Robust Visual Domain Adaptation with Low-Rank Reconstruction

I-Hong Jhuo[†‡*], Dong Liu[§], D. T. Lee[†‡], Shih-Fu Chang[§]
[†]Dept. of CSIE, National Taiwan University, Taipei, Taiwan
[‡]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[§]Dept. of Electrical Engineering, Columbia University, New York, NY, USA
ihjhuo@gmail.com, {dongliu, sfchang}@ee.columbia.edu, dtlee@ieee.org

Abstract

Visual domain adaptation addresses the problem of adapting the sample distribution of the source domain to the target domain, where the recognition task is intended but the data distributions are different. In this paper, we present a low-rank reconstruction method to reduce the domain distribution disparity. Specifically, we transform the visual samples in the source domain into an intermediate representation such that each transformed source sample can be linearly reconstructed by the samples of the target domain. Unlike the existing work, our method captures the intrinsic relatedness of the source samples during the adaptation process while uncovering the noises and outliers in the source domain that cannot be adapted, making it more robust than previous methods. We formulate our problem as a constrained nuclear norm and $\ell_{2,1}$ norm minimization objective and then adopt the Augmented Lagrange Multiplier (ALM) method for the optimization. Extensive experiments on various visual adaptation tasks show that the proposed method consistently and significantly beats the state-of-the-art domain adaptation methods.

1. Introduction

Visual classification is often faced with the dilemma of data deluge and the label scarcity. While exploiting the vast amount of unlabeled data directly (e.g., via the semi-supervised learning paradigm [27]) is valuable in its own right, it is beneficial to leverage labeled data samples of relevant categories across data sources. For example, it is increasingly popular to enrich our limited collection of training data samples with those from the Internet. One problem with this strategy, however, comes from the possible misalignment of the *target* domain under consideration and the *source* domain that provides the extra data and labels. Physically, this misalignment results from bias of each visual domain in terms of a variety of visual cues, such as the visual resolution, viewpoint, illumination, and so on. Figure 1 shows some example discrepancy of visual appearance of *bookcase* images among three domains.

Figure 1. *Bookcase* images of different domains from the domain adaptation benchmark dataset [21]. The images in the first column are from the *amazon* domain, while the images in the second and third columns are from the *dslr* and *webcam* domain, respectively. As can be seen, the visual appearance of the images from different domains vary a lot.

This misalignment corresponds to the shift in data distribution in a certain feature space. To be precise, the marginal distribution of the samples in the source domain and that in the target are different. This makes direct incorporation of data from the source domain harmful: in theory, the disparity violates the basic assumption underpinning supervised learning; in practice, the resulting performance degrades considerably on the target test samples [21], refuting the value of introducing the auxiliary data.

The above theoretic and practical paradox has inspired recent research efforts into the *domain adaptation* problem in computer vision and machine learning [9, 12, 13, 14, 15]. Formally, domain adaptation addresses the problem where the marginal distribution of the samples X_s in the source domain and the samples X_t in the target domain are different, while the conditional distributions of labels provided samples, $P(Y_s|X_s)$ and $P(Y_t|X_t)$ (Y_s and Y_t denoting la-

*The work is supported by NSC Study Abroad Program grants 100-2917-I-002-043.

bels in either domain) are similar [13]. The goal is to effectively adapt the sample distribution of the source domain to the target domain.

Existing solutions to this problem vary in setting and methodology. Depending on how the source information is exploited, the division is between classifier-based and representation-based adaptation. The former advocates implicit adaptation to the target distribution by adjusting a classifier from the source domain (e.g., [1, 11, 12, 14]), whereas the latter attempts to achieve alignment by adjusting the representation of the source data via learning a transformation [15, 21]. Orthogonal to this, the extant proposals can also be classified into supervised (e.g., [11, 12, 14, 15, 21]) and unsupervised (e.g., [13]) adaptation, based on whether labels have been exploited during the adaptation.

The common issues with the prior proposals are twofold. First, during the adaptation, they typically deal with source samples separately without accounting for the mutual dependency. This may (either implicitly or explicitly) cause the adapted distribution to be arbitrarily scattered around and any structural information beyond single data samples of the source data may become undermined. Second, they blindly translate all samples including the noises and particularly possible outliers from the source domain to the target. The latter can lead to significantly distorted or corrupted models when the recognition models are learned.

