision boundary. In this way, the learned decision function (e.g. SVM) is imposed with transfer capability and can be used for classification of both domains. It is straightforward and easy to understand. However, for determining the decision boundary, a number of labeled data are necessary, which may increase the cost of data labeling. For realizing unsupervised transfer learning, feature based

representation and transformation methods have been proposed. These methods aim at aligning the domain shift by adapting features from the source domain to target domain without training classifiers. Though these methods have been proven to be effective for domain adaptation, two issues still exist. First, for representation based adaptation, the noise and outliers from source data may also be transferred to target data during naïve transformation due to overfitting, which leads to significantly distorted or corrupted data structure. Second, the learned subspace is suboptimal, which limits the transfer ability due to the fact that the subspace and the representation (e.g. global low-rank, local sparse coding etc.) are learned independently. Therefore, subspace learning that help most to representation transfer should be conducted simultaneously.

As described in Fig.2, in this paper, we propose a novel model which targets at learning discriminative subspaces $\mathbf{P}_S$ and $\mathbf{P}_T$ of both domains by using a newly proposed *domain-class-consistency* metric, and a robust transfer representation by using $l_{2,1}$-norm minimization. The model not only reduce the domain distribution mismatch, but also maximize the separability of different classes (i.e. c_1, c_2, c_3.) within the same domain. Also, in the model, we formulate to maximize the inter-class total distance within the same domain, such that the inter-class difference within a domain can cover the between-domain discrepancy. In this way, the distribution mismatch has weaken and the inter-class difference can be enhanced. For example, in face recognition, the difference between two images of the same person captured under different illumination condition may be larger than that of two persons captured under the same condition. Additionally, by imposing $l_{2,1}$-norm constraint on the transfer representation coefficient $\mathbf{Z}$ between source and target data points, the outliers in the source domain can be well removed without incorrectly transferring into the target domain, and only a few valuable data points are utilized. With the above description, we call our method discriminative robust transfer learning (DKTL).

The rest of this paper is organized as follows. Section II summarizes the related work in transfer learning and domain adaptation. The proposed model and optimization algorithms are presented in Section III. The experiments on several datasets for transfer learning tasks are conducted in Section IV. Finally, Section V concludes this paper.

II. Related Work

A number of transfer learning or domain adaptation methods have been proposed. Yang et al. [8] proposed an adaptive support vector machine (ASVM), which aims at learning the perturbation term for adapting the source classifier to the target classifier. Collobert et al. [11] proposed a transductive SVM (T-SVM), which utilized the labeled and unlabeled samples simultaneously. Duan et al. [12] proposed a domain adaptation machine (DAM) method which integrates SVM for classifier adaptation. With the SVM based classifier adaptation idea, they also proposed an adaptive multiple kernel learning method

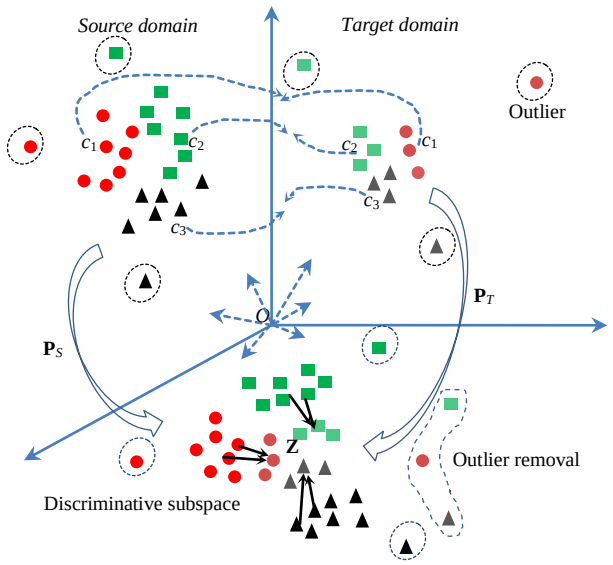


Fig. 2. Schematic diagram of the proposed DKTL method

(AMKL) [13] and a domain transfer MKL (DTMKL or DTSVM) [7] methods, by integrating multiple kernels for improving the robustness and classification accuracy. Zhang et al. [10, 14] also proposed a domain adaptation ELM method for classifier adaptation. They also proposed a robust extreme domain adaptation (EDA) [15] method by using Laplacian graph regularization for local structure preservation and achieve state-of-the-art results. Shekhar et al. [16] proposed a domain adaptive dictionary learning method (SDDL) for representation classifier adaptation. Zhu and Shao [17] also proposed a cross-domain dictionary learning method (WSCDDL) for weakly-supervised transfer learning based on representation classifier adaptation.

In feature adaptation category, Gopalan et al. [4] proposed a SGF method for unsupervised domain adaptation via low dimensional subspace transfer. The idea behind SGF is that it samples a group of subspaces along the geodesic between source and target data, and project the source data into the subspaces for discriminative classifier learning. Gong et al. [18] proposed an unsupervised domain adaptation method (GFK) for visual domain adaptation, in which geodesic flow kernel is used to model the domain shift by integrating an infinite number of subspaces where the changes in geometric and statistical properties are characterized. Fernando et al. [19] proposed principal component subspace alignment (SA) for subspace transfer. More recently, low rank representation (LRR) based domain adaptation is proposed, with two representative work can be referred as [20, 21], in which a common point is that LRR is used for aligning the domain shifts and the effectiveness is demonstrated. As referenced in Liu et al.[22, 23], LRR can get the block diagonal solution and performs perfectly for subspace segmentation when the subspaces are independent and the data sampling is sufficient. Instead, when handling disjoint subspace problems and insufficient data, LRR will not work well. Therefore, LRR based domain adaptation capability will be limited under such

