Semi-supervised Domain Adaptation with Subspace Learning for Visual Recognition

Ting Yao [†], Yingwei Pan [‡], Chong-Wah Ngo [§], Houqiang Li [‡], and Tao Mei [†]
[†] Microsoft Research, Beijing, China
[‡] University of Science and Technology of China, Hefei, China
[§] City University of Hong Kong, Kowloon, Hong Kong
{tiyao, tmei}@microsoft.com, panyw.ustc@gmail.com, cscwngo@cityu.edu.hk, lihq@ustc.edu.cn

Abstract

In many real-world applications, we are often facing the problem of cross domain learning, i.e., to borrow the labeled data or transfer the already learnt knowledge from a source domain to a target domain. However, simply applying existing source data or knowledge may even hurt the performance, especially when the data distribution in the source and target domain is quite different, or there are very few labeled data available in the target domain. This paper proposes a novel domain adaptation framework, named Semi-supervised Domain Adaptation with Subspace Learning (SDASL), which jointly explores invariant low-dimensional structures across domains to correct data distribution mismatch and leverages available unlabeled target examples to exploit the underlying intrinsic information in the target domain. Specifically, SDASL conducts the learning by simultaneously minimizing the classification error, preserving the structure within and across domains, and restricting similarity defined on unlabeled target examples. Encouraging results are reported for two challenging domain transfer tasks (including image-to-image and image-to-video transfers) on several standard datasets in the context of both image object recognition and video concept detection.

1. Introduction

In the standard machine learning technologies, the training and test data are assumed to be drawn from the same distribution. When the distribution changes, the need to rebuilt most statistical models from scratch using the newly collected training data, however, makes the task intellectually expensive or unpractical for many real-world applications. As a result, domain adaptation would be desirable.

In general, domain adaptation involves two distinct types of datasets, one from a source domain and the other from a target domain. The source domain contains a large amount of labeled data such that a classifier can be reliably built, while the target domain refers broadly to a dataset that is assumed to have different characteristics from the source. Thus, simply applying the classifier learnt in the source domain may hurt the performance in the target domain, a phenomenon known as "domain shift" [29]. Furthermore, the labeled target data are often very few and they alone are not sufficient to construct a good classifier. Therefore, our main objective is to attain good performance on the target domain by utilizing the source data or adapting classifiers trained in the source domain. In addition, how to effectively leverage unlabeled target data also remains an important issue for domain adaptation.

In the literature, there have been several techniques being proposed for addressing the challenge of domain shift by learning a common feature representation [3, 5]. The objective is to identify a new feature representation that is invariant across domains. With this, the source and the target domain exhibit more shared characteristics. However, in general, these approaches highly depend on the heuristic selection of pivot features appearing frequently in both domains. Furthermore, the criterion of feature selection may be sensitive to different applications. On the other hand, it is assumed that visual data exist in the low-dimensional subspaces, which can provide a meaningful description of the underlying domain shift [11, 12, 13, 22]. Given the data from two domains, we are investigating in this paper how to obtain the projections of mapping the data from source and target domains onto a subspace. The new feature representations in this subspace should be able to reduce the data distribution mismatch as much as possible, meanwhile preserving the structure property of the original data. Furthermore, to tackle with the challenge of target labeled data insufficiency, the unlabeled target data is also leveraged on smooth assumption encoded in a regularizer, which has been shown effective for semi-supervised learning [17].

By consolidating the idea of semi-supervised learning and subspace learning for domain adaptation, this paper presents a novel Semi-supervised Domain Adaptation with Subspace Learning (SDASL) framework for visual recognition. It attempts to learn a subspace which can manifest the underlying difference and commonness between different domains. When projected onto this subspace, the data distribution mismatch of the source and target domains can be reduced and data structure properties are preserved as well. Standard machine learning methods can then be used in the subspace to train classifier for both domains. More specifically, three regularizers are jointly employed in our framework, including the *structural risk regularizer* which seeks a decision boundary that achieves a small classification error, the *structure preservation regularizer* that restricts the distance between mappings of similar samples in both source and target domains, and the *manifold regularizer* based on the smoothness assumption that the target classifier shares similar decision values on the similar target unlabeled samples. It is worth noticing that the proposed framework is unified and any other criterion for domain adaptation can be easily incorporated. We demonstrate the effectiveness of our proposed approach on both image-to-image and image-to-video transfers, and show its superiority to several state-of-the-arts.

The remaining sections are organized as follows. Section 2 describes related work on domain adaptation. Section 3 presents our semi-supervised domain adaptation with subspace learning framework including overall objective function and its algorithm for visual recognition. Section 4 provides empirical evaluations, followed by the discussion and conclusions in Section 5.

2. Related Work

The research on domain adaptation has proceeded along three different dimensions: unsupervised domain adaptation [1, 16, 25, 26], supervised domain adaptation [2, 8, 21, 24, 28], and semi-supervised domain adaptation [9, 13, 14].

Unsupervised domain adaptation refers to the setting when the labeled target data is not available. Shi *et al.* [25] defined an information-theoretic measure which balances between maximizing domain similarity and minimizing expected classification error on the target domain. Long *et al.* [16] jointly performed feature matching and instance weighting to learn a new feature representations that is robust to domain difference. In another work by Wang *et al.* [26], the problem was considered in terms of unsupervised manifold alignment, where the source and target domains were aligned by preserving the neighborhood structure of the data points. Similar in spirit, Baktashmotlagh *et al.* [1] made use of the Riemannian metric on the statistical manifold as a measure of distance between the source and target distributions for domain adaptation.