In this paper, we propose a novel visual domain adaptation method which not only tries to keep the intrinsic relatedness of source samples during adaptation but also achieve a more robust adaptation by accounting for noises and removing outliers. As illustrated in Figure 2, the basic idea is to transform the data samples in the source domain into an intermediate representation such that each transformed sample can be linearly reconstructed by samples of the target domain. Upon this linear relationship, we capture the intrinsic relatedness of the source samples using a low-rank structure and meanwhile identify the outlying samples using a sparse structure. The whole transformation procedure is unsupervised without utilizing any label information. We then formulate our proposal into a constrained nuclear norm and $\ell_{2,1}$-norm minimization problem and adopt the Augmented Lagrange Multiplier (ALM) [16] method for optimization. Extensive experimental results on various visual recognition tasks very well verify the effectiveness of our method. In addition, we extend our method to the scenario considering multiple related source domains, and propose a multi-task low-rank domain adaptation method, which can simultaneously adapt multiple source domains into the target domain via low-rank reconstruction.

2. Related Work

The domain adaptation problem has recently been extensively studied in the literature [7, 21, 22, 8]. Notably,

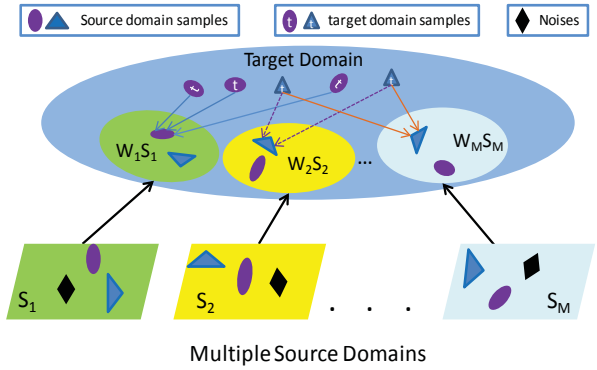

Figure 2. Illustration of our proposed method. Each source domain S_i contains two classes of samples (marked as purple ellipses and blue triangles) as well as some noisy samples (marked as black diamonds). The samples in the target domain are marked with letter 't'. Our method transforms each source domain S_i into an intermediate representation W_iS_i such that each transformed sample can be linearly reconstructed by the target samples. Within each source domain S_i, we enforce the reconstruction of source samples to be related to each other under a low-rank structure while allowing the existence of a sparse set of noisy samples. Furthermore, by enforcing different source domains $W_1S_1, ..., W_MS_M$ to be jointly low rank, we form a compact source sample set whose distribution is close to the target domain. The whole procedure is unsupervised without utilizing any label information. This figure is best viewed in color.

Daume III [9] *et al*. proposed the Feature Replication (FR) by using the simple augmented features of the source and target for SVM training. Yang *et al*. [25] proposed the adaptive SVM (A-SVM) method in which the target classifier $f^t(x)$ was adapted from the auxiliary classifier $f^s(x)$, whereby the training boiled down to learn the perturbation $\triangle f(x)$ such that $f^t(x) = f^s(x) + \triangle f(x)$. Similarly, Jiang *et al*. [14] proposed the Cross-Domain SVM (CDSVM) method, which defined a weight for each source sample based on k-nearest neighbors and then re-trained the SVM classifier to update the weights. There are also some other works [11, 12] using multiple kernel learning to align the distributions between source and target domain. In addition, Saenko *et al*. [21] proposed a metric learning method to adapt the acquired visual models in source domain to a new domain and minimize the variance between different feature distributions. The most relevant one to our proposal is [13], which proposed an unsupervised incremental learning algorithm. Specifically, they proposed to create a sequence of intermediate representation subspaces (hence incremental) between the source and target domains to account for the domain shift, by which the source label information can be "propagated" to the target domain. In contrast, we focus on direct transformation here but emphasize sample correlation and noise/outlier removal here, though during the transformation our setting is also unsupervised.

Methodologically, our work is also related to low-rank matrix recovery [18, 2, 3]. In particular, Robust PCA [24] aimed to decompose a corrupted low-rank matrix X into a clean low-rank matrix Z and a sparse matrix E that accounted for sparse errors. Moreover, Chen *et al*. [6] proposed to use a low-rank structure to capture the correlation of different tasks for multi-task learning [5] while using the $\ell_{2,1}$ norm to remove outliers. Differently, our proposed method takes advantages of the low rank and group sparsity structure to seek for a transformation function that can bridge the distribution gaps between the different domains.

3. Robust Domain Adaptation via Low-Rank Reconstruction

In this section, we will introduce our visual domain adaptation method based on low-rank reconstruction. We consider two scenarios in the realistic visual domain adaptation applications. The first scenario is the single source domain adaptation, in which there is only one source domain to be adapted to the target domain. The second is the multiple source domain adaptation, which simultaneously adapt multiple source domains to the target domain.