strong, independent subspace assumption. As indicated by recently proposed sparse subspace clustering (SSC) [24, 25] for clustering data points that lie in a union of multiple low-dimensional subspaces or near the intersections of subspaces, by minimizing the reconstruction error $\|\mathbf{X}-\mathbf{XZ}\|_{\mathrm{F}}$ with sparsity constraint of $\mathbf{Z}$ imposed, we are inspired to have an instinctive idea that expect $\|\mathbf{X}_T-\mathbf{X}_S\mathbf{Z}\|_{\mathrm{F}}$ to be minimized for adapting source data to target data lying in different subspaces. Additionally, the minimization problem only guarantees the domain consistency, but the separability in the subspaces may be broken. Therefore, discriminative subspaces for different domains are also simultaneously learned for botch domain and class consistency (i.e. DCC).

III. Proposed Discriminative Kernel Transfer Learning

A. Notations

In this paper, the source and target domain are defined by subscript "*S*" and "*T*". The training set of source and target domain is defined as $\mathbf{X}_S\in\Re^{D\times N_S}$ and $\mathbf{X}_T\in\Re^{D\times N_T}$, where D denotes dimension, N_S and N_T are the number of samples. Let $\mathbf{P}\in\Re^{D\times d}$ represents the basis transformation that maps the original space of the source and target data into some subspace of dimension d, respectively. The reconstruction coefficient matrix is denoted as $\mathbf{Z}$. $\mathbf{I}$ denotes the identity matrix. $\|\cdot\|_p$, $\|\cdot\|_{q,p}$ and $\|\cdot\|_{\mathrm{F}}$ denote l_p-norm, $l_{q,p}$-norm and Frobenius norm. The superscript $^{\mathrm{T}}$ denotes the transpose, $Tr(\cdot)$ denotes the trace operator of a matrix.

B. Problem Formulation

As illustrated in Figure 2, we tend to learn a representation matrix $\mathbf{Z}$ for representing target data $\mathbf{X}_T$ by using the source data $\mathbf{X}_S$ in their discriminative subspace projected by the basis transformation $\mathbf{P}$, respectively. Therefore, the general framework of the proposed DKTL can be formulated as

$$\begin{aligned}&\min_{\mathbf{P},\mathbf{Z}} E(\mathbf{X}_S,\mathbf{X}_T,\mathbf{P},\mathbf{Z})+\lambda\cdot\Omega(\mathbf{P})+\tau\cdot R(\mathbf{Z})\\ &s.t.\ \mathbf{P}^{\mathrm{T}}\mathbf{P}=\mathbf{I},\ \lambda,\tau>0\end{aligned}\tag{1}$$

where $E(\cdot)$ represents the domain-inconsistency term (i.e. representation error), $\Omega(\cdot)$ denotes the class-inconsistency term (i.e. discriminative regularizer) between domains, $R(\cdot)$ represents the robust regularization term on the representation coefficients with outlier removal property, λ and τ represent the positive regularization parameters.

From the optimization problem (1), it is obvious that by jointly minimizing the domain-inconsistency and class-inconsistency, the domain-class-consistency can strengthen such that the proposed DKTL not only realize the domain transfer, but also enhance the class consistency simultaneously. Therefore, the proposed model is more robust for classification oriented transfer learning tasks.

Suppose $\mathbf{P}$ be represented by a linear combination of the transformed training samples $\varphi(\mathbf{X})=[\varphi(\mathbf{X}_S),\varphi(\mathbf{X}_T)]$, which can be written as

$$\mathbf{P}=\varphi(\mathbf{X})\mathbf{\Phi}\tag{2}$$

where $\mathbf{\Phi}\in\Re^{N\times d}$ denotes the combination coefficients, $\varphi(\cdot)$ is some unknown mapping function, and $N=N_S+N_T$,

Specifically, with eq.(2) and the mapping function $\varphi(\cdot)$, the first term in eq.(1) can be formulated as

$$\begin{aligned}E(\mathbf{X}_S,\mathbf{X}_T,\mathbf{P},\mathbf{Z})&=\left\|\mathbf{P}^{\mathrm{T}}\varphi(\mathbf{X}_T)-\mathbf{P}^{\mathrm{T}}\varphi(\mathbf{X}_S)\mathbf{Z}\right\|_{\mathrm{F}}^2\\&=\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_T)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_S)\mathbf{Z}\right\|_{\mathrm{F}}^2\end{aligned}\tag{3}$$

where $\mathbf{Z}\in\Re^{N_S\times N_T}$ denotes the representation matrix between domains. Obviously, smaller E corresponds to stronger domain consistency.

The second term in eq.(1) is formulated as

$$\begin{aligned}\Omega(\mathbf{P})&=\sum_{c=1}^{C}\left\|\mathbf{P}^{\mathrm{T}}\varphi(\boldsymbol{\mu}_S^c)-\mathbf{P}^{\mathrm{T}}\varphi(\boldsymbol{\mu}_T^c)\right\|_2^2-\sum_{t\in\{S,T\}}\sum_{c,k=1,c\neq k}^{C}\left\|\mathbf{P}^{\mathrm{T}}\varphi(\boldsymbol{\mu}_t^c)-\mathbf{P}^{\mathrm{T}}\varphi(\boldsymbol{\mu}_t^k)\right\|_2^2\\&=\sum_{c=1}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\boldsymbol{\mu}_S^c)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\boldsymbol{\mu}_T^c)\right\|_2^2-\\&\qquad\sum_{t\in\{S,T\}}\sum_{c,k=1,c\neq k}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\boldsymbol{\mu}_t^c)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\boldsymbol{\mu}_t^k)\right\|_2^2\end{aligned}\tag{4}$$