In contrast, when the labeled target data is available, we refer to the problem as supervised domain adaptation. Yang *et al.* [28] proposed adaptive support vector machine (A-SVM) to learn a new SVM classifier for the target domain, which is adapted from an existing classifier trained with the samples from a source domain. Pan *et al.* [21] proposed a new dimensionality reduction method called maximum mean discrepancy embedding (MMDE) for domain adaptation, which aims to learn a shared latent space where distance between distributions can be reduced while the data variance can be preserved. Bergamo *et al.* [2] exploited the availability of strongly-labeled target training data to simultaneously determine the correct labels of the source training examples and incorporate this labeling information to improve the classifier by using transductive learning. Later in [8], Duan *et al.* constructed a parameterized augmented space as the common space motivated by a domain adaptation method proposed by Daumé III in [5] and the parameters were learnt through optimizing a large margin classification model. The work of Saenko *et al.* [24] was one of the earliest papers to investigate domain adaptation in visual recognition by metric learning techniques, which aim to learn a transformation that minimizes the effect of domain-induced changes.

Semi-supervised domain adaptation methods have also been proposed. Jiang *et al.* [14] proposed to not only include weighted source domain instances but also weighted unlabeled target domain instances in training, which essentially combines instance weighting with bootstrapping. Duan *et al.* [9] proposed to utilize the unlabeled target data to more precisely measure the data distribution mismatch between the source and target domains based on the maximum mean discrepancy [4]. In [13], Guo *et al.* developed a subspace co-regularized method for multilingual text classification problem. It aims to minimize the training error on the labeled data in each language while penalizing the distance between the subspace representations of the two languages of both labeled and unlabeled documents.

In short, our work in this paper belongs to semi-supervised domain adaptation. Besides of the use of unlabeled target examples as in these aforementioned semi-supervised methods, our approach additionally incorporates the objective of obtaining a subspace on which data distribution mismatch is reduced and original structure properties are preserved.

3. Semi-supervised Domain Adaptation with Subspace Learning

The main goal of semi-supervised domain adaptation with subspace learning (SDASL) is to bridge the domain gap by jointly constructing good subspace feature representations to minimize domain divergence and leveraging unlabeled target data in conjunction with labeled data. The

training of SDASL is performed simultaneously by minimizing the classification error, preserving the structure relationships within and across domains, and restricting similarity defined on unlabeled target instances. In particular, the objective function of SDASL is composed of three components, i.e., *structural risk*, *structure preservation* within and across domains, and *manifold regularization*. Of the three, the former two aim to explore invariant low dimensional structures across domains and meanwhile minimizing the structural risk of the learnt models on the subspace, while the last exploits the intrinsic information in the target domain. After we obtain the predictive function on the subspace, the label of a new coming target instance can be determined accordingly. In the following, we will first introduce the annotations used in this paper, followed by constructing the three learning components of SDASL. Then the joint overall objective and its optimization strategy are provided. Finally, the whole SDASL algorithm for visual recognition is presented.

For simplicity, we focus on the scenario when transferring only from one source. However, the proposed method can be extended to multiple sources. Suppose we are given plenty of labeled source data and only a limited number of labeled target data. Additionally we are given unlabeled target data. Our goal is to assist tasks in a label-scarce target domain by transferring the knowledge in the label-rich source domain.

3.1. Notations

Suppose there are l_s labeled samples in the source domain, represented as: $\mathbf{X}_S = \{\mathbf{x}_1^S, \mathbf{x}_2^S, \ldots, \mathbf{x}_{l_s}^S\}^\top \in \mathbb{R}^{l_s \times d_s}$, where d_s is the dimensionality of source data. Similarly, assume there are l_t ($l_t \ll l_s$) labeled instances and u_t unlabeled examples in the target domain, denoted as: $\mathbf{X}_T = \{\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_{l_t}^T\}^\top \in \mathbb{R}^{l_t \times d_t}$ and $\mathbf{X}_T^U = \{\mathbf{x}_1^U, \mathbf{x}_2^U, \ldots, \mathbf{x}_{u_t}^U\}^\top \in \mathbb{R}^{u_t \times d_t}$, respectively. The corresponding labels of $\mathbf{X}_S$ and $\mathbf{X}_T$ are given as column vectors $\mathbf{Y}_S \in \{-1, +1\}^{l_s}$ and $\mathbf{Y}_T \in \{-1, +1\}^{l_t}$, respectively.

3.2. Structural Risk

Deriving from the idea of subspace learning by assuming that the feature representations in different views are generated from this latent subspace, we project the original features into the low-dimensional subspace to explore the invariant structures across domains and minimize domain divergence. Accordingly, the linear predictive functions are defined as

$$\begin{cases} f_S(\mathbf{x}^S) = \mathbf{x}^S \mathbf{m}_S \mathbf{w}_S + b_S \\ f_T(\mathbf{x}^T) = \mathbf{x}^T \mathbf{m}_T \mathbf{w}_T + b_T \end{cases}, \tag{1}$$

where $\mathbf{w}_S, \mathbf{w_T} \in \mathbb{R}^d$ and b_S, b_T are the model weight and bias parameters, respectively. $\mathbf{m}_S$ and $\mathbf{m}_T$ are the feature mapping matrices, with $\mathbf{m}_S \in \mathbb{R}^{d_S \times d}$ and $\mathbf{m}_T \in \mathbb{R}^{d_T \times d}$, where d is the dimension of the subspace. The mapping matrices $\mathbf{m}_S$ and $\mathbf{m}_T$ are designed to be orthogonal in order to make each mapping basis uncorrelated to each other, i.e., $\mathbf{m}_S^\top \mathbf{m}_S = \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