3.1. Single Source Domain Adaptation

Suppose we have a set of n samples $S = [s_1, \ldots, s_n] \in \mathbb{R}^{d\times n}$ in a single source domain, and a set of p samples $T = [t_1, \ldots, t_p] \in \mathbb{R}^{d\times p}$ in the target domain, where d is the dimension of the feature vector. Our goal is to find a transformation matrix $W \in \mathbb{R}^{d\times d}$ to transform the source domain S into an intermediate representation matrix such that the following relation holds:

$$WS = TZ, \tag{1}$$

where $WS = [Ws_1, \ldots, Ws_n] \in \mathbb{R}^{d\times n}$ denotes the transformed matrix reconstructed by the target domain and $Z = [z_1, ..., z_n] \in \mathbb{R}^{p\times n}$ is the reconstruction coefficient matrix with each $z_i \in \mathbb{R}^p$ being the reconstruction coefficient vector corresponding to the transformed sample Ws_i. In this way, each transformed source sample will be linearly reconstructed by the target samples, which may significantly reduce the disparity of the domain distributions. However, the above formula finds the reconstruction of each source sample independently, and hence may not capture any structure information of the source domain S. Another issue with the reconstruction in Eq. (1) is that it cannot handle the undesirable noises and outliers in the source domain that have no association w.r.t. the target domain. Such noise and outlier information is frequently observed in visual domain adaptation, especially when the source samples are collected from the web. To effectively solve the above issues, we formulate the domain adaptation problem as the following objective function:

$$\begin{aligned} \min_{W,Z,E} \quad & \mathrm{rank}(Z) + \alpha\|E\|_{2,1}, \\ s.t. \quad & WS = TZ + E, \\ & WW^\top = I, \end{aligned} \tag{2}$$

where $\mathrm{rank}(\cdot)$ denotes the rank of a matrix, $\|E\|_{2,1} = \Sigma_{j=1}^{n}\sqrt{\sum_{i=1}^{d}(E_{ij})^2}$ is called $\ell_{2,1}$ norm, and $\alpha > 0$ is the tradeoff parameter. The constraint $WW^\top = I$ is imposed to ensure the obtained W is a basis transformation matrix.

Now we explain the rationality of the above objective function. First, the minimization of $\mathrm{rank}(Z)$ tends to find a reconstruction coefficient matrix with the lowest rank structure. This essentially couples the reconstruction of different source samples together, which captures the relatedness of all the source samples. Second, the minimization of $\|E\|_{2,1}$ encourages the error columns of E to be zero, based on the assumption that some samples in the source domain are noises or outliers, while the others are clean enough to be successfully adapted. By decomposing the noise and outlier information in the source domain into the matrix E, the adaptation becomes more robust to noises and outliers.

The above optimization problem is difficult to solve due to the discrete nature of the rank function. Fortunately, the following optimization provides a good surrogate for problem (2):

$$\begin{aligned} \min_{W,Z,E} \quad & \|Z\|_* + \alpha\|E\|_{2,1}, \\ s.t. \quad & WS = TZ + E, \\ & WW^\top = I, \end{aligned} \tag{3}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, i.e., the sum of the singular values of the matrix.

Once we obtain the optimal solution $(\hat{W}, \hat{Z}, \hat{E})$, we can transform the source data into the target domain in the following way:

$$\hat{W}S - \hat{E} = [\hat{W}s_1 - \hat{e}_1, \ldots, \hat{W}s_n - \hat{e}_n], \tag{4}$$

where $\hat{e}_i$ denotes the ith column of matrix $\hat{E}$. Finally, the transformed source samples will be mixed with the target samples T as the augmented training samples for training the classifiers, which will be used to perform recognition on the unseen test samples in the target domain.

3.2. Multiple Source Domain Adaptation

While most domain adaptation methods only adapt the information from a single source domain to the target domain [11, 14, 21], we often wish to simultaneously adapt multiple source domains into the target domain to improve the generalization ability of the visual classifiers. However, this is a challenging task since the distributions of the individual source domains may be significantly different from

Algorithm 1 Solve Problem (5) by Inexact ALM

Input: Target domain $T \in R^{d\times p}$ and multiple source domains, $\{S_i \in R^{d\times n}\}_{i=1}^{M}$, parameters: α, β .

Initialize: $Q = J = 0, E_i = 0, Y_i = 0, U_i = 0, V_i = 0, \mu = 10^{-7}, W_i = I, i = 1, 2, ..., M$.

while not converged **do**

Fix the others and update $F_1, ..., F_M$ by $F_i = \arg\min_{F_i} \frac{1}{\mu}\|F_i\|_* + \frac{1}{2}\|F_i - (Z_i + \frac{Y_i}{\mu})\|_F^2$.

