Joint Feature Selection and Structure Preservation for Domain Adaptation

Jingjing Li, Jidong Zhao and Ke Lu
University of Electronic Science and Technology of China, Chengdu, China
lijin117@yeah.net, {jdzhao, kel}@uestc.edu.cn

Abstract

The essence of domain adaptation is to explore common latent factors shared by the involved domains. These factors can be specific features or geometric structures. Most of previous methods exploit either the shared features or the shared geometric structures separately. However, the two strategies are complementary with each other and jointly exploring them is more optimal. This paper proposes a novel approach, named joint *Feature Selection and Structure Preservation* (FSSP), for unsupervised domain adaptation. FSSP smoothly integrates structure preservation and feature selection into a unified optimization problem. Intensive experiments on text categorization, image classification and video event recognition demonstrate that our method performs better, even with up to 30% improvement in average, compared with the state-of-the-art methods.

1 Introduction

As poet Sándor Petőfi once wrote, "life is dear, love is dearer," a scientist in the filed of machine learning might say, "data is dear, labeled data is dearer." How to acquire more labeled data from existing ones has been a crucial research topic recently. Domain adaptation [Pan and Yang, 2010] proves to be effective for leveraging labeled data in the well-labeled source domain to transfer classification discriminability to the unlabeled target domain.

Domain adaptation deals with the problem where data from two domains have common class label but divergent data distributions. Since traditional machine learning algorithms would fail to handle the situation, domain adaptation, one category of transfer learning [Pan and Yang, 2010], has been widely studied in many real world applications, e.g., image classification [Long *et al.*, 2014b], text categorization [Ding *et al.*, 2015] and video event recognition [Duan *et al.*, 2012b].

The basic assumption of domain adaptation is that some common latent factors are shared by the involved domains. Therefore, the mechanism of domain adaptation is to explore these common latent factors, and utilize them to mitigate both the marginal and conditional distributions across domains, which can be done by one of the two strategies, i.e., instance re-weighting and feature extraction. Most approaches in the first group [Chu *et al.*, 2013] try to train a sophisticated classifier on the source domain, e.g., multiple kernel SVM, which can be used in the target domain. Whereas approaches in the second group aim to preserve important data properties, e.g., statistical property and geometric structure. In most cases, feature extraction is more effective than training a complex classifier, and deep learning [Donahue *et al.*, 2013] is a great example. In this paper, therefore, we focus on the feature extraction. However, previous methods in this group usually preserve the statistical property and geometric structure independently, e.g., [Ding *et al.*, 2015; Pan *et al.*, 2011] explore the statistical property by maximizing the empirical likelihood, while [Gong *et al.*, 2012; Long *et al.*, 2014a] optimize predefined objective function by exploring the geometric structure. In fact, these two properties are complementary with each other and jointly exploring them could benefit from both sides. The statistical properties and geometric structure are two observations of the data from different viewpoints. Each viewpoint has its theoretical basement and exists unilateralism. Different viewpoints are not mutually exclusive and combining them normally could transcend the specific limitations of each perspective [Zhu and Lafferty, 2005].

In this paper, we explore the benefit of integrating the optimization of statistical property and geometric structure into a unified framework. On one hand, we seek a common subspace shared by the involved domains where common latent features can be uncovered and the data distribution gap across two domains can be mitigated. On the other hand, we deploy a graph structure to characterize the sample relationship. The motivation of this paper is illustrated in Fig. 1. Furthermore, if we simply select features from the source domain and the target domain by a general projection matrix, the selected results may be different for each dimension of the subspace learned from different domains [Gu *et al.*, 2011]. The upper part of Fig. 1 shows an illustrative toy example. It can be seen that the selected features (colored ones) from different domains are different too. Therefore, it is hard to distinguish which features (corresponding rows) in both domains are really redundant. However, after row-sparsity regularization, the selection tends to be clear. Finally, the main contributions of this paper are summarized as follows:

1) A unified framework of feature selection and geometric

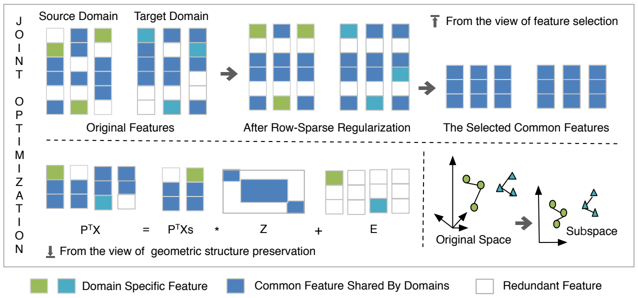

Figure 1: Illustration of our approach. The upper part shows our idea from the view of feature selection, and the bottom half shows it from the perspective of geometric structure preservation. In this paper, we are going to optimize both of them in a unified framework.

structure preservation is proposed for unsupervised domain adaptation, and it achieves state-of-the-art performance on several standard benchmarks even with up to 30% improvement in average compared with baselines.

2) In the aspect of feature selection, we deploy $\ell_{2,1}$-norm on the projection matrix, which leads to achieving row-sparsity and, as a result, selecting relevant features across the involved domains.

3) In the aspect of geometric structure preservation, not only the structure of samples is preserved by a nearest neighbor graph, but also the structure of features in the embedded space is preserved by a representation matrix.