Furthermore, the training objective corresponds to an empirical risk minimization with a regularization penalty over the model parameters $\{\mathbf{w}_S, b_S, \mathbf{m}_S, \mathbf{w}_T, b_T, \mathbf{m}_T\}$ as

$$\begin{aligned} \min_{\substack{\mathbf{w}_S, b_S, \mathbf{m}_S \\ \mathbf{w}_T, b_T, \mathbf{m}_T}} & \|\mathbf{X}_S \mathbf{m}_S \mathbf{w}_S + b_S - \mathbf{Y}_S\|^2 + \alpha_S \|\mathbf{w}_S\|^2 \\ & + \|\mathbf{X}_T \mathbf{m}_T \mathbf{w}_T + b_T - \mathbf{Y}_T\|^2 + \alpha_T \|\mathbf{w}_T\|^2 \\ & s.t. \quad \mathbf{m}_S^\top \mathbf{m}_S = \mathbf{I}, \quad \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I} \end{aligned}, \tag{2}$$

where α_S and α_T are tradeoff parameters. The objective decomposes into the empirical risk with a least square loss of the labeled examples from both source and target domains, and the regularization penalty $\|\mathbf{w}_S\|^2$ and $\|\mathbf{w}_T\|^2$. The parameter α_S and α_T are the tradeoff parameters.

3.3. Structure Preservation

One of the key goals in most state-of-the-art multi-view learning [10] is to seek for a joint latent space that corresponding views are mapped to nearby locations. This also indicates that similar views should have similar mappings. Similarly, to tackle with the challenge of domain shift, we incorporate a discriminative regularization term in the objective function to take into account of the structure within and across domains. That is, the distance between the mappings in the latent subspace of the same category from source and target domains should be as small as possible.

Technically, positives from both domains are represented as: $\mathbf{A} = \begin{bmatrix} \mathbf{X}_S^+ \mathbf{m}_S \\ \mathbf{X}_T^+ \mathbf{m}_T \end{bmatrix}$, where $\mathbf{X}_S^+$ and $\mathbf{X}_T^+$ denote the positives in the source and target domain, respectively. The distance between positives from source and target domains is measured by $tr(\mathbf{A}^\top \mathbf{L}_1 \mathbf{A})$, where $\mathbf{L}_1 = \mathbf{D}_1 - \mathbf{1}\mathbf{1}^\top$, $\mathbf{1}$ denotes a column vector with all 1 entries, and $\mathbf{D}_1$ is the diagonal matrix that contains the row sums of $\mathbf{1}\mathbf{1}^\top$.

To learn a shared latent space across different domains, we integrate the structure preservation within and across domains as a regularization for domain adaptation.

3.4. Manifold Regularization

Manifold regularization has been shown effective for semi-supervised learning [17]. This regularizer is to measure the smoothness of the predicted class labels along the inherent structure of unlabeled target data. In other words, the outputs of the predictive function are restricted to have similar values for similar examples.

The estimation of the manifold regularization can be measured by the appropriate pairwise similarity between

the unlabeled target samples. Specifically, it can be given by

$$\sum_{i,j=1}^{u_t} \mathbf{S}_{ij} \left\| \mathbf{x}_i^U \mathbf{m}_T \mathbf{w}_T - \mathbf{x}_j^U \mathbf{m}_T \mathbf{w}_T \right\|^2, \tag{3}$$

where $\mathbf{S} \in \mathbb{R}^{u_t \times u_t}$ denotes the affinity matrix defined on the unlabeled target samples. Under the manifold criterion, it is reasonable to minimize Eq.(3), because it will incur a heavy penalty if the difference between the outputs of function $f_T(\mathbf{x}^T)$ for similar examples is big.

There are many ways of defining the affinity matrices $\mathbf{S}$. Inspired by [10], the elements are computed by Gaussian functions in this work, i.e.,

$$\mathbf{S}_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i^U - \mathbf{x}_j^U\|^2}{\sigma^2}} & if\ \mathbf{x}_i^U \in N_k(\mathbf{x}_j^U)\ or\ \mathbf{x}_j^U \in N_k(\mathbf{x}_i^U) \\ 0 & otherwise \end{cases}, \tag{4}$$

where σ is the bandwidth parameter, and $N_k(\mathbf{x}_i^U)$ represents the set of k nearest neighbors of $\mathbf{x}_i^U$.

By defining the graph Laplacian $\mathbf{L}_2 = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix with its elements defined as $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$, the regularization can be computed as $(\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)^\top \mathbf{L}_2 (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)$.

This regularization term can be added to our optimization framework, which can utilize unlabeled target examples that have auxiliary similarity information. It can also be considered as a generalization of the semi-supervised version of [17] to the domain adaptation.

3.5. Overall Objective Function

The overall objective function integrates the optimization objectives throughout subsections (3.2-3.4). Hence we get the following optimization problem

$$\begin{aligned} \min_{\left\{\begin{smallmatrix}\mathbf{w}_S, b_S, \mathbf{m}_S \\ \mathbf{w}_T, b_T, \mathbf{m}_T\end{smallmatrix}\right\}} & \|\mathbf{X}_S \mathbf{m}_S \mathbf{w}_S + b_S - \mathbf{Y}_S\|^2 + \alpha_S \|\mathbf{w}_S\|^2 \\ & + \|\mathbf{X}_T \mathbf{m}_T \mathbf{w}_T + b_T - \mathbf{Y}_T\|^2 + \alpha_T \|\mathbf{w}_T\|^2 \\ & + \gamma tr(\mathbf{A}^\top \mathbf{L}_1 \mathbf{A}) + \eta (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T)^\top \mathbf{L}_2 (\mathbf{X}_T^U \mathbf{m}_T \mathbf{w}_T) \\ & s.t. \quad \mathbf{m}_S^\top \mathbf{m}_S = \mathbf{I}, \quad \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I} \end{aligned}, \tag{5}$$

where γ and η are tradeoff parameters.