Fix the others and update $W_1, ..., W_M$ by $W_i = (S_iS_i^\top)^{-1}[(J_i + TZ_i + E_i)S_i^\top - (U_i + V_i)\frac{S_i^\top}{\mu}]$, $W_i \leftarrow$ orthogonal(W_i).

Fix the others and update $Z_1, ..., Z_M$ by $Z_i = (I + T^\top T)^{-1}[T^\top(W_iS_i - E_i) + \frac{1}{\mu}(T^\top V_i - Y_i) + F_i]$.

Fix the others and update $J_1, ..., J_M$ by $J_i = \arg\min_{J_i} \frac{\beta}{\mu}\|J_i\|_* + \frac{1}{2}\|J_i - (W_iS_i + \frac{U_i}{\mu})\|_F^2$.

Fix the others and update $E_1, ..., E_M$ by $E_i = \arg\min_{E_i} \frac{\alpha}{\mu}\|E_i\|_{2,1} + \frac{1}{2}\|E_i - (J_i - TZ_i + \frac{V_i}{\mu})\|_F^2$.

Update Multipliers
$Y_i = Y_i + \mu(Z_i - F_i)$,
$U_i = U_i + \mu(W_iS_i - J_i)$,
$V_i = V_i + \mu(W_iS_i - TZ_i - E_i)$.

Update the parameter μ by $\mu = \min(\mu\rho, 10^{10})$, where $\rho = 1.2$.

Check the convergence condition: $\forall i = 1, ..., M$, $Z_i - F_i \longrightarrow 0$, $W_iS_i - J_i \longrightarrow 0$, $W_iS_i - TZ_i - E_i \longrightarrow 0$.

end while

Output: $Z_i, E_i, W_i, i = 1, 2, ..., M$.

each other. In the following, we propose a multi-task low-rank reconstruction method that jointly adapt the multiple source domains into the target domain.

Suppose we have M source domains, $S_1, S_2, ..., S_M$, where each $S_i \in \mathbb{R}^{d\times n}$ is the feature matrix of the ith source domain. Our multi-task low-rank domain adaptation method can be formulated as:

$$\begin{aligned} \min_{Z_i,E_i,W_i} & \sum_{i=1}^{M}(\|Z_i\|_* + \alpha\|E_i\|_{2,1}) + \beta\|Q\|_*, \\ s.t. \quad & W_iS_i = TZ_i + E_i, \\ & W_iW_i^\top = I, i = 1, ..., M, \end{aligned} \tag{5}$$

where $\alpha, \beta > 0$ are two tradeoff parameters, W_i, Z_i and E_i are the transformation matrix, coefficient matrix and sparse error matrix of the ith source domain respectively. The matrix Q is a matrix formed by $Q = [W_1S_1|\ W_2S_2|\ ...|\ W_MS_M] \in \mathbb{R}^{d\times(M\times n)}$, where $W_iS_i \in \mathbb{R}^{d\times n}$ represents the ith transformed source domain.

Comparing with the single domain adaptation formulation in Eq. (3), the proposed multi-task domain adaptation objective is characterized by: 1) For each source domain S_i, the low rank and sparsity constraints are still used for seeking the transformation matrix W_i, which preserves the relatedness structure and provides noise tolerant properties. 2) The combined Q is enforced to be low rank, which is specifically added to discover a low-rank structure across different source domains and thus further reduce the distribution disparity in a collective way.

Like the case with a single source domain, after obtaining the optimal solution $(W_i, Z_i, E_i), i = 1, \ldots, M$, we can transform each source domain as $W_iS_i - E_i$ and then combine all source domains together with the target domain T as the training data for training classifiers.

3.3. Optimization

The problem (5) is a typical mixed nuclear norm and $\ell_{2,1}$ norm optimization problem [16]. However, it differs from the existing optimization formulations in that it has the matrix orthogonality constraints $W_iW_i^\top = I, i = 1, \ldots, M$. Following most existing orthogonality preserving methods in the literature [23], we use matrix orthogonalization to deal with these constraints. The basic idea is to first solve each W_i without the orthogonality constraint, and then convert each obtained W_i into an orthogonal matrix via matrix factorization such as SVD. Therefore, the optimization can still be easily solved by the existing nuclear norm and $\ell_{2,1}$ norm optimization methods.