The rest of this paper is organized as follows. Section 2 presents some brief discussion with related works. Section 3 introduces the proposed method in detail. Experiments are reported in Section 4, and Section 5 is the conclusion.

2 Related Works and Discussions

This paper focuses on domain adaptation [Pan and Yang, 2010] where the source domain and the target domain share the same task but have different data distributions.

According to the recent work [Long *et al.*, 2014b], most of the unsupervised domain adaptation approaches work by learning a new feature representation to reduce the data distribution differences among domains. The new feature representation can be learned by: 1) exploring domain-invariant common factors [Ding and Fu, 2014; Ding *et al.*, 2015], 2) minimizing proper distance measures [Gong *et al.*, 2012; Long *et al.*, 2014a], and 3) re-weighting relevant features with sparsity-promoting regularization [Gu *et al.*, 2011; Long *et al.*, 2014b]. Actually, these three groups can be concisely summed up in two bases: feature selection, which consists of 1) and 3), and geometric structure preservation. This paper aims to take full advantage of both feature selection and geometric structure preservation, incorporate them into a unified framework and jointly optimize them.

To our knowledge, this work is among the very leading works for domain adaptation to joint feature selection and geometric structure preservation. Notably, experiments

Table 1: Notations and corresponding descriptions, in which m, n and d denote the number of samples, dimensionality of original space and subspace, respectively.

Notation	Description	Notation	Description
$\mathbf{X}_s \in \mathbb{R}^{n*m_s}$,$\mathbf{y_s}$	source data/labels	$\mathbf{X} \in \mathbb{R}^{n*m}$	$\mathbf{X}_s$ and $\mathbf{X}_t$
$\mathbf{X}_t \in \mathbb{R}^{n*m_t}$,$\mathbf{y_t}$	target data/labels	$\mathbf{L} \in \mathbb{R}^{m*m}$	graph Laplacian
$\mathbf{Z} \in \mathbb{R}^{m_s*m}$	reconstruction	$\mathbf{P} \in \mathbb{R}^{n*d}$	projection matrix
$\mathbf{E} \in \mathbb{R}^{d*m}$	sparse error	$\mathbf{Y} \in \mathbb{R}^{d*m}$	eigenvector matrix
$\mathbf{G} \in \mathbb{R}^{n*n}$	sub-gradient	λ, β, γ	penalty parameters

demonstrate that our work can get better recognition accuracy than baselines with a significant advantage.

3 The Proposed Approach

3.1 Notations

In this paper, we use bold low-case symbols to represent vectors, bold upper-case symbols to represent matrices, specifically, $\mathbf{I}$ represents the identity matrix. A sample is denoted as a vector, e.g., $\mathbf{x}$, and the i-th sample in a set is represented by the symbol $\mathbf{x}_i$. For a matrix $\mathbf{M}$, its $\ell_{2,1}$-norm is defined as: $\|\mathbf{M}\|_{2,1} = \sum_j \sqrt{\sum_i (\mathbf{M}_{ij})^2}$. We also use the Frobenius norm $\|\mathbf{M}\|_\mathrm{F} = \sqrt{\sum_i \delta_i(\mathbf{M})^2}$, where $\delta_i(\mathbf{M})$ is the i-th singular value of the matrix $\mathbf{M}$. The trace of matrix $\mathbf{M}$ is represented by $\mathbf{tr}(\mathbf{M})$. For clarity, we also show the frequently used notations in Table 1.

3.2 Problem Definition

Definition 1 *A domain $\mathbb{D}$ is defined by a feature space $\mathcal{X}$ and its probability distribution $P(\mathbf{X})$, where $\mathbf{X} \in \mathcal{X}$. For a specific domain, a classification task $\mathbb{T}$ consists of class information $\mathcal{Y}$ and a classifier $f(\mathbf{x})$, that is $\mathbb{T} = \{\mathcal{Y}, f(\mathbf{x})\}$.*

We use subscripts s and t to indicate the source domain and the target domain, respectively. This paper focuses on the following problem:

Problem 1 *Given a labeled source domain $\mathbb{D}_s$ and an unlabeled target domain $\mathbb{D}_t$, where $\mathbb{D}_s \neq \mathbb{D}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P(\mathbf{X}_\mathrm{s}) \neq P(\mathbf{X}_\mathrm{t})$ and $P(\mathbf{y}_s|\mathbf{X}_s) \neq P(\mathbf{y}_t|\mathbf{X}_t)$, find a subspace spanned by $\mathbf{P}$ in which the common latent features shared by involved domains are uncovered, the data manifold structure is preserved, and the domain shift is minimized.*