Next we show that the optimal $\{\mathbf{w}_S, \mathbf{w}_T, b_S, b_T\}$ can be solved in terms of $\mathbf{m}_S$ and $\mathbf{m}_T$. We minimize the objective function in Eq.(5) by setting its derivative with respect to $\mathbf{w}_S$, $\mathbf{w}_T$, b_S and b_T to zero, which results in:

$$\begin{aligned} & \left\{ b_V = \frac{1}{l_V} \mathbf{1}^\top (\mathbf{Y}_V - \mathbf{X}_V \mathbf{m}_V \mathbf{w}_V) \right\}_{V \in \{S,T\}} \\ & \mathbf{w}_S = \left[(\mathbf{X}_S \mathbf{m}_S)^\top \mathbf{H}_S \mathbf{X}_S \mathbf{m}_S + \alpha_S \mathbf{I} \right]^{-1} \mathbf{m}_S^\top \mathbf{Z}_S \\ & \mathbf{w}_T = \left[(\mathbf{X}_T \mathbf{m}_T)^\top \mathbf{H}_T \mathbf{X}_T \mathbf{m}_T + \alpha_T \mathbf{I} + \eta \mathbf{C} \right]^{-1} \mathbf{m}_T^\top \mathbf{Z}_T, \end{aligned} \tag{6}$$

where $\mathbf{Z}_V$, $\mathbf{H}_V$, and $\mathbf{C}$ are defined as

$$\begin{aligned} & \left\{ \mathbf{Z}_V = \mathbf{X}_V^h \mathbf{H}_V \mathbf{Y}_V, \quad \mathbf{H}_V = \mathbf{I} - \frac{1}{l_V} \mathbf{1}\mathbf{1}^\top \right\}_{V \in \{S,T\}} \\ & \mathbf{C} = (\mathbf{X}_T^U \mathbf{m}_T)^\top \mathbf{L}_2 \mathbf{X}_T^U \mathbf{m}_T \end{aligned}. \tag{7}$$

Note that we use $V \in \{S, T\}$ for simplicity, i.e., V can be replaced by any symbol of S and T.

Substituting Eq.(6) into Eq.(5), we get the objective function:

$$\begin{aligned} L(\mathbf{m}_S, \mathbf{m}_T) = & \gamma tr((\tilde{\mathbf{X}}_S \mathbf{m}_S + \tilde{\mathbf{X}}_T \mathbf{m}_T)^\top \mathbf{L}_1 (\tilde{\mathbf{X}}_S \mathbf{m}_S + \tilde{\mathbf{X}}_T \mathbf{m}_T)) \\ & + \mathbf{Y}_S^\top \mathbf{H}_S \mathbf{Y}_S - \mathbf{Z}_S^\top \mathbf{m}_S (\mathbf{m}_S^\top \bar{\mathbf{M}}_S \mathbf{m}_S + \alpha_S \mathbf{I})^{-1} \mathbf{m}_S^\top \mathbf{Z}_S \\ & + \mathbf{Y}_T^\top \mathbf{H}_T \mathbf{Y}_T - \mathbf{Z}_T^\top \mathbf{m}_T (\mathbf{m}_T^\top \bar{\mathbf{M}}_T \mathbf{m}_T + \alpha_T \mathbf{I})^{-1} \mathbf{m}_T^\top \mathbf{Z}_T \end{aligned}, \tag{8}$$

where $\tilde{\mathbf{X}}_S = [\mathbf{X}_S^+, \mathbf{0}]^\top$ and $\tilde{\mathbf{X}}_T = [\mathbf{0}, \mathbf{X}_T^+]^\top$. $\bar{\mathbf{M}}_S$ and $\bar{\mathbf{M}}_T$ are defined as

$$\bar{\mathbf{M}}_S = \mathbf{X}_S^\top \mathbf{H}_S \mathbf{X}_S \ \text{ and } \ \bar{\mathbf{M}}_T = \mathbf{X}_T^\top \mathbf{H}_T \mathbf{X}_T + \eta \mathbf{X}_T^{U\top} \mathbf{L}_2 \mathbf{X}_T^U. \tag{9}$$

From the above, the overall objective function can be rewritten as

$$\min_{\{\mathbf{m}_S, \mathbf{m}_T\}} L(\mathbf{m}_S, \mathbf{m}_T) \quad s.t.\ \mathbf{m}_S^\top \mathbf{m}_S = \mathbf{I},\ \mathbf{m}_T^\top \mathbf{m}_T = \mathbf{I}. \tag{10}$$

The optimization above is a non-convex problem. However, the gradient of the objective function with respect to $\mathbf{m}_S$ and $\mathbf{m}_T$ can be easily obtained and we have

$$\begin{aligned} \nabla_{\mathbf{m}_V} L(\mathbf{m}_S, \mathbf{m}_T) = & -2\mathbf{Z}_V \mathbf{Z}_V^\top \mathbf{m}_V (\mathbf{m}_V^\top \bar{\mathbf{M}}_V \mathbf{m}_V + \alpha_V \mathbf{I})^{-1} \\ & + 2\bar{\mathbf{M}}_V \mathbf{m}_V (\mathbf{m}_V^\top \bar{\mathbf{M}}_V \mathbf{m}_V + \alpha_V \mathbf{I})^{-1} \mathbf{m}_V^\top \mathbf{Z}_V \mathbf{Z}_V^\top \mathbf{m}_V \\ & (\mathbf{m}_V^\top \bar{\mathbf{M}}_V \mathbf{m}_V + \alpha_V \mathbf{I})^{-1} + 2\gamma \tilde{\mathbf{X}}_V^\top \mathbf{L}_1 (\tilde{\mathbf{X}}_S \mathbf{m}_S + \tilde{\mathbf{X}}_T \mathbf{m}_T), \end{aligned} \tag{11}$$

for $V \in \{S, T\}$.