To solve the optimization problem in (5), we first convert it into the following equivalent form:

$$\begin{aligned} \min_{J,F_i,Z_i,E_i,W_i} & \sum_{i=1}^{M}(\|F_i\|_* + \alpha\|E_i\|_{2,1}) + \beta\|J\|_*, \\ s.t. \quad & W_iS_i = TZ_i + E_i, \\ & Q = J, \\ & Z_i = F_i, i = 1, ..., M, \end{aligned} \tag{6}$$

where $J = [J_1, \cdots, J_M]$ with each J_i corresponding to W_iS_i and the orthogonality constraints are ignored. The above equivalent problem can be solved by the Augmented Lagrange Multiplier (ALM) method [16] which minimizes

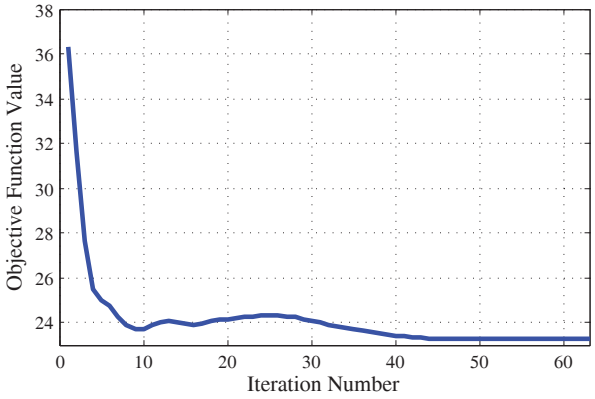

Figure 3. The convergence curve of Algorithm 1 on the three-domain object recognition experiment (see section 4.2).

the augmented lagrange function in the following form:

$$\min_{J_i,F_i,Z_i,E_i,W_i,Y_i,U_i,V_i} \beta\|J\|_* + \sum_{i=1}^{M}(\|F_i\|_* + \alpha\|E_i\|_{2,1}) + \sum_{i=1}^{M}(\langle U_i, W_iS_i - J_i\rangle + \langle Y_i, Z_i - F_i\rangle + \frac{\mu}{2}\|Z_i - F_i\|_F^2 + \langle V_i, W_iS_i - TZ_i - E_i\rangle + \frac{\mu}{2}\|W_iS_i - J_i\|_F^2 + \frac{\mu}{2}\|W_iS_i - TZ_i - E_i\|_F^2), \tag{7}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product operator, $\mu > 0$ is a penalty parameter and $Y_1, ..., Y_M, U_1, ..., U_M$ and $V_1, ..., V_M$ are the Lagrange multipliers. In this paper, we select the *inexact* ALM method for the optimization to take advantage of its fast convergence speed. The procedure of the optimization procedure can be shown in Algorithm 1. Note that the sub-problems involved in the optimization all have closed-form solutions. Specifically, step 2 and step 5 can be solved by adopting singular value thresholding operator [4] meanwhile the step 6 can be solved by the analytic solution in [17].

We implement the Algorithm 1 on the MATLAB platform of a Six-Core Intel Xeon Processor X5660 with 2.8 GHz CPU and 32 GB memory, and observe that the iterative optimization converges fast. For example, in the three-domain object recognition experiment (see Section 4.2) involving 496 samples, one iteration between step 1 and step 10 in Algorithm 1 can be finished within 2 seconds. Furthermore, as each optimization sub-problem in Algorithm 1 will monotonically decrease the objective function, the algorithm will converge. Figure 3 shows the convergence process of the iterative optimization, which is captured when adapting the *dslr* source domain to *webcam* target domain on the three-domain objective recognition dataset. As can be seen, the objective function converges to the minimum after about 40 iterations.

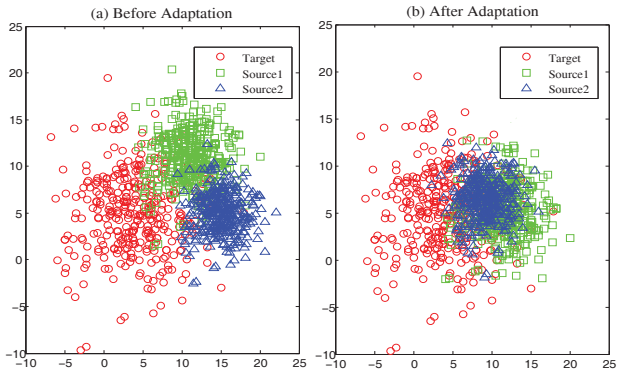

Figure 4. Toy experiment illustrating the effectiveness of our proposed method. In (a), the red samples denote the target domain while the green and blue samples denote two different source domains. As can be seen, the distributions of the three domains are significantly separated from each other. In (b), our method is able to map the samples in each source domain into a compact region of the target domain such that their distributions become more consistent. This figure is best viewed in color.