3.3 Problem Formulation

The basic assumption behind domain adaptation is that the involved domains share some common latent factors, these factors can be specific features or geometric structures, and in most cases, are both of them. If we assume that there exists a common latent subspace shared by both domains where the shared features can be uncovered, then we can find the subspace spanned by an appropriate basis $\mathbf{P}$ where each sample from the target domain can be drawn from one subspace segmentation in the source domain. Thus, the goal of **Problem 1** can be formulated as optimizing the following objective:

$$\min_{\mathbf{P},\mathbf{Z}} \|\mathbf{P}^\top \mathbf{X}_t - \mathbf{P}^\top \mathbf{X}_s \mathbf{Z}\|_\mathrm{F}^2, \tag{1}$$

where $\mathbf{P}$ is the projection matrix, $\mathbf{Z}$ is the reconstruction coefficient matrix corresponding to $\mathbf{X}_s$, and $\mathbf{X}_s$ serves as a dictionary [Qiu *et al.*, 2012]. Since $\mathbf{X}_s$ can represent $\mathbf{X}_t$ by appropriate $\mathbf{P}$ and $\mathbf{Z}$, and it is no doubt that $\mathbf{X}_s$ can represent itself too. Therefore, we combine $\mathbf{X}_s$ and $\mathbf{X}_t$ together as complete $\mathbf{X}$ to dig out more shared information. From [Yin *et al.*, 2015], there are two explanations for $\mathbf{Z}$ based on the model. Firstly, the ij-th element of $\mathbf{Z}$ reflects the similarity between the sample pair $\mathbf{x}_i$ and $\mathbf{x}_j$. Secondly, the i-th column of $\mathbf{Z}$ severs as a better representation of $\mathbf{x}_i$ such that the desired pattern, say subspace structure, is more prominent. From this perspective, $\mathbf{Z}$ preserves the embedding manifold structure of samples. Furthermore, in order to learn a robust and efficient subspace, we introduce the Frobenius norm of $\mathbf{Z}$ according to [Lu *et al.*, 2012a].

As we have discussed in the introduction, the selected features by a general $\mathbf{P}$ may be different for each dimension of the learned subspace, especially in the case of domain adaptation where $\mathbf{X}_s$ and $\mathbf{X}_t$ have divergent data distributions. This motivates us to deploy $\ell_{2,1}$-norm on $\mathbf{P}$ [Lu *et al.*, 2012b], which leads to selecting common features shared by the domains. As a result, we can further formulate our objective function as follows:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Z},\mathbf{E}} \|\mathbf{P}\|_{2,1} + \tfrac{\beta}{2}\|\mathbf{Z}\|_{\mathrm{F}}^2 + \gamma\|\mathbf{E}\|_1 \\ s.t.\ \mathbf{P}^\top\mathbf{X} = \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} + \mathbf{E},\end{aligned} \tag{2}$$

where $\mathbf{E}$ is used to detect the sample specific errors. $\beta > 0$ and $\gamma > 0$ are penalty parameters. Please note that $\mathbf{E}$ can be very helpful when samples are corrupted, and it is also very useful to handle outliers because it is very difficult to guarantee that every sample in the target domain can be appropriately reconstructed by the source domain.

Finally, as we have discussed above, the common latent factors shared by domains are not only specific features, but also geometric structure. With the goal to jointly select features and preserve geometric structure, we introduce a graph based regularization term into our objective. Thus, the final objective function can be formulated as follows:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Z},\mathbf{E}} \|\mathbf{P}\|_{2,1} + \tfrac{\lambda}{2}\mathbf{tr}(\mathbf{P}^\top\mathbf{X}\mathbf{L}\mathbf{X}^\top\mathbf{P}) + \tfrac{\beta}{2}\|\mathbf{Z}\|_{\mathrm{F}}^2 + \gamma\|\mathbf{E}\|_1 \\ s.t.\ \mathbf{P}^\top\mathbf{X} = \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} + \mathbf{E}, \mathbf{P}^\top\mathbf{X}\mathbf{D}\mathbf{X}^\top\mathbf{P} = \mathbf{I},\end{aligned} \tag{3}$$

where $\lambda > 0$ is a penalty parameter. $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian [Chung, 1997] and $\mathbf{D} = \sum_{\mathbf{j}} \mathbf{W}_{\mathbf{ij}}$ is a diagonal matrix. $\mathbf{I}$ is the identity matrix with proper size. The constraint $\mathbf{P}^\top\mathbf{X}\mathbf{D}\mathbf{X}^\top\mathbf{P} = \mathbf{I}$ is introduced to avoid trivial solutions. $\mathbf{W}$ is a symmetric adjacency matrix with $\mathbf{W}_{ij}$ characterizes the appropriate connection among the samples [Li *et al.*, 2016], it can be computed by various criteria [Yan *et al.*, 2007]. In this paper, we use $\mathbf{W}$ to characterize the sample relationship and apply the heat kernel method to get $\mathbf{W}$ as follows:

$$\mathbf{W}_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}) & \text{, if } \mathbf{x}_i \in k\mathrm{NN}(\mathbf{x}_j) \\ 0 & \text{, otherwise} \end{cases}, \tag{4}$$

where $k\mathrm{NN}(\mathbf{x}_j)$ is the k-nearest neighbors of $\mathbf{x}_j$.

3.4 Problem Optimization

Since the constraint in Eq. (3) is not convex, we convert Eq. (3) to the following equivalent equation to make it easier to optimize.