where $\varphi(\boldsymbol{\mu}_S^c)=\frac{1}{N_S^c}\sum_{i=1}^{N_S^c}\varphi(\mathbf{X}_{S,i}^c)$ and $\varphi(\boldsymbol{\mu}_T^c)=\frac{1}{N_T^c}\sum_{i=1}^{N_T^c}\varphi(\mathbf{X}_{T,i}^c)$ represent the centroid of class c of source and target training data, respectively. The first term in eq.(4) denotes the intra-class inconsistency between-domains and the second term in eq.(4) denotes the inter-class consistency within-domains. By minimizing the difference between the intra-class inconsistency and the inter-class consistency, the generalized class-consistency can better improve the classification oriented domain transfer performance.

The third term in eq.(1) can be formulated as

$$R(\mathbf{Z})=\|\mathbf{Z}\|_{q,p}\tag{5}$$

where $\|\cdot\|_{q,p}$ represents $l_{q,p}$-norm. Given a matrix $\mathbf{Q}\in\Re^{m\times n}$, then there is

$$\|\mathbf{Q}\|_{q,p}=\left(\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|Q_{i,j}|^q\right)^{p/q}\right)^{1/p}\tag{6}$$

As can be seen from eq.(6), a common Frobenius norm is achieved when $p=q=2$. Intrinsically, different approaches may be induced by selecting different p and q values. Generally, for sparsity pursuit, $q\geq 2$ and $0\leq p\leq 2$ may be constrained. If $p=0$, the problem is not convex and therefore $p=1$ is used for sparse approximation in this paper. Since q is used to measure the row vector norm, and $q=2$ is set because larger q does not improve the results [26]. Therefore, the eq.(5) can be formulated as $R(\mathbf{Z})=\|\mathbf{Z}\|_{2,1}$ for better sparsity and robustness

for outliers. The property of $l_{2,1}$-norm guarantees that the outliers in source data cannot be selected in representation transfer. In this way, the potential outliers in source domain cannot be transferred to target domain via $l_{2,1}$-norm minimization used in the model.

Finally, by substituting eqs.(3), (4) and (5) into eq.(1), the proposed DKTL model can be formulated as

$$\begin{aligned}&\min_{\mathbf{\Phi},\mathbf{Z}}\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_T)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_S)\mathbf{Z}\right\|_{\mathrm{F}}^2+\lambda\cdot\\&\left(\sum_{c=1}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_S^c\right)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_T^c\right)\right\|_2^2-\right.\\&\left.\sum_{t\in\{S,T\}}\sum_{c,k=1,c\neq k}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_t^c\right)-\mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_t^k\right)\right\|_2^2\right)+\tau\cdot\|\mathbf{Z}\|_{2,1}\\&s.t.\ \mathbf{\Phi}^{\mathrm{T}}\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X})\mathbf{\Phi}=\mathbf{I},\ \lambda,\tau>0\end{aligned}\tag{7}$$

According to the Mercer kernel theorem, let $\mathbf{K}=\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X})=\kappa(\mathbf{X},\mathbf{X})$, $\mathbf{K}_T=\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_T)=\kappa(\mathbf{X},\mathbf{X}_T)$, $\mathbf{K}_S=\varphi(\mathbf{X})^{\mathrm{T}}\varphi(\mathbf{X}_S)=\kappa(\mathbf{X},\mathbf{X}_S)$, $\mathbf{K}_{\mu,S}^c=\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_S^c\right)=\kappa\left(\mathbf{X},\boldsymbol{\mu}_S^c\right)$, $\mathbf{K}_{\mu,T}^c=\varphi(\mathbf{X})^{\mathrm{T}}\varphi\left(\boldsymbol{\mu}_T^c\right)=\kappa\left(\mathbf{X},\boldsymbol{\mu}_T^c\right)$, then the proposed DKTL model (7) can be reformulated as

$$\begin{aligned}&\min_{\mathbf{\Phi},\mathbf{Z}}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right\|_{\mathrm{F}}^2+\lambda\cdot\left(\frac{1}{C}\sum_{c=1}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,S}^c-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,T}^c\right\|_2^2-\right.\\&\left.\frac{2}{C(C-1)}\sum_{t\in\{S,T\}}\alpha_t\sum_{c,k=1,c\neq k}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^c-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^k\right\|_2^2\right)+\tau\cdot\|\mathbf{Z}\|_{2,1}\\&s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}=\mathbf{I},\ \alpha_S+\alpha_T=1,\lambda,\tau,\alpha_S,\alpha_T>0\end{aligned}\tag{8}$$

where $\mathbf{K}$, $\mathbf{K}_S$, $\mathbf{K}_T$ denote the kernel Gram matrix, $\mathbf{K}_{\mu,S}^c$ and $\mathbf{K}_{\mu,T}^c$ denote the kernel mean vectors with respect to class c, $\kappa(\cdot)$ represents the kernel function. In this paper, the Gaussian kernel function is used, and it can be represented with kernel parameter σ by

$$\kappa(\mathbf{x},\mathbf{y})=\exp\left(-\|\mathbf{x}-\mathbf{y}\|_2^2/2\sigma^2\right)\tag{9}$$

From eq.(8), it is clear that this is a non-convex optimization problem with respect to two variables $\mathbf{\Phi}$ and $\mathbf{Z}$. For solutions, a variable alternating optimization algorithm is proposed.