4. Experiments

In this section, we will evaluate the effectiveness of our proposed method, referred to as Robust Domain Adaptation with Low-rank Reconstruction (RDALR), on various challenging visual domain adaptation tasks including object recognition and video event detection. In each task, the performance of the following domain adaptation methods will be compared. (1) Naive Combination (NC). We directly augment the target domain with samples from the source domain without any transformation. (2) Adaptive SVM (A-SVM) [25]. In this method, a SVM classifier is first trained in the source domain, and then adjusted to fit the training samples in the target domain. (3) Noisy Domain Adaptive Reconstruction (NDAR). In this case, we do not consider to remove the noise and outlier information in the source domain, and this can be achieved by removing the E_i term in Eq. (5). (4) Our proposed RDALR method. (5) The state-of-the-art domain adaptation methods in the recent literature [10, 13, 21].

We use the one-vs-all SVM as the classifier for cross domain classification. After the domain adaptation, the training samples in the source domain (after transformation) and the target domain are combined together for SVM training, and the obtained SVM classifiers will be used to perform testing on the unseen samples in the target domain. To determine the appropriate parameter setting for our method, we vary the values of α and β over the grid of $\{10^{-4}, 10^{-3}, \ldots, 1\}$ and then choose the optimal values based on five-fold cross validation. Similarly, the optimal parameter C in A-SVM and SVM is selected from $\{2^{-5}, 2^{-2}, ..., 2^{3}\}$ based on cross validation.

Table 1. Performance comparison (%) of single source domain adaptation on the three-domain object recognition benchmark.

		Compared Methods						Our Method
Source	Target	DAML [21]	ITML [10]	UDA [13]	NC	A-SVM	NDAR	RDALR
webcam	dslr	27	18	19 ± 1.2	22.13 ± 1.1	25.96 ± 0.7	30.11 ± 0.8	32.89 ± 1.2
dslr	webcam	31	23	26 ± 0.8	32.17 ± 1.4	33.01 ± 0.8	35.33 ± 1.2	36.85 ± 1.9
amazon	webcam	48	41	39 ± 2.0	41.29 ± 1.3	42.23 ± 0.9	47.52 ± 1.1	50.71 ± 0.8

Table 2. Performance comparison (%) of multiple source domain adaptation on the three-domain object recognition benchmark.

		Compared Methods				Our Method
Source	Target	UDA [13]	NC	A-SVM	NDAR	RDALR
amazon, dslr	webcam	31 ± 1.6	20.62 ± 1.8	30.36 ± 0.6	33.23 ± 1.6	36.85 ± 1.1
amazon, webcam	dslr	25 ± 0.4	16.38 ± 1.1	25.26 ± 1.1	29.21 ± 0.9	31.17 ± 1.3
dslr, webcam	amazon	15 ± 0.4	16.87 ± 0.7	17.31 ± 0.9	19.08 ± 1.1	20.89 ± 0.9

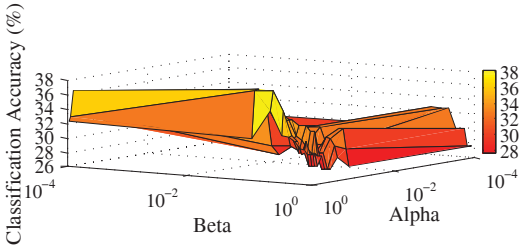

Figure 5. Classification accuracy of our method as a function of the combination of parameter α and β. This figure is generated when performing multiple source domains adaptation from *amazon* and *dslr* to *webcam*.

4.1. An Illustrative Toy Example

In this subsection, we use the toy data to illustrate that our proposed method is able to transform the samples from multiple source domains into the target domain such that the distribution variance is significantly reduced. As shown in Figure 4(a), we randomly generate three clouds of samples, each of which contains about 400 samples. We simply treat the red samples as the target domain while assuming the blue and green samples are two different source domains. Obviously, the distributions of the three domains are significantly different despite some partial overlapping. We apply our method to map the two source domains into target domain simultaneously while removing the undesirable noise information, and the result can be shown in Figure 4(b). As can be seen, the two source domains are mixed together into the target domain in a compact region, which demonstrates the effectiveness of our proposed method in reducing the difference of domain distributions in domain adaptation.