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Z},\mathbf{E}} \|\mathbf{P}\|_{2,1} + \tfrac{\lambda}{2}\|\mathbf{P}^\top\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 + \tfrac{\beta}{2}\|\mathbf{Z}\|_{\mathrm{F}}^2 + \gamma\|\mathbf{E}\|_1 \\ s.t.\ \mathbf{P}^\top\mathbf{X} = \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} + \mathbf{E},\end{aligned} \tag{5}$$

where $\mathbf{Y}$ is a matrix whose rows are eigenvectors of the eigen-problem $\mathbf{W}\mathbf{Y} = \mathbf{\Lambda}\mathbf{D}\mathbf{Y}$, and $\mathbf{\Lambda}$ is a diagonal matrix of which diagonal elements are eigenvalues. The equivalence proof of Eq. (3) and Eq. (5) can be found in [Cai *et al.*, 2007; Gu *et al.*, 2011].

Now, Eq. (5) can be optimized by the augmented Lagrangian multiplier (ALM) [Lin *et al.*, 2010]. First, we transform Eq. (5) into the augmented Lagrangian function:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{Z},\mathbf{E}} \|\mathbf{P}\|_{2,1} + \tfrac{\lambda}{2}\|\mathbf{P}^\top\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 + \tfrac{\beta}{2}\|\mathbf{Z}\|_{\mathrm{F}}^2 + \gamma\|\mathbf{E}\|_1 + \\ \mathbf{tr}(\mathbf{U}^\top(\mathbf{P}^\top\mathbf{X} - \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} - \mathbf{E})) + \tfrac{\mu}{2}\|\mathbf{P}^\top\mathbf{X} - \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} - \mathbf{E}\|_{\mathrm{F}}^2,\end{aligned} \tag{6}$$

where $\mu > 0$ is a penalty parameter and $\mathbf{U}$ is a Lagrange multiplier. Since we cannot directly optimize all the variables in Eq. (6) at the same time, we introduce the alternating direction method of multipliers (ADMM) [Hestenes, 1969]. By deploying ADMM, we can alternately update each variable one by one in an iterative manner. Thus, Eq. (6) can be solved by the following steps:

1) To solve $\mathbf{Z}$, by taking the derivative of Eq. (6) w.r.t $\mathbf{Z}$, and setting the derivative to zero, we get:

$$\mathbf{Z} = (\mathbf{X}_{\mathrm{s}}^\top\mathbf{P}\mathbf{P}^\top\mathbf{X}_{\mathrm{s}} + \tfrac{\beta}{\mu}\mathbf{I})^{-1}\mathbf{X}_{\mathrm{s}}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{X} - \mathbf{E} + \mathbf{U}/\mu). \tag{7}$$

2) For $\mathbf{E}$, by ignoring the irrelevant terms w.r.t. $\mathbf{E}$, we can optimize $\mathbf{E}$ by:

$$\mathbf{E} = \arg\min_{\mathbf{E}} \tfrac{\gamma}{\mu}\|\mathbf{E}\|_1 + \|\mathbf{E} - (\mathbf{P}^\top\mathbf{X} - \mathbf{P}^\top\mathbf{X}_{\mathrm{s}}\mathbf{Z} + \mathbf{U}/\mu)\|_{\mathrm{F}}^2. \tag{8}$$

3) To solve $\mathbf{P}$, by taking the derivative of Eq. (6) w.r.t $\mathbf{P}$, and setting the derivative to zero, we get:

$$\mathbf{P} = \mathbf{\Phi}^{-1}((\mathbf{X} - \mathbf{X}_{\mathrm{s}}\mathbf{Z})(\mathbf{E}^\top - \mathbf{U}^\top/\mu) + \tfrac{\lambda}{\mu}\mathbf{X}\mathbf{Y}^\top), \tag{9}$$

where $\mathbf{\Phi} = (\mathbf{X} - \mathbf{X}_{\mathrm{s}}\mathbf{Z})(\mathbf{X} - \mathbf{X}_{\mathrm{s}}\mathbf{Z})^\top + \frac{\lambda}{\mu}\mathbf{X}\mathbf{X}^\top + \frac{2}{\mu}\mathbf{G}$. Please note that $\|\mathbf{P}\|_{2,1}$ is not smooth, therefore, as a surrogate, we compute its sub-gradient $\mathbf{G}$, where $\mathbf{G}$ is diagonal and its i-th diagonal element can be calculated by

$$\mathbf{G}_{ii} = \begin{cases} 0 & \text{, if } \mathbf{p}^i = \mathbf{0} \\ \frac{1}{2\|\mathbf{p}^i\|} & \text{, otherwise} \end{cases}, \tag{10}$$

where $\mathbf{p}^i$ denotes the i-th row of $\mathbf{P}$.

Problem 1 specified that the aim of this work is to find a subspace, spanned by the appropriate basis $\mathbf{P}$, in which the common latent features shared by the involved domains can be uncovered, the data manifold structure can be preserved, and the domain shift can be minimized. However, Eq. (9) shows that the optimization of $\mathbf{P}$ involves some unknown variables, e.g., $\mathbf{Z}$, $\mathbf{E}$ and $\mathbf{Y}$. To address this problem, we apply Principal Component Analysis (PCA) [Turk and Pentland, 1991] to the initialization of our algorithm. For clarity,

Algorithm 1. *Joint Feature Selection and Structure Preservation for Unsupervised Domain Adaptation*

Input: Sample sets $\mathbf{X}_t$ and $\mathbf{X}_s$, label information of $\mathbf{X}_s$,
balanced parameter λ, β and γ.