C. Optimization

The optimization of DKTL model (8) is presented in this section. From eq.(8), there are two variables $\mathbf{\Phi}$ and $\mathbf{Z}$ in the model. When fix one of them, the model is convex with respect to the other one. Therefore, a variable alternating optimization algorithm is proposed for solving the minimization problem.

✧ Update $\mathbf{\Phi}$:

By fixing $\mathbf{Z}$, the problem (8) with respect to $\mathbf{\Phi}$ becomes

$$\begin{aligned}&\min_{\mathbf{\Phi}}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right\|_{\mathrm{F}}^2+\lambda\cdot\left(\frac{1}{C}\sum_{c=1}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,S}^c-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,T}^c\right\|_2^2-\right.\\&\left.\frac{2}{C(C-1)}\sum_{t\in\{S,T\}}\alpha_t\sum_{c,k=1,c\neq k}^{C}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^c-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^k\right\|_2^2\right)\\&s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}=\mathbf{I},\ \alpha_S+\alpha_T=1,\lambda,\tau,\alpha_S,\alpha_T>0\end{aligned}\tag{10}$$

The problem (10) can be further simplified as

$$\begin{aligned}&\min_{\mathbf{\Phi}}Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}\mathbf{\Phi}\right)\\&s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}=\mathbf{I}\end{aligned}\tag{11}$$

where $\mathbf{A}$, $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ can be represented as

$$\mathbf{A}=\mathbf{A}_1+\lambda\cdot\mathbf{A}_2-\lambda\cdot\mathbf{A}_3\tag{12}$$

$$\mathbf{A}_1=\left(\mathbf{K}_T-\mathbf{K}_S\mathbf{Z}\right)\left(\mathbf{K}_T-\mathbf{K}_S\mathbf{Z}\right)^{\mathrm{T}}\tag{13}$$

$$\mathbf{A}_2=\frac{1}{C}\sum_{c=1}^{C}\left(\mathbf{K}_{\mu,S}^c-\mathbf{K}_{\mu,T}^c\right)\left(\mathbf{K}_{\mu,S}^c-\mathbf{K}_{\mu,T}^c\right)^{\mathrm{T}}\tag{14}$$

$$\mathbf{A}_3=\frac{2}{C(C-1)}\sum_{t\in\{S,T\}}\alpha_t\sum_{c,k=1,c\neq k}^{C}\left(\mathbf{K}_{\mu,t}^c-\mathbf{K}_{\mu,t}^k\right)\left(\mathbf{K}_{\mu,t}^c-\mathbf{K}_{\mu,t}^k\right)^{\mathrm{T}}\tag{15}$$

The deduction of eq.(11) from eq.(10) is presented in Appendix A. Therefore, the optimal $\mathbf{\Phi}$ can be spanned by the first l eigenvectors with respect to the first l smallest eigenvalues of the matrix $\mathbf{K}^{-1}\mathbf{A}$. The optimization of (11) is shown in Appendix B. Note that in computing $\mathbf{A}_1$, the initialized $\mathbf{Z}$ is needed. Therefore, we initialize $\mathbf{Z}=\mathbf{K}_S^{\mathrm{T}}\mathbf{K}_T$ for a warm start.

The solving process of $\mathbf{\Phi}$ is summarized in Algorithm 1.

Algorithm 1. Solving $\mathbf{\Phi}$

Input: $\mathbf{K}_S$, $\mathbf{K}_T$, $\mathbf{K}_{\mu,S}^c$, $\mathbf{K}_{\mu,T}^c$, λ, d;

Procedure:

1. Initialize $\mathbf{Z}=\mathbf{K}_S^{\mathrm{T}}\mathbf{K}_T$;
2. Compute $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ using eqs.(13), (14), (15), respectively;
3. Compute $\mathbf{A}$ using eq.(12);
4. Perform Eigen-value decomposition of $\mathbf{K}^{-1}\mathbf{A}=\mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}}$;
5. Get $\mathbf{\Phi}=\mathbf{U}(:,\upsilon)$, where υ is the index of the d smallest Eigen-values;

Output: $\mathbf{\Phi}$

✧ Update Z:

By fixing $\mathbf{\Phi}$, the problem (8) is transformed into the following problem

$$\min_{\mathbf{Z}}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right\|_{\mathrm{F}}^2+\tau\cdot\|\mathbf{Z}\|_{2,1}\tag{16}$$

The second term in eq.(16) can be written as [27]

Algorithm 2. Solving **Z**

Input: $\mathbf{K}_S$, $\mathbf{K}_T$, $\mathbf{\Phi}$, τ;

Procedure:

1. Initialize $\mathbf{Z}=\mathbf{K}_S^{\mathrm{T}}\mathbf{K}_T$;
2. Compute $\mathbf{\Theta}$ using eq.(18);
3. Compute **Z** using eq.(21);

Output: **Z**;

$$\|\mathbf{Z}\|_{2,1}=Tr\left(\mathbf{Z}^{\mathrm{T}}\mathbf{\Theta}\mathbf{Z}\right) \tag{17}$$

where $\mathbf{\Theta}\in\Re^{N_S\times N_S}$ is a diagonal matrix, whose the i-th diagonal element is calculated as

$$\Theta_{ii}=\frac{1}{2\|\mathbf{Z}_i\|_2} \tag{18}$$

where $\mathbf{Z}_i$ represents the i-row of $\mathbf{Z}$.