4.2. Experiment on Three-Domain Object Benchmark [21]

We first test the proposed method on the visual domain adaptation benchmark dataset [21] that is collected from three different domains, *amazon*, *dslr* and *webcam*. This dataset consists of 31 different object categories varying from bike and notebook to bookcase and keyboard, and the total number of images is $4,652$. The *dslr* and *webcam* domain have around 30 images per category while the amazon domain has an average of 90 images per category. For low-level features, we adopt the SURF feature from [21] and all images are represented by 800-dimension Bag-of-Words (BoW) feature.

Since our method can handle single domain and multiple domain adaptation, we use two different experimental settings to test the performance. Following the experiment setting in [21], for source domain samples we randomly select 8 images per category in webcam/dslr, and select 20 images per category in amazon. Meanwhile, we select 3 images per category as the target domain for amazon/webcam/dslr. These images are used for domain adaptation and classifier training, while the remaining unseen images in the target domain are used as the test set for performance evaluation. To make our results comparable to the previous works, we also use the SVM with RBF kernel as the classifier where the average classification accuracy over the 31 object categories on the test set is used as the evaluation metric. Each experiment is repeated 5 times based on the 5 random splits and the average classification accuracy and the standard derivation over all categories are reported.

- **Single source domain adaptation**. Table 1 shows the performance of different methods, in which we also quote the results directly from [10, 13, 21]. From the results, we have the following observations: (1) All domain adaptation methods produce better results than NC, which confirms the superiority of domain adaptation. (2) Our RDALR method significantly outperforms the Domain Adaptive Metric Learning (DAML) [21], A-SVM and Unsupervised Domain Adaptation (UDA) [13] methods, which verifies that the low-rank reconstruction can better reduce the disparity of domain distributions comparing

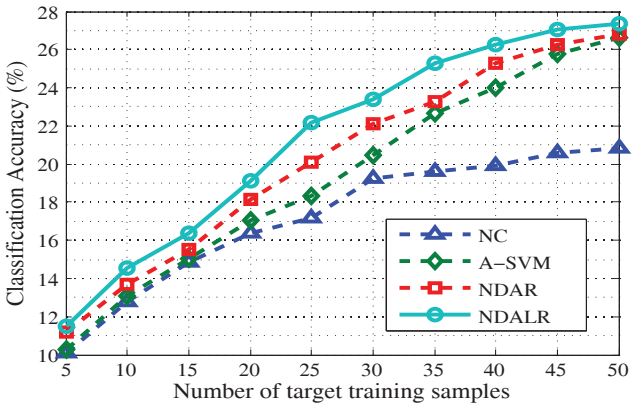

Figure 6. Performance of different methods on Caltech 256 dataset, where the per category number of training images in the target domain varies from 5 to 50. The per category number of images from the Bing source domain is fixed at 10.

with the state-of-the-art methods in the literature. (3) RDALR clearly outperforms the NDAR, since the latter does not remove the undesired noise information in the source domain.

- **Multiple source domain adaptation**. We then evaluate the performance of multiple source domain adaptation on the dataset. We use the same setting as in the single domain adaptation experiment. However, the difference is that the samples in the target domain are combined with samples from multiple source domains for training the classifiers. Table 2 shows three different combination of multiple source domains. One closely relevant work is the UDA [13] method, where the authors attempt to learn an intermediate representative subspace on Grassmann manifold between the source and target domain. As shown, our method outperformed this baseline method by 6.35%, 5.27%, and 5.39% under the three different domain combinations. It also outperforms all other prior methods (NC, A-SVM, and NDAR) in all cases, usually with large margins. This demonstrates that our method is effective for multiple source domain adaptation. Figure 5 shows the performance under various parameter combinations in multiple source domains adaptation experiment (from amazon and dslr to webcam).

4.3. Experiment on Caltech 256

We evaluate our method on a large scale domain adaptation dataset established by [1]. This dataset uses the Caltech 256 as the target domain while adopting the web images crawled from Bing image search engine as the source domain. The Caltech 256 target domain has $30,607$ images falling into 256 object categories. The Bing source domain contains about $120,924$ weakly-labeled images crawled by using each text labels of Caltech 256 as the search keywords. Since the search results contain a lot of irrelevant images, such a domain adaptation task is very challenging.

For each image, we extract the SIFT feature from the keypoints detected by the Difference of Gaussian (DOG) detector [19], and then represent each image as $5,000$-dimensional BoW feature. On the Caltech 256 target domain, we randomly select $\{5, 10, \ldots, 50\}$ images from each category as the training data, and use the rest as the test data. On the Bing source domain, we randomly select from each category 10 images which will be used for the domain adaptation and use the linear SVM as the classifier in the experiment. The average classification accuracy over the 256 categories on the test set are reported as the evaluation metric. Each experiment is repeated three times based on three random splits, and the average result is reported.