Initialize: $\mathbf{Z} = \mathbf{0}$, $\mathbf{E} = \mathbf{0}$, $\mathbf{U} = \mathbf{0}$,
$\mu = 10^{-4}, \mu_{max} = 10^{6}, \rho = 1.3, \epsilon = 10^{-5}$.

Output: Label information of $\mathbf{X}_t$.

1. Initialize $\mathbf{P}_0$ by PCA.
2. Computer $\mathbf{W}$, $\mathbf{D}$ and $\mathbf{L}$.
3. Learn $\mathbf{Y}$ by solving the eigen-problem $\mathbf{WY} = \mathbf{\Lambda DY}$.

Repeat

4. Get $\mathbf{Z}$ and $\mathbf{E}$ by Eq. (7) and Eq. (8), respectively.
5. Optimize $\mathbf{P}$ and $\mathbf{G}$ by Eq. (9) and Eq. (10), respectively.
6. Update the multiplier via
$\mathbf{U}_{new} = \mathbf{U}_{old} + \mu(\mathbf{P}^\top\mathbf{X} - \mathbf{P}^\top\mathbf{X}_s\mathbf{Z} - \mathbf{E})$.
7. Update μ via $\mu_{new} = min(\rho\mu_{old}, \mu_{max})$.
8. Check the convergence condition
$\|\mathbf{P}^\top\mathbf{X} - \mathbf{P}^\top\mathbf{X}_s\mathbf{Z} - \mathbf{E}\|_\infty < \epsilon$.

until *Convergence*

9. Project both $\mathbf{X}_t$ and $\mathbf{X}_s$ to the learned subspace by $\mathbf{P}$, that is $\mathbf{P}^\top\mathbf{X}_t$ and $\mathbf{P}^\top\mathbf{X}_s$.
10. Classify $\mathbf{X}_t$ in the subspace by Nearest Neighbor classifier, and $\mathbf{X}_s$ is used as reference.

Algorithm 1 shows the details of our method. Limited by space, please refer to [Nie *et al.*, 2010] for a similar convergence analysis of this algorithm.

3.5 Computational Complexity

The computational cost of **Algorithm 1** is composed of several major parts listed as follows:

- The eigen-problem solved in step 3.
- Matrix inversion and multiplication in step 4 and 5.

Here we analyze the computational complexity by the big O notation. For simplicity and without loss of generality, we assume the matrix which we handled are with the size of $n \times m$, and d is the dimensionality of the learned subspace where $d \ll min(m, n)$. The eigen-decomposition costs $O(dm^2)$, matrix inversion and multiplication cost a maximum of $O(m^3)$. Thus, the total cost of **Algorithm 1** is much less than $O(km^3)$ because lots of matrix operations are performed in the embedded low-dimensional space, where k indicates the number of matrix operations. When m is very large, we could adopt divide-and-conquer to address the large-scale data problem.

4 Experiments

In this section, we evaluate our algorithm on several standard benchmarks which consist of text dataset, image dataset and video dataset. We compare our algorithm with several state-of-the-art domain adaptation approaches, e.g., GFK [Gong *et al.*, 2012], TJM [Long *et al.*, 2014b], TCA [Pan *et al.*, 2011], TSL [Si *et al.*, 2010], and DLRC [Ding *et al.*, 2015]. Since we apply PCA [Turk and Pentland, 1991] to the initialization of our algorithm and 1-Nearest Neighbor as the classifier (NNC), we also compare our method with both of them. Specifically, for PCA, we use the model trained on $\mathbf{X}_s$ to recognize $\mathbf{X}_t$, and for NNC, we use $\mathbf{X}_s$ as reference to classify $\mathbf{X}_t$ in the original data space. To fully demonstrate the superiority of our method, we also compare our method with several approaches in the group of instance re-weighting strategy on the evaluations of video event recognition. All of the reported results are the classification accuracy on the target domain, which is also widely used in literature [Gong *et al.*, 2012; Long *et al.*, 2014b]:

Figure 2: Some selected samples from Caltech-256, Amazon, DSLR, Webcam, MRSC and VOC2007.

$$accuracy = \frac{|\mathbf{x} : \mathbf{x} \in \mathbf{X_t} \wedge \bar{\mathbf{y_t}} = \mathbf{y_t}|}{|\mathbf{x} : \mathbf{x} \in \mathbf{X_t}|} \quad (11)$$

where $\bar{\mathbf{y_t}}$ is the predicted label of the target domain by each approach, and $\mathbf{y}_t$ is the real label vector.

Each of the hyper-parameters used in our experiments is the optimal one chosen from a large range. We chose an acceptable common set of them for consistency. For the sake of fairness, all of the datasets used in our experiments were downloaded from the webpages of the related works, and we strictly followed the same experimental settings with them.

4.1 Data Description

Amazon, **Caltech-256**, **DSLR**, and **Webcam** (4 datasets Domain Adaptation, **4DA**) is the most popular benchmark in the field of domain adaptation. 4DA experimental setting was firstly introduced in [Gong *et al.*, 2012], which is an extension of 3DA benchmark introduced in [Saenko *et al.*, 2010]. 3DA includes object categories from Amazon (A, images downloaded from amazon.com), DSLR (D, high-resolution images by a digital SLR camera) and Webcam (W, low-resolution images by a web camera). A, D and W are three different domains, and each domain consists of 31 categories, e.g., monitor, keyboard and laptop. 4,652 images are the total number of 3DA. 4DA contains an additional domain, Caltech-256 (C) [Griffin *et al.*, 2007], which has 30,607 images and 256 categories. Some of the selected samples from 4DA are shown in Fig. 2. Our experimental configuration on 4DA is identical with [Gong *et al.*, 2012]. Specifically, 10 common classes shared by four datasets are selected. There are 8 to 151 samples per category per domain, and 2,533 images in total. Furthermore, 800 dimensional SURF features are extracted as our low-level input. Then the low-level input is normalized to unit.