By substituting eq.(17) into eq.(16), we have

$$\min_{\mathbf{Z}}\left\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T-\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right\|_{\mathrm{F}}^2+\tau\cdot Tr\left(\mathbf{Z}^{\mathrm{T}}\mathbf{\Theta}\mathbf{Z}\right) \tag{19}$$

As can be seen from model (19), it is differentiable with respect to **Z**. Let its derivative be 0, we have

$$\left(\mathbf{K}_S^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S+\tau\cdot\mathbf{\Theta}\right)\cdot\mathbf{Z}=\mathbf{K}_S^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T \tag{20}$$

Then, the close-form solution of **Z** can be expressed as

$$\mathbf{Z}=\left(\mathbf{K}_S^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S+\tau\cdot\mathbf{\Theta}\right)^{-1}\mathbf{K}_S^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T \tag{21}$$

The optimization of **Z** is summarized in Algorithm 2.

For **Z**, the closed form solution can be achieved. However, in computing $\mathbf{\Theta}$, the initialized **Z** is needed. Therefore, for achieving an optimal $\mathbf{Z}^*$, some iterations are necessary.

Therefore, the whole optimization process of the proposed DKTL model (8) is summarized in Algorithm 3.

IV. Experiments

In this section, the experiments on several benchmark datasets, two office object datasets, Multi-PIE face dataset, and three handwritten digits datasets, have been conducted for evaluating the proposed DKTL method. It is worth noting that for classification tasks, the regularized least square (RLS) method is used.

A. Cross-domain Object Recognition

In experiments, we test our method in two standard domain adaptation benchmark data: 3DA and 4DA. The 3DA and 4DA datasets are illustrated as follows

3DA: *Amazon*, *DSLR* and *Webcam* [9].

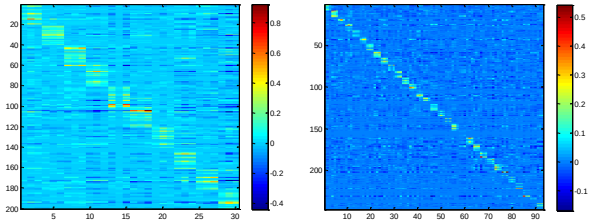
Some examples from four domains are shown in Fig. 1.

Algorithm 3. Proposed DKTL

Input: $\mathbf{K}_S$, $\mathbf{K}_T$, $\mathbf{K}_{\mu,S}^c$, $\mathbf{K}_{\mu,T}^c$, λ, τ, d, $T_{\max}$;

1. Initialize $\mathbf{Z}=\mathbf{K}_S^{\mathrm{T}}\mathbf{K}_T$ and t=1;
2. **While** not converged ($t<T_{\max}$) **do**
3. Update $\mathbf{\Phi}$ by using Algorithm 1;
4. Update **Z** by using Algorithm 2;
5. Compute the objective function value using eq.(8)
6. $t=t+1$;
7. **Until** Convergence;

Output: $\mathbf{Z}^*$ and $\mathbf{\Phi}^*$;



(a) **Z** on 4DA data (b) **Z** on 3DA data

Fig. 3. Visualization of solved representation coefficients **Z**

It's clear that 3DA dataset includes 4106 samples from three domains, where each domain contains 31 object classes, such as back-pack, keyboard, earphone, etc. By following [9], the 800-bin SURF features are used. 5 random splits of the training data in the source and target domain are implemented and the mean accuracies over 31 categories for a single source domain and multiple source domains adaptation are reported in Table I and Table II, respectively. We compare with five methods, including ASVM [8], GFK [18], SGF [4], RDALR [21] and LTSL [20]. From the results, we can observe that LSDT with nonlinear kernel function performs much better results than other methods for single source domain adaptation. For multiple source domains adaptation, DKTL outperforms other methods. Additionally, LTSL outperforms RDALR method to a large extent. Therefore, LTSL is compared in the following experiments.

4DA: *Amazon*, *DSLR*, *Webcam* and *Caltech* [18].

In 4DA dataset, four domains with 2433 samples are included, where each domain contains 10 common object classes selected from 3DA dataset and an extra Caltech 256 dataset. In experiments, the deep convolutional activation feature (DeCAF) of 4DA data is exploited [28]. The CNN with 5 convolutional layers and 3 fully-connected layers is trained on ImageNet-1000 [29]. For deep feature representation of 4DA, the outputs of the 7$^{\text{th}}$ fully-connected layer are used as deep features of the 4DA dataset.

We strictly follow the experimental setting by Gong et al.[18], 20 random splits of training data are used, and the mean classification accuracies on CNN deep features are reported in Table III. We have compared to state-of-the-art methods. From Table III, we clearly observe that DKTL performs much better than state-of-the-art LTSL results and also prior to other methods.

TABLE I. CLASSIFICATION ACCURACY (%) OVER 31 OBJECT CATEGORIES OF SINGLE SOURCE DOMAIN ADAPTATION IN 3DA SETTING

Source domain	Target domain	ASVM [8]	GFK [18]	SGF [4]	RDALR [21]	LTSL [20]	DKTL
Amazon	Webcam	42.2±0.9	46.4±0.5	45.1±0.6	50.7±0.8	**53.5**±0.4	53.0±0.8
DSLR	Webcam	33.0±0.8	61.3±0.4	61.4±0.4	36.9±1.9	62.4±0.3	**65.7±0.4**
Webcam	DSLR	26.0±0.7	66.3±0.4	63.4±0.5	32.9±1.2	63.9±0.3	**73.3±0.5**