3.6. SDASL Algorithm

To address the difficult non-convex problem (10) due to the orthogonal constrains, we use a gradient descent optimization procedure with curvilinear search for a local optimal solution and readers can refer to [27] for details.

After the optimization of $\mathbf{m}_S$ and $\mathbf{m}_T$, we can obtain the linear predictive functions defined in Eq.(1) with the model parameters $\{\mathbf{w}_V, b_V\}_{V \in \{S,T\}}$ calculated by Eq.(6). Next, given a target test visual instance, $\hat{\mathbf{x}} \in \mathbb{R}^{d_t}$, we compute the prediction values using the linear function as

$$f(\hat{\mathbf{x}}) = \hat{\mathbf{x}} \mathbf{m}_T \mathbf{w}_T + b_T. \tag{12}$$

The label of instance $\hat{\mathbf{x}}$ is $\text{sign}(f(\hat{\mathbf{x}}))$, where $\text{sign}(\bullet)$ is the signum function. The whole SDASL algorithm is given as Algorithm 1.

Algorithm 1 Semi-supervised Domain Adaptation with Subspace Learning (SDASL)

Input: $0 < \mu < 1, \varepsilon \geq 0$.
Initialize the mapping matrices $\mathbf{m}_S$ and $\mathbf{m}_T$ using Principal Component Analysis (PCA).
for $iter = 1$ to T_{max} **do**
  Compute gradients:
    $\mathbf{G}_S = \nabla_{\mathbf{m}_S} L(\mathbf{m}_S, \mathbf{m}_T)$
    $\mathbf{G}_T = \nabla_{\mathbf{m}_T} L(\mathbf{m}_S, \mathbf{m}_T)$
  if $\|\mathbf{G}_S\|_F^2 + \|\mathbf{G}_T\|_F^2 \leq \varepsilon$ **then**
    Exit.
  end if
  Compute skew-symmetric matrices:
    $\mathbf{P}_S = \mathbf{G}_S \mathbf{m}_S^\top - \mathbf{m}_S \mathbf{G}_S^\top$
    $\mathbf{P}_T = \mathbf{G}_T \mathbf{m}_T^\top - \mathbf{m}_T \mathbf{G}_T^\top$
  Set $\tau = 1$
  repeat
    $\tau = \mu\tau$
    Compute new trial point:
      $\mathbf{Q}_S(\tau) = (\mathbf{I} + \frac{\tau}{2}\mathbf{P}_S)^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{P}_S)\mathbf{m}_S$
      $\mathbf{Q}_T(\tau) = (\mathbf{I} + \frac{\tau}{2}\mathbf{P}_T)^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{P}_T)\mathbf{m}_T$
  until Armijo-Wolfe conditions [19] meet
  Update the transformation matrices:
    $\mathbf{m}_S = \mathbf{Q}_S(\tau)$
    $\mathbf{m}_T = \mathbf{Q}_T(\tau)$
end for
Compute $\mathbf{w}_S$, $\mathbf{w}_T$, b_S and b_T via Eq.(6).
Output:
Predictive function: $\forall \hat{\mathbf{x}}, \quad f(\hat{\mathbf{x}}) = \hat{\mathbf{x}}\mathbf{m}_T\mathbf{w}_T + b_T$.

4. Experiments

We conducted our experiments for both image-to-image and image-to-video transfer tasks, i.e., object recognition on the image dataset studied in [24], and video concept detection on the challenge TRECVID 2011 Semantic Indexing (SIN) task with the assistance of images from ImageNet [6].

4.1. Image-to-image transfer

The first experiment was conducted on the Office dataset released in [24]. It contains three image datasets from three different domains. The images in the first domain dslr are captured with a digital SLR camera and have high resolution. The second domain amazon consists of images downloaded from online merchants (www.amazon.com). These images are of products at medium resolution. The images in the third domain webcam are collected by a web camera. Thus, the images are of low resolution. Each domain contains 31 categories and in total there are 4,652 images in all the three domains. Figure 1 shows image examples of category "bike" and "desk chair" from the three domains and illustrates the difference or shift between domains.

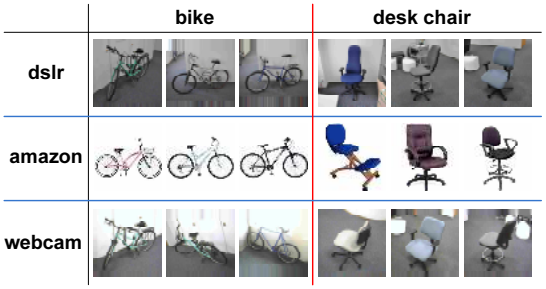

Figure 1. Image examples of category "bike" and "desk chair" come from (top row) dslr (high-resolution images captured with a digital SLR camera), (middle row) amazon (medium-resolution images downloaded from online merchants), and (bottom row) webcam (low-resolution images recorded by a web camera).

Compared Approaches. We compare the following approaches for performance evaluation:

- *SVM-S*. A SVM classifier trained only on the labeled examples in the source domain.
- *SVM-T*. A SVM classifier learnt entirely from the labeled examples in the target domain.
- *SVM-ST*. An aggregate SVM classifier trained from all the labeled samples in both source and target domains.
- *A-SVM* [28] aims to learn a new decision boundary that is close to the original one (learnt on the source labeled data) as well as separating the target data.
- *FR* [5] is to augment features for transfer learning. The augmented features are used to construct a kernel function for SVM training. With this, the impact of the examples from target domain is twice as those from source domain on the predictions of target test data.
- *Metric* [24] is to learn a transformation based on the information theoretic metric learning method of [15] by leveraging both similarity constrains within the same categories and dissimilar constrains between the different categories.
- *HFA* [8] learns the classifier and transformations to a common latent subspace between source and target in a max-margin framework.
- *GFK* [11] integrates an infinite number of subspaces along the geodesic flow between the source and target domains to characterize changes in between.
- *SCMV* [13] assumes the representations of the same object from different domains in the subspace should be similar. The learning of the subspace and classifier

is conducted by simultaneously minimizing the training losses on the labeled data in both domains and penalizing the distance between the two projected subspace representations of the same object.