Reuters-215782 is a challenging text dataset with several different categories. The widely used 3 largest top categories of Reuters-215782 are **orgs**, **people**, and **place**, each of the top categories consists of many subcategories. As suggested in [Ding *et al.*, 2015], we evaluate our approach on the preprocessed version of this dataset with the same settings of [Gao *et al.*, 2008].

Table 2: Recognition results (%) of domain adaptation on 4DA dataset. Since DLRC did not use DSLR as source domain for the reason of small sample number, we only report 9 results for DLRC.

Source	Target	PCA	NNC	TCA	GFK	TSL	TJM	DLRC	Ours
Caltech-256	Amazon	37.58	23.70	38.70	41.05	45.25	46.76	49.75	**75.78**
	Webcam	38.98	25.76	39.06	40.68	33.37	39.98	41.76	**75.25**
	DSLR	42.04	25.48	41.44	38.81	44.15	44.59	47.85	**76.43**
Amazon	Caltech-256	38.22	26.00	37.36	40.28	37.51	39.45	42.75	**79.16**
	Webcam	35.93	29.83	37.67	39.00	34.49	42.03	42.93	**75.93**
	DSLR	29.94	25.48	33.32	36.35	27.81	45.22	41.86	**74.52**
Webcam	Caltech-256	26.71	10.95	29.30	30.73	28.97	30.19	33.85	**81.83**
	Amazon	27.77	14.82	30.05	29.76	30.15	29.96	38.57	**82.98**
	DSLR	73.25	24.20	87.29	80.83	86.57	89.17	**94.31**	93.63
DSLR	Caltech-256	26.18	10.60	31.81	30.28	28.49	31.43	–	**82.81**
	Webcam	66.78	31.53	86.13	75.59	83.75	85.42	–	**93.22**
	Amazon	29.12	11.69	32.29	32.06	29.06	32.78	–	**81.21**
Average		39.38	21.67	43.70	42.95	42.46	46.42	48.18	**81.06**

MRSC+VOC consists of two different datasets: MRSC and VOC2007. MRSC dataset contains 4,323 images from 18 different classes, and it was originally provided by Microsoft Research Cambridge. VOC2007 dataset contains 5,011 images labeled by 20 classes. The two datasets share 6 common classes, i.e., "aeroplane", " bicycle", "bird", "car", "cow", and "sheep". We build our dataset by selecting all images with the common concepts. Specifically, 1,269 images from MSRC are selected to form the training data, and 1,530 images from VOC2007 are chosen to form the test data. We denote this evaluation as MSRC $\rightarrow$ VOC, and build another evaluation VOC $\rightarrow$ MSRC by switching the source/target pair. Then following the same experimental settings of [Long *et al.*, 2014b], we resize all images to be 256 pixels, 128-dimensional dense SIFT features are extracted as input. Some selected samples from this dataset can also be seen in Fig. 2.

The large scale **Columbia Consumer Video dataset (CCV)** [Jiang *et al.*, 2011] contains 9,317 web videos over 20 semantic categories, where 4,659 videos are used for training and the remaining 4,658 videos are used for testing. Binary labels (presence or absence) for each visual concept for each video were assigned. In our experiment, we use the subset which contains visual events, i.e., "Basketball", "Baseball", "Soccer", "IceSkating", "Skiing", "Swimming", "Biking", "Graduation", "Birthday", "WeddingReception", "WeddingCeremony", "WeddingDance", "MusicPerformance", "NonMusicPerformance" and "Parade". Furthermore, for a fair comparison, we deploy the same experimental settings as in [Duan *et al.*, 2012a]. Specifically, we merge the first seven events as "sports", and also merge three "Weeding-xxx" events as "wedding" and two "xxx-Performance" events as "performance". Finally, we have 5,610 videos from the six event classes in total, i.e., "sports", "graduation", "birthday", "wedding", "performance" and "parade". For each video, we use the 5,000 dimensional SIFT features offered by [Jiang *et al.*, 2011].

4.2 Implementation Details and Results

For consistency, we choose a common set of hyper-parameter settings for our FSSP on different evaluations. Specifically, we empirically set $\lambda = 0.1$, $\beta = 0.1$ and $\gamma = 1$. The dimensionality of subspace is set to 30, and the number of neighbors is set to 5. The nearest neighbor graph is learned in an unsupervised manner.

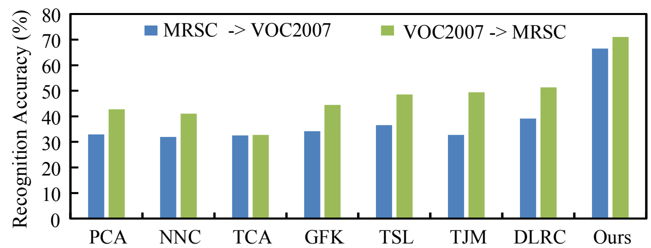

Figure 3: Recognition results on **MRSC+VOC2007**. PCA and NNC are traditional learning methods, while others are transfer learning approaches.