TABLE II. CLASSIFICATION ACCURACY (%) OVER 31 OBJECT CATEGORIES OF MULTIPLE SOURCE DOMAINS ADAPTATION IN 3DA SETTING

Source domain	Target domain	ASVM [8]	GFK [18]	SGF [4]	RDALR [21]	LTSL [20]	DKTL
Amazon, DSLR	Webcam	30.4±0.6	34.3±0.6	31.0±1.6	36.9±1.1	55.3±0.3	60.0**±0.5**
Amazon, Webcam	DSLR	25.3±1.1	52.0±0.8	25.0±0.4	31.2±1.3	57.7±0.4	**63.7±0.7**
DSLR, Webcam	Amazon	17.3±0.9	21.7±0.5	15.0±0.4	20.9±0.9	20.0±0.2	**22.0±0.4**

TABLE III. CLASSIFICATION ACCURACY (%) OF DIFFERENT DOMAIN ADAPTATION BASED ON CNN FEATURE IN 4DA SETTING

Method	A→D	C→D	A→C	W→C	D→C	D→A	W→A	C→A	C→W	A→W
Na ïveComb	94.1±0.8	92.8±0.7	83.4±0.4	81.2±0.4	82.7±0.4	90.9±0.3	90.6±0.2	90.3±0.2	90.6±0.8	91.1±0.8
SGF [4]	92.0±1.3	92.4±1.1	77.4±0.7	76.8±0.7	78.2±0.7	88.0±0.5	86.8±0.7	89.3±0.4	87.8±0.8	88.1±0.8
GFK [18]	94.3±0.7	91.9±0.8	79.1±0.7	76.1±0.7	77.5±0.8	90.1±0.4	85.6±0.5	88.4±0.4	86.4±0.7	88.6±0.8
LTSL [20]	94.5±0.5	93.5±0.8	85.4±0.1	82.6±0.3	84.8±0.2	91.9±0.2	91.0±0.2	90.9±0.1	90.8±0.7	91.5±0.5
DKTL	**96.6±0.5**	**94.3±0.6**	**86.7±0.3**	**84.0±0.3**	**86.1±0.4**	**92.5±0.3**	**91.9±0.3**	**92.4±0.1**	**92.0±0.9**	**93.0±0.8**

TABLE IV. COMPARISON WITH OTHER METHODS FOR FACE RECOGNITION ACROSS POSES AND EXPRESSION

Tasks	Source	Target	Na ïveComb	ASVM [8]	SGF [4]	GFK [18]	LTSL [20]	DKTL
Session 1	frontal	60° pose	52.0	52.0	53.7	56.0	61.0	**66.0**
Session 2	frontal	60° pose	55.0	56.7	55.0	58.7	62.7	**71.0**
Session 1+2	frontal	60° pose	54.5	55.1	53.8	56.3	60.2	**69.5**
Cross session	Session 1	Session 2	93.6	97.2	92.5	96.7	97.2	**99.4**

TABLE V. HANDWRITTEN DIGITS RECOGNITION PERFORMANCE ACROSS DIFFERENT DOMAINS

Source domain	Target domain	Na ïveComb	A-SVM [8]	SGF [4]	GFK [18]	LTSL [20]	DKTL
MINIST	USPS	78.8±0.5	78.3±0.6	79.2±0.9	82.6±0.8	78.4±0.7	**88.0±0.4**
SEMEION	USPS	83.6±0.3	76.8±0.4	77.5±0.9	82.7±0.6	83.4±0.3	**85.8±0.4**
MINIST	SEMEION	51.9±0.8	70.5±0.7	51.6±0.7	70.5±0.8	50.6±0.4	**74.9±0.4**
USPS	SEMEION	65.3±1.0	74.5±0.6	70.9±0.8	76.7±0.3	64.5±0.7	**81.6±0.4**
USPS	MINIST	71.7±1.0	73.2±0.8	71.1±0.7	74.9±0.9	71.2±1.0	**79.0±0.6**
SEMEION	MINIST	67.6±1.2	69.3±0.7	66.9±0.6	74.5±0.6	66.8±1.2	**77.3±0.7**

B. Cross-poses Face Recognition

The CMU Multi-PIE face dataset [30] is a comprehensive face dataset of 337 subjects, in which the images are captured across 15 poses, 20 illuminations, 6 expressions and 4 different sessions. For our purpose, we select the first 60 subjects from session 1 and session 2 in experiments. Session 1 contains 7 images per subject with 7 poses under neutral expression, while session 2 was prepared with the same poses as session 1 under smile expression. Four cross-domain recognition tasks are as follows.

- ✧ Session 1 (cross-poses): one frontal face and an extreme pose with 60° angle for each subject are used as source and target data, respectively. The remaining faces are used as probe faces.
- ✧ Session 2 (cross-poses): the same configuration as session 1 is conducted on session 2.
- ✧ Session 1+2 (cross-poses): Two frontal faces and two faces with extreme 60° pose from both sessions are selected as source and target data. The remaining faces with poses are used as probe faces.
- ✧ Cross session: The faces in session 1 with neural expression are taken as source data, while the faces in session 2 with smile expression are taken as target data.

Fig. 4 describes some examples faces of one subject which consists of two sessions (neutral vs. smile expressions) From Fig. 4, we can observe the highly nonlinear domain mismatch between frontal faces and posed faces, while the domain mismatch between neutral and smile faces with the same view is slightly insignificant.

Fig. 4. Example of one subject. Session 1 (the 1st row with neutral expression) and Session 2 (the 2nd row with smile expression)

The face recognition results by using different methods are shown in Table IV. From the results, we can see that the proposed DKTL method outperforms LTSL and others. This demonstrates that linear subspace transfer may not deal with such nonlinear rotation well. For cross-session task, the recognition gap is small due to that expression change is easier to be adapted than pose.