- *SDASL* is our approach described in Algorithm 1.

Parameter Setting. In this experiment, we focus on one source to one target domain adaptation on image object detection task. In each setting of our experiments, we pick one of the three domain as the source domain and another one as the target domain. Then six source-target domain pairs are generated by the aforementioned three domains, i.e., *a2w*, *w2a*, *a2d*, *d2a*, *w2d* and *d2w*, where *a*, *w* and *d* represents amazon, webcam and dslr, respectively. All the examples in the source domain are used as the source training data. The instances in the target domain are evenly split into two halves. One is used as the target training data and the other is as the target test collection. Furthermore, to simulate a semi-supervised learning scenario, we divide the selected training data into two subsets: one subset is used as the labeled set (five in our experiments) in which we consider that the labels are known; for the remaining training examples, the labels are hidden and this subset is used as unlabeled set. We take the output of 1000-way fc8 classification layer by using DeCAF [7] as image representation.

To ensure the performance of these methods comparable, we use the same RBF kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\delta\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ with $\delta = 0.1$ in all SVM-based methods. Following the setting in [24, 11], *Metric* and *GFK* use 1-nearest neighbor as its classifier. For the proposed *SDASL*, we empirically set μ=0.3 in the curvilinear search. The parameters α_S and α_T are both fixed to 1.0. The tradeoff parameters γ and η are selected from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and the optimal values are determined by using a validation set. The averaged accuracies over all categories on target domain are finally reported.

Performance Comparison. Table 1 summarizes the classification accuracy obtained by all the above methods averaged over all 31 categories for six pairings of the source and target domains. The highest performances are in bold font and the symbol $\uparrow$ indicates that performance is significantly better than others, according to the randomization test [23] with 100,000 randomization iterations and at 0.05 significance level. It is also worth noting that the performances are given by choosing 100 as the dimensionality of the latent subspace for methods *SCMV* and *SDASL*. According to [8] and [11], the dimensions are set to 1,000 and 10 for *HFA* and *GFK*, respectively. The other six methods perform on the original 1,000 dimensional visual features.

Overall, our proposed *SDASL* consistently outperforms the other runs across different pairings of the source and target domains. In particular, the accuracy of *SDASL* for the adaptation from amazon to webcam can achieve 0.8540, which makes the improvement over *SVM-T* by 5.5%. More importantly, by learning a low-dimensional latent subspace, the dimension of the mappings of visual feature is reduced by several orders of magnitude. Furthermore, *SDASL* by additionally incorporating manifold regularization leads to a performance boost against *SCMV* which only restricts the distance between the two projected subspace representations. *SDASL* improves *Metric* and *HFA*, which basically indicates the advantage of exploiting the unlabeled target examples. Our *SDASL* also exhibits better performance than *GFK*, where the very low-dimensional subspaces may not represent high-dimensional input data accurately.

There is a significant performance gap between *SVM-T* and *SVM-S*, which demonstrates that the SVM classifier learnt with the source examples performs much worse than developing a new classifier with very few target training examples due to the domain gap. The exceptions are the two transfers between webcam and dslr, in which *SVM-S* slightly improves *SVM-T*. This is arisen from the fact that the datasets webcam and dslr are statistically similar as the same objects are captured with different camera for each dataset. Therefore, the knowledge transfer in between is more confident and closely related. Similar in spirit, take webcam as the target domain, *SVM-S* model learnt in dslr exhibits better performance than that learnt in amazon.

Another interesting observation is that *SVM-T* outperforms *SVM-ST* when amazon (with much more images) is as source domain. The result basically indicates that the risk of bias on source domain could be increased by incorporating sufficient source positive examples, and thus may hurt the performance on target domain. In contrast, the improvement of *FR* is more obvious on all six transfer pairs, which verifies the advantage of augmenting the influence of target examples. Instead of explicitly being affected by the data distribution difference, *A-SVM* is to adapt a source model so that the decision boundary is adjusted to fit the target domain. With this, the adaptation can benefit from source domain and lead to performance gain.

In addition, we conducted experiments by using the outputs of fc6 and fc7 layers in DeCAF [7] as image representations on all the compared methods, respectively. The performance trends are similar with that of fc8 and our method outperforms all the baselines on *a2w*, *w2a*, *a2d* and *d2a* adaptations. On the other *w2d* and *d2w* settings, *SVM-ST* and *SVM-S* achieve the best two performances, which is similar to the observations in [7], followed by our method. This is still due to the fact that in webcam and dslr, the major difference between the images of same objects is caused by the use of different camera devices. Thus, the problem of domain shift is not severe, and can be handled by fc6 and fc7 representations.