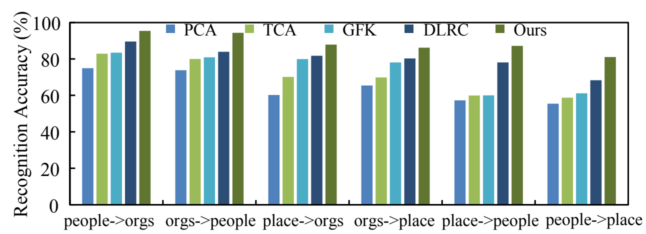

Figure 4: Recognition results on **Reuters-215782**. For better visual effect, we selected the results of 5 methods to show. DLRC represents state-of-the-art performance of baselines.

For **MRSC+VOC**, we perform two evaluations: 1) MRSC $\rightarrow$ VOC2007 and 2) VOC2007 $\rightarrow$ MRSC. For each evaluation, the first dataset serves as the source domain and is used for training, the second dataset serves as the target domain and is used for testing. The experimental results on this dataset are shown in Fig. 3.

For **Reuters-215782**, we perform six evaluations, i.e., people $\rightarrow$ orgs, orgs $\rightarrow$ people, place $\rightarrow$ orgs, orgs $\rightarrow$ place, place $\rightarrow$ people, and people $\rightarrow$ place. In each evaluation, the first dataset serves as the source domain and is used for training, the second dataset serves as the target domain and is used for

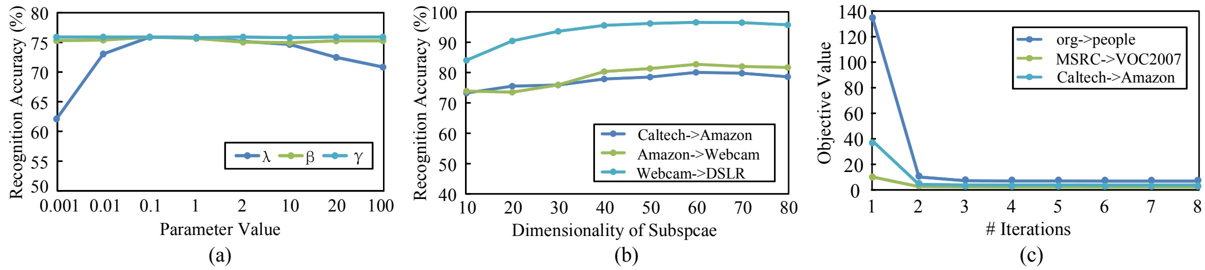

Figure 5: (a) shows the results on *Caltech* $\rightarrow$ *Amazon* with different penalty parameters. If one parameter is used for testing, the others are set as $\lambda = 0.1$, $\beta = 0.1$ and $\gamma = 1$. (b) shows the results on different datasets with varying dimensionality of subspace. (c) are the convergence curves of different evaluations.

Table 3: Experimental results (%) of domain adaptation on visual event recognition in videos.

Event	DASVM	MKMM	DAM	DSM	Ours
sports	49.79	54.50	41.27	42.31	**87.07**
graduation	7.76	7.23	7.69	7.85	**91.59**
birthday	5.63	5.67	5.61	8.16	**91.34**
wedding	10.21	17.45	14.37	20.24	**86.08**
performance	31.90	29.89	28.70	47.96	**85.19**
parade	8.45	8.52	9.43	8.37	**91.74**
average	18.96	20.54	17.85	22.48	**88.83**

testing. Limited by space, we only compare our method with PCA, TCA, GFK and DLRC on this dataset. The experimental results are shown in Fig. 4.

For **4DA**, two different datasets are randomly selected as the source domain and the target domain, respectively, thus leading to $4 \times 3 = 12$ evaluations. The recognition results are reported in Table 2.

For **CCV**, we compare our method with four widely cited domain adaptation approaches in the field: DASVM [Bruzzone and Marconcini, 2010], MKMM [Schweikert *et al.*, 2009], DAM [Duan *et al.*, 2009], and DSM [Duan *et al.*, 2012a]. These methods include algorithms in the group of instance re-weighting strategy. Thus, the experiments on this dataset can also demonstrate the superiority of our method compared with the instance re-weighting ones. The experimental results are shown in Table 3.

4.3 Discussions

From the experimental results, several observations can be drawn as follows:

1) Transfer learning methods perform much better than traditional (non-transfer) ones, which means transfer learning, or domain adaptation, is valuable and practical for real world applications.

2) All of the subspace learning approaches work much better than NNC, which means perform domain adaptation through a dimensionality reduction procedure is not trivial and the results are promising.

3) The baselines either try to maximize the empirical likelihood to explore specific features, e.g., TCA and DLRC, or aim to preserve geometric structure by minimizing proper distance measures, e.g., GFK, and each of them is the representative method in their own field, but none of them performs better than our FSSP in average. It exactly demonstrates the motivation of our work, that is jointly optimizing feature selection and geometric structure preservation is more optimal than optimizing them separately.