C. Cross-domain Handwritten Digits Recognition

Three handwritten digits datasets, MINIST [31], USPS [32] and SEMEION [32] are used for evaluating the proposed cross

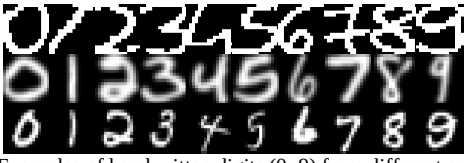


Fig. 5. Examples of handwritten digits (0~9) from different sources: SEMEION (1st row), USPS (2nd row) and MINIST (3rd row)

domain learning method. The classification accuracies over 10 classes from digit 0~9 are reported for different tasks. The MINIST handwritten digits dataset consists of 70,000 instances with each image size of 28×28, the USPS dataset contains 9298 examples with each image size of 16×16, and the SEMEION dataset contains 2593 images with each image size of 16×16. For dimension consistency, the size of MINIST digit images is manually resized as 16×16. The example images for each class of MINIST, USPS and SEMEION are shown in Fig.12, from which we can clearly observe the significant domain mismatch across different domains.

In experiment, cross-domain tests are explored. Each dataset is viewed as one domain, and there are 6 combinations of cross-domain tasks. For the purpose of our experiments, we randomly select 100 samples per class from a source domain for training and 10 samples per class from the target domain for testing. In this way, 5 random splits are generated and the average accuracies with parameter tuning are reported in Table V, in which A-SVM [8], SGF [4], GFK [18] and LTSL [20] are compared with our proposed DKTL method. From the results, we can see that the proposed method outperforms other methods to a large extent.

With the above experiments on several benchmark datasets, we can observe the competitive effectiveness of the proposed DKTL method via $l_{2,1}$-norm minimization. The proposed joint *domain-class-consistency* by using a kernel sparse representation and discriminative cross-domain subspace learning shows a new perspective and interest of transfer learning.

V. Conclusion

In this paper, we propose a discriminative kernel transfer learning (DKTL) via $l_{2,1}$-norm minimization. In the model, the domain-class-consistency (DCC) is proposed which takes into account the domain consistency and class consistency simultaneously. To this end, in subspace learning, the discriminative learning for strengthening the importance of between-domain intra-class consistency and within-domain inter-class inconsistency is integrated. For reducing the domain inconsistency, we tend to learn a representation coefficient matrix between the source data and the target data in the learned discriminative subspace. To avoid that the potential outliers in source domain are transferred to the target domain in representation, the $l_{2,1}$-norm constraint is imposed, such that a few valuable source data points are selected during representation transfer learning. Extensive experiments on several benchmark datasets demonstrate that the effectiveness and superiority of the proposed DKTL method.

Appendix A

Deduction of (11)

The eq. (10) can be re-written as

$$\min_{\mathbf{\Phi}} Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right)\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_T - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_S\mathbf{Z}\right)^{\mathrm{T}} + \lambda\cdot\left(Tr\left(\frac{1}{C}\cdot \sum_{c=1}^{C}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,S}^{c} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,T}^{c}\right)\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,S}^{c} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,T}^{c}\right)^{\mathrm{T}}\right)\right.$$

$$\left. - Tr\left(\frac{2}{C(C-1)}\sum_{t\in\{S,T\}}\alpha_t\cdot \sum_{c,k=1,c\neq k}^{C}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^{c} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^{k}\right)\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^{c} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}_{\mu,t}^{k}\right)^{\mathrm{T}}\right)\right)$$

$$s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi} = \mathbf{I},\ \alpha_S + \alpha_T = 1, \lambda, \tau, \alpha_S, \alpha_T > 0$$

$$\Rightarrow \min_{\mathbf{\Phi}} Tr\left(\mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{K}_T - \mathbf{K}_S\mathbf{Z}\right)\left(\mathbf{K}_T - \mathbf{K}_S\mathbf{Z}\right)^{\mathrm{T}}\mathbf{\Phi}\right) + \lambda\cdot\left(Tr\left(\mathbf{\Phi}^{\mathrm{T}}\frac{1}{C}\cdot \sum_{c=1}^{C}\left(\mathbf{K}_{\mu,S}^{c} - \mathbf{K}_{\mu,T}^{c}\right)\left(\mathbf{K}_{\mu,S}^{c} - \mathbf{K}_{\mu,T}^{c}\right)^{\mathrm{T}}\mathbf{\Phi}\right) - Tr\left(\mathbf{\Phi}^{\mathrm{T}}\frac{2}{C(C-1)}\sum_{t\in\{S,T\}}\alpha_t\cdot \sum_{c,k=1,c\neq k}^{C}\left(\mathbf{K}_{\mu,t}^{c} - \mathbf{K}_{\mu,t}^{k}\right)\left(\mathbf{K}_{\mu,t}^{c} - \mathbf{K}_{\mu,t}^{k}\right)^{\mathrm{T}}\mathbf{\Phi}\right)\right)$$

$$s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi} = \mathbf{I},\ \alpha_S + \alpha_T = 1, \lambda, \tau, \alpha_S, \alpha_T > 0$$

$$\Rightarrow \min_{\mathbf{\Phi}} Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}_1\mathbf{\Phi}\right) + \lambda\cdot\left(Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}_2\mathbf{\Phi}\right) - Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}_3\mathbf{\Phi}\right)\right)$$

$$s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi} = \mathbf{I},\ \alpha_S + \alpha_T = 1, \lambda, \tau, \alpha_S, \alpha_T > 0$$

$$\Rightarrow \min_{\mathbf{\Phi}} Tr\left(\mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{A}_1 + \lambda\cdot\mathbf{A}_2 - \lambda\cdot\mathbf{A}_3\right)\mathbf{\Phi}\right)$$

$$= \min_{\mathbf{\Phi}} Tr\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}\mathbf{\Phi}\right)$$

$$s.t.\ \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi} = \mathbf{I},\ \alpha_S + \alpha_T = 1, \lambda, \tau, \alpha_S, \alpha_T > 0$$

where $\mathbf{A}$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$ are represented in eqs.(12), (13), (14) and (15), respectively.