Effect of Each Regularizer. As three components, i.e., structural risk on the subspace (SR) in Section 3.2, structure

Table 1. Classification accuracy of different approaches averaged over all 31 categories. The highest performances are in bold font and the numbers in the brackets represent the feature dimension used in each approach. The symbol ↑ indicates statistically better performance than the others according to the randomization test [23]. (a: amazon, w: webcam, and d: dslr)

	SVM-T (d=1000)	*SVM-S* (d=1000)	*SVM-ST* (d=1000)	*A-SVM*[28] (d=1000)	*FR*[5] (d=1000)	*Metric*[24] (d=1000)	*HFA*[8] (d=1000)	*GFK*[11] (d=10)	*SCMV*[13] (d=100)	*SDASL* (d=100)
a2w	0.8094	0.5198	0.7822	0.8096	0.8193	0.7995	0.7846	0.8323	0.8168	**0.8540** ↑
w2a	0.6393	0.4859	0.6506	0.6513	0.6506	0.6591	0.6570	0.6581	0.6605	**0.6726** ↑
a2d	0.8293	0.5488	0.7764	0.7886	0.8374	0.8333	0.8374	0.8258	0.8415	**0.8577** ↑
d2a	0.6393	0.4661	0.6443	0.6485	0.6478	0.6499	0.6556	0.6645	0.6457	**0.6676** ↑
w2d	0.8293	0.8415	0.8455	0.8333	0.8415	0.8496	0.8536	0.8323	0.8455	**0.8618** ↑
d2w	0.8094	0.8218	0.8292	0.8267	0.8317	0.8391	0.7921	0.8387	0.8342	**0.8465** ↑

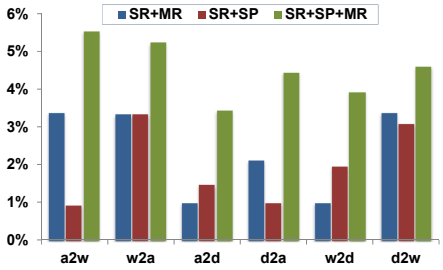

Figure 2. Performance improvements over *SVM-T* with different combination of regularization across six pairing of the source and target domain. The performances are compared against the results w.r.t their overall classification accuracy.

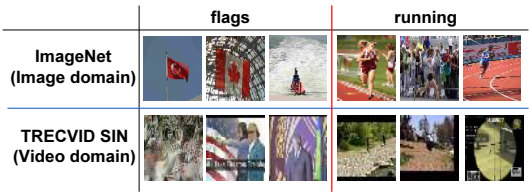

Figure 3. Examples of concept "flags" and "running" come from (top row) ImageNet (image domain) and (bottom row) TRECVID Semantic Indexing (SIN) (video domain) illustrating the apparent difference between image and video domains.

preservation (SP) in Section 3.3, and manifold regularization (MR) in Section 3.4, are jointly explored and optimized in our proposed *SDASL*. The degree of contribution from each component is here investigated. Figure 2 shows the degree of improvement over the run *SVM-T* with three different combinations of regularization, i.e., SR+MR, SR+SP, and SR+SP+MR. The optimization procedure on each combination all uses curvilinear search presented in [27]. The results across six pairings of the source and target domains consistently indicate that learning using three components leads to a larger performance boost compared to using two components. Furthermore, learning utilizing SR+MR also exhibits better performance than SR+SP when amazon and webcam are picked as target domain, in contrast, SR+SP outperforms SR+MR when dslr is used as target domain. This observation is not surprise because dataset dslr has much less training examples than amazon and webcam weakening the effect of unlabeled target data in MR.

4.2. Image-to-video transfer

The second experiment was conducted on ImageNet [6] (Web images with clean labels) and TRECVID 2011 SIN dataset (TV11) [20] (Web video). We use TV11 as the target domain, while examples from ImageNet as the source domain training data. Among the 50 concepts officially evaluated in TV11, there are 22 concepts which share common definition with ImageNet and these concepts are evaluated in our experiments. Figure 3 shows examples of concepts "flags" and "running" from the two domains. As illustrated in the figure, the image and video domains are quite different making the transferring from image-to-video much harder than image-to-image.

Parameter Setting. The source training data consists of all the images of these 22 concepts collected from ImageNet. In total, there are 27,452 source training images. TV11 has 46,133 training shots for all the 22 concepts and we randomly select 3, 5, 7, 10, 20, 50, and 100 positive samples per concept from the training set as the labeled target training data. The remaining training examples are used as the unlabeled set. In addition, TV11 also contains 137,327 video shots as testing samples and they are all used as target test data for each concept.

We use bag-of-visual-words (BoW) to represent the diverse content of images/keyframes, because of their consistent good performances reported in [18]. BoW is generated from SIFT of local interest points extracted by Difference-of-Gaussian (DoG) and Hessian Affine detectors. Specifically, we generate a visual vocabulary of 500 words for each kind of keypoints using k-means and then encode each image/keyframe with 1000-dimensional vector by concatenating the BoW histograms from the two vocabularies. Following TRECVID evaluation, the Inferred average precision (InfAP) which is an approximation of Average Precision (AP) on partially labeled testing data is computed over the top 2,000 retrieved shots. Note that we do not compare with [8, 11, 24] here because [8] is with high computational

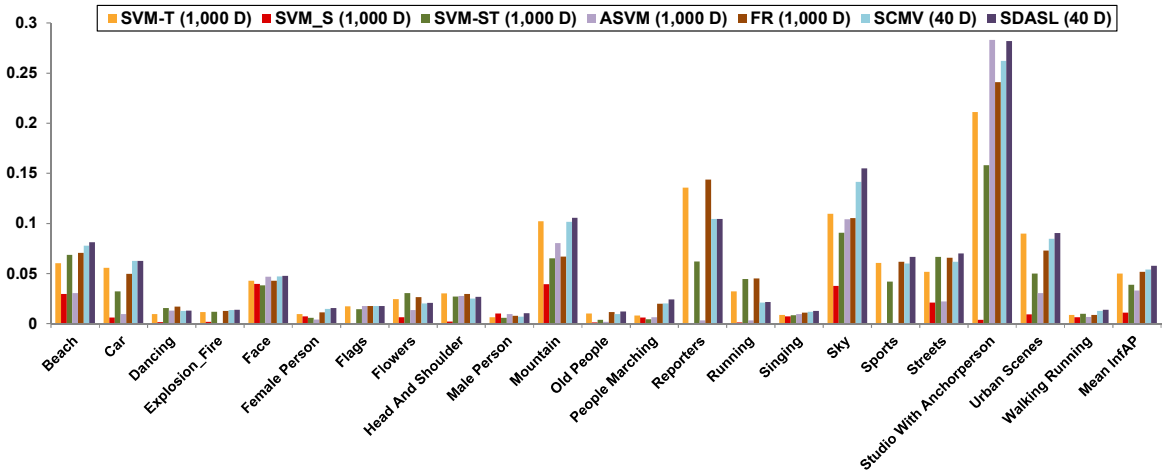

Figure 4. Per-concept InfAPs of different approaches with 100 target positive examples for all the 22 concepts. The numbers in the brackets represent the feature dimension used in each approach.