4) TJM performs joint feature matching and instance reweighting, but it does not consider the data structure of samples. As a result, it performs worse than our FSSP.

5) From the results reported in Table 2, it can be seen that our FSSP significantly advances state-of-the-art baselines with 30% accuracy rates in average. It is quite impressive since 4DA is one of the most popular and challenging benchmarks in the literature.

6) It can be seen from Table 3 that our method outperforms the baselines notably. As [Duan *et al.*, 2012a] pointed out, video event recognition is challenging because irrelevant source domains may be harmful for the classification performances in the target domain. Most baselines perform bad because the so-called negative transfer [Pan and Yang, 2010]. Our method performs well because we only select the relevant features, and use relevant neighbors for reconstruction. The dimensionality reduction which we performed can further filter negative information. Finally, the results can also demonstrate the effectiveness of feature extraction compared with instance re-weighting.

7) Fig. 5(a) shows the parameters sensitivity of our method. It can be seen that our algorithm is robust with different values of β and γ when λ is fixed, but λ needs to be carefully chosen from [0.01, 1]. Fig. 5(b) shows that our method performs smoothly with varying dimensionality of subspace. However, computational costs will grow with the dimensionality increasing. Fig. 5(c) shows that our algorithm converges very fast, usually within about 5-round iterations.

5 Conclusion

This paper proposes a unified framework of joint feature selection and geometric structure preservation for unsupervised domain adaptation. Experiments on both visual dataset and text dataset demonstrate that the joint optimization is much better than separate ones.

Acknowledgments

This work is supported in part by the National Science Foundation of China under Grants 61273254 and 61371183.

References

[Bruzzone and Marconcini, 2010] Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE TPAMI*, 32(5):770–787, 2010.

[Cai *et al.*, 2007] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression: A unified approach for sparse subspace learning. In *ICDM 2007*, pages 73–82. IEEE, 2007.

[Chu *et al.*, 2013] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR 2013*, pages 3515–3522. IEEE, 2013.

[Chung, 1997] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119. IEEE, 2014.

[Ding *et al.*, 2015] Zhengming Ding, Ming Shao, and Yun Fu. Deep low-rank coding for transfer learning. In *AAAI 2015*, pages 3453–3459. AAAI Press, 2015.

[Donahue *et al.*, 2013] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[Duan *et al.*, 2009] Lixin Duan, Ivor W Tsang, Dong Xu, and Stephen J Maybank. Domain transfer svm for video concept detection. In *CVPR 2009*, pages 1375–1381. IEEE, 2009.

[Duan *et al.*, 2012a] Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR 2012*, pages 1338–1345. IEEE, 2012.

[Duan *et al.*, 2012b] Lixin Duan, Dong Xu, IW-H Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE TPAMI*, 34(9):1667–1680, 2012.

[Gao *et al.*, 2008] Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *ACM SIGKDD 2008*, pages 283–291. ACM, 2008.

[Gong *et al.*, 2012] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR 2012*, pages 2066–2073. IEEE, 2012.

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[Gu *et al.*, 2011] Quanquan Gu, Zhenhui Li, and Jiawei Han. Joint feature selection and subspace learning. In *IJCAI 2011*, volume 22, page 1294. Citeseer, 2011.

[Hestenes, 1969] Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

[Jiang *et al.*, 2011] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR 2011*, page 29. ACM, 2011.

[Li *et al.*, 2016] Jingjing Li, Yue Wu, Jidong Zhao, and Ke Lu. Multi-manifold sparse graph embedding for multi-modal image classification. *Neurocomputing*, 173:501–510, 2016.

[Lin *et al.*, 2010] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[Long *et al.*, 2014a] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *IEEE TKDE*, 26(7):1805–1818, 2014.

[Long *et al.*, 2014b] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *CVPR 2014*, pages 1410–1417. IEEE, 2014.

[Lu *et al.*, 2012a] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *ECCV 2012*, pages 347–360. Springer, 2012.

[Lu *et al.*, 2012b] Ke Lu, Zhengming Ding, and Sam Ge. Sparse-representation-based graph embedding for traffic sign recognition. *IEEE TITS*, 13(4):1515–1524, 2012.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *NIPS 2010*, pages 1813–1821, 2010.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE TNN*, 22(2):199–210, 2011.

[Qiu *et al.*, 2012] Qiang Qiu, Vishal M Patel, Pavan Turaga, and Rama Chellappa. Domain adaptive dictionary learning. In *ECCV 2012*, pages 631–645. Springer, 2012.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV 2010*, pages 213–226. 2010.

[Schweikert *et al.*, 2009] Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *NIPS 2009*, pages 1433–1440, 2009.

[Si *et al.*, 2010] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE TKDE*, 22(7):929–942, 2010.

[Turk and Pentland, 1991] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[Yan *et al.*, 2007] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE TPAMI*, 29(1):40–51, 2007.

[Yin *et al.*, 2015] Ming Yin, Junbin Gao, Zhouchen Lin, Qinfeng Shi, and Yi Guo. Dual graph regularized latent low-rank representation for subspace clustering. *IEEE TIP*, 24(12):4918–4933, 2015.

[Zhu and Lafferty, 2005] Xiaojin Zhu and John Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML 2005*, pages 1052–1059. ACM, 2005.