Appendix B

Optimization of (11)

According to eq.(11), the Lagrange multiplier function *Lag* can be expressed as

$$Lag(\mathbf{\Phi}, \rho) = \mathbf{\Phi}^{\mathrm{T}}\mathbf{A}\mathbf{\Phi} - \rho\cdot\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi} - \mathbf{I}\right) \tag{22}$$

where ρ>0 represents the Lagrange multiplier.

By setting the derivative of (22) with respect to $\mathbf{\Phi}$ as 0, one can obtain

$$\mathbf{A}\mathbf{\Phi} = \rho\cdot\mathbf{K}\mathbf{\Phi} \rightarrow \mathbf{K}^{-1}\mathbf{A}\mathbf{\Phi} = \rho\cdot\mathbf{\Phi} \tag{23}$$

From (23), we can get that $\mathbf{\Phi}$ can be solved by using the following Eigen-value decomposition

$$\mathbf{K}^{-1}\mathbf{A} = \mathbf{U}\Sigma\mathbf{U}^{\mathrm{T}} \tag{24}$$

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 61401048, the Fundamental Research Funds for the Central Universities and the Hong Kong Scholar Program under Grant XJ2013044.

REFERENCES

[1] H. Daumé "Frustratingly easy domain adaptation," *ACL*, 45: 256-263, 2007.

[2] G. Griffin, A. Holub, and P. Perona. "Caltech-256 object category dataset," Tech.rep. 2007.

[3] S.J.Pan, Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, 2010.

[4] R. Gopalan, R. Li, and R. Chellappa. "Domain adaptation for object recognition: An unsupervised approach," *ICCV*, 2011.

[5] B. Kulis, K. Saenko, and T. Darrell. "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," *CVPR*, 20-25, 2011.

[6] J. Hoffman, E. Rodner, J. Donahue, B. Kulis, K. Saenko. "Asymmetric and Category Invariant Feature Transformations for Domain Adaptation," *Int. J. Comput. Vis.*, 109: 28-41, 2014.

[7] L. Duan, W. Tsang, and D. Xu. "Domain Transfer Multiple Kernel Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3): 465-479, 2012.

[8] J. Yang, R. Yan, and A. Hauptmann. "Cross-domain video concept detection using adaptive SVMs," *ACM MM*. 2007.

[9] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. "Adapting visual category models to new Domains," *ECCV*, 2010.

[10] L. Zhang and D. Zhang, "Domain Adaptation Extreme Learning Machines for Drift Compensation in E-nose Systems," *IEEE Trans. Instru. Meas.*, vol. 64, no. 7, pp. 1790-1801, 2015.

[11] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Large scale transductive SVMs," *Journal of Machine Learning Research*, vol. 7, pp. 1687-1712, 2006.

[12] L. Duan, D. Xu, and I. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 504-518, 2012.

[13] L. Duan, D. Xu, W. Tsang, and J. Luo. Visual Event Recognition in Videos by Learning from Web Data. *IEEE Trans. Pattern Anal. Mach. Intell*., 34(9): 1667-1680, 2012.

[14] L. Zhang and D. Zhang, "Domain Adaptation Transfer Extreme Learning Machines," *Proc. Int' Conf. Elm*, vol. 1, pp. 103-119, 2014.

[15] L. Zhang and D. Zhang, "Robust Visual Knowledge Transfer via EDA," arXiv, 2015.

[16] S. Shekhar, V.M. Patel, H.V. Nguyen, and R. Chellappa. "Generalized Domain-Adaptive Dictionaries," *CVPR*, 361-368, 2013.

[17] F. Zhu and L. Shao, "Weakly-Supervised Cross-Domain Dictionary Learning for Visual Recognition," *International Journal of Computer Vision*, vol. 109, no. 1, pp. 42-59, 2014.

[18] B. Gong, Y. Shi, F. Sha, and K. Grauman. "Geodesic flow kernel for unsupervised domain adaptation," *CVPR*, 2066-2073, 2012.

[19] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised Visual Domain Adaptation Using Subspace Alignment," *ICCV*, pp. 2960-2967, 2013.

[20] M. Shao, D. Kit, and Y. Fu. "Generalized Transfer Subspace Learning Through Low-Rank Constraint," *Int. J. Comput. Vis.*, 109: 74-93, 2014.

[21] I.H. Jhuo, D. Liu, D. Lee, and S.F. Chang. "Robust visual domain adaptation with low-rank reconstruction," *CVPR*, 2168-2175, 2012.

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

[23] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1): 171-184, 2013.

[24] E. Elhamifar and R. Vidal. Sparse subspace clustering. *CVPR*, 2790-2797, 2009.

[25] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11): 2675-2781, 2013.

[26] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," *IEEE Trans. Cybernetics*, vol. 44, no. 6, pp. 793-804, 2014.

[27] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and Robust Feature Selection via Joint $l_{2,1}$-Norm Minimization," *NIPS*, 2010.

[28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," *ICML*, 2014.

[29] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet classification with deep convolutional neural networks," *NIPS*, 2012.

[30] R. Gross, I. Matthews, J.F. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Computing*, 28(5): 807-813, 2010.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[32] A. Frank and A. Asuncion, (2010) UCI machine learning repository [Online]. Available: http://archive.ics.uci.edu/ml