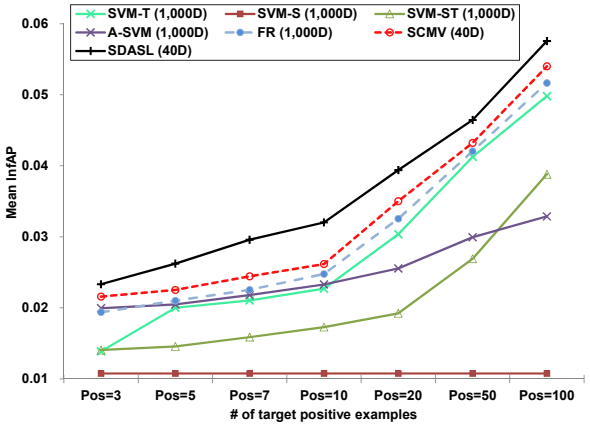

Figure 5. The performance (Mean InfAP) of different approaches with the increase of target positive training examples. The numbers in the brackets represent the feature dimension.

cost when training on thousands of examples for each concept and [11, 24] cannot cope with this evaluation criteria.

Performance Comparison. Figure 5 shows the performance of different approaches in terms of Mean InfAP against the number of target positive training examples. Overall, *SDASL* with different target positive examples consistently exhibits significantly better performance than other approaches. The three adaptation methods as *SDASL*, *SCMV*, and *FR* perform better than the two baseline methods as *SVM-S* and *SVM-T*. The result basically indicates the benefit of re-using source data. Furthermore, when the number of the target positive training examples exceeds 10, *SVM-T* outperforms *A-SVM* instead. This somewhat reveals the weakness of *A-SVM*, which restricts the target decision boundary not far away from the source one. In practice, this constrain may even deteriorate the adaptation performance, especially when the domain gap is large.

Figure 4 further details the InfAP performance for all the 22 queries. Basically different approach respond differently to concepts. For instance, concept "Male Person" is better classified with *SVM-S*. On the other hand, the concept "Studio with Anchorperson" shows much better result with *SVM-T*. In the experiment, *SDASL* successfully brings up the InfAP performance of these concepts. Among all the 22 concepts, *SDASL* achieves the best performance for 16 concepts. To verify that the performance of different approaches is not by chance, we conducted significance test using the randomization test [23]. The number of iterations used in the randomization is 100,000 and at 0.05 significance level. *SDASL* is found to be significantly better than others.

5. Discussion and Conclusion

In this paper, we have presented semi-supervised domain adaptation with subspace learning for visual recognition. Particularly, we explore a new feature representation in the subspace which could reduce the data distribution mismatch across domains and preserve structure properties of the original data. Meanwhile, as the unlabeled target examples exhibit the underlying intrinsic information in the target domain, these examples are further employed to generalize the visual concept classifier. Experiments conducted on both image-to-image and image-to-video transfers validate our proposal and analysis. Our future works are as follows. First, we will extend this work to multiple source domains, where the subspace is explored holistically among all the domains. Moreover, several issues arisen from this extension need to be further addressed, e.g., how to select good source domains making the task more effective. Second, as the proposed framework is unified and any other criterion can be easily incorporated, we will investigate other robust principles to further improve the performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61272290, No. 61325009), the National Hi-Tech Research and Development Program (863 Program) of China under Grant 2014AA015102, and the 973 Program under Grant No. 2015CB351803.

References

[1] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Domain adaptation on the statistical manifold. In *CVPR*, 2014.

[2] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010.

[3] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[4] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22:49–57, 2006.

[5] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. 2013.

[8] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, 2012.

[9] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[10] Z. Fang and Z. Zhang. Discriminative feature selection for multi-view cross-domain learning. In *CIKM*, 2013.

[11] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

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

[13] Y. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *ICML*, 2012.

[14] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.

[15] B. Kulis, P. Jain, and K. Grauman. Fast similarity search for learned metrics. *IEEE Trans. on PAMI*, 39(12):2143–2157, 2009.

[16] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer joint matching for unsupervised domain adaptation. In *CVPR*, 2014.

[17] S. Melacci and M. Belkin. Laplacian support vector machines trained in the primal. *Journal of Machine Learning Research*, 12:1149–1184, 2011.

[18] C.-W. Ngo, Y.-G. Jiang, X. Y. Wei, W. L. Zhao, Y. Liu, S. A. Zhu, and S.-F. Chang. Vireo/dvmm at trecvid 2009: High-level feature extraction, automatic video search, and content-based copy detection. In *NIST TRECVID workshop*, 2009.

[19] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.

[20] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, et al. Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *NIST TRECVID workshop*, 2011.

[21] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, 2008.

[22] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. on Neural Networks*, 99:1–12, 2009.

[23] J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990.

[24] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[25] Y. Shi and F. Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *ICML*, 2012.

[26] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI*, 2009.

[27] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constrains. *Mathematical Programming*, 142:397–434, 2013.

[28] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM MM*, 2007.

[29] T. Yao, C.-W. Ngo, and S. Zhu. Predicting domain adaptivity: Redo or recycle? In *ACM MM*, 2012.