Figure 6 shows the experiment results under different numbers of training images in the target domain for all the methods in comparison. As can be seen, the classification results of all methods keep improving as the number of training images in the target domain is increased. However, our proposed method achieves the most significant performance improvements compared to the other methods. Moreover, the performance gains become salient as the number of training images in the target domain increases until the training size becomes large (at which point the target domain may become self sufficient). Again, the experiment results verify the effectiveness of our method.

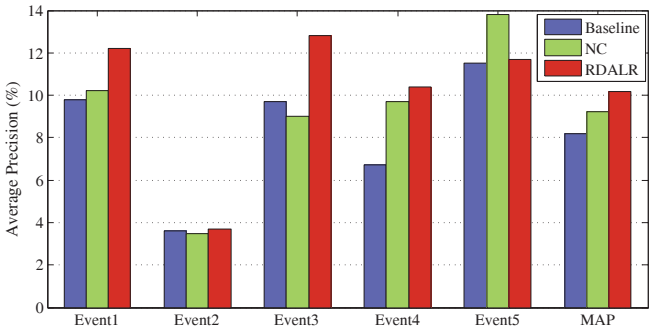

Figure 7. Per-event performance comparison on the TRECVID MED 2011 dataset. This figure is best viewed in color.

4.4. Experiment on TRECVID MED 2011

The TRECVID 2011 Multimedia Event Detection (MED) [26] development dataset contains $10,804$ video clips from $17,566$ minutes video programs falling into five event class and the background class. The five events are "Attempting a board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony" and "Working on a woodworking project", respectively. The dataset is partitioned into a training set with 8783 videos and a testing set with 2021 videos. It is worth noting the small training sample

problem exists in the MED dataset. Specifically, the training set contains around $8,273$ background videos that do not belong to any of the five events, and the average number of training videos for each event is 100. This makes the event detection a challenging task, and also provides a testbed to evaluate our proposed domain adaptation method.

In this experiment, we use the TRECVID MED dataset as our target domain while using the videos crawled from the web as the source domain. For each event, we use the event name as keyword to crawl videos from the YouTube website, and thus obtain a source domain containing 520 YouTube video clips. It is worth noting the crawled videos are very diversified and usually include some noisy videos that are totally irrelevant, posing a great challenging to any domain adaptation method.

Given a video clip, we sample one frame from every two seconds. For each frame, we extract 128-dimensional SIFT feature from the keypoints detected by two kinds of detectors: DoG and Hessian Affine [20]. Then, k-means method is applied to group the SIFT features into $5,000$ clusters. Finally, we aggregate the $5,000$-dimensional features from all sampled frames in a video clip together as the clip-level feature representation. We use the linear SVM as the classifier in the experiment.

Following the TRECVID evaluation, we use average precision (AP) to evaluate the performance for each event, and then calculate the Mean Average Precision (MAP) across the five events as the overall evaluation metric. Figure 7 shows the results on each event, where the baseline results are obtained by training the SVM classifiers only on the target training data without using any YouTube videos. From the results, we have the following observations: (1) Although NC produces higher MAP than the baseline, it performs even worse than the baseline method on event "Feeding an animal" and "Landing a fish". (2) Our method achieves the best average performance compared to the other methods. Moreover, it shows performance improvement on four out of the five events. Specifically, on event "Landing a fish ", our method outperforms the baseline and NC method by 3.29% and 4.22%. This demonstrates the great potential of our method for video event detection. The reason for performance degradation on event "working on a woodworking project" may be caused by the unexpected large cross-domain content variance. Another potential reason is that the visual features used as input to the recognition models may be inadequate (e.g., missing temporal and audio features) to capture the event properties that can persist over different domains.

5. Conclusion

We have introduced a robust visual domain adaptation method to reduce the distribution disparity between the source and target domain. The basic idea is to transform the source samples into an intermediate representation such that each of them can be linearly reconstructed by the target samples. The proposed method captures the intrinsic relatedness of the source samples using a low-rank structure and meanwhile identifies the noise and outlier information using a sparse structure, which allows our method to achieve superior robustness in the domain adaptation task. We demonstrate the effectiveness of our proposed method on extensive domain adaptation benchmarks. In the future, we plan to adopt the low-rank reconstruction as a preprocessing step of the semi-supervised learning so that the distributions of the unlabeled and labeled samples can be more consistent.

References

[1] A. Bergamo, and L. Torresani. Exploiting weakly-labeled web images to improve object classification:A domain adaptation approach. In *NIPS'10*. 2, 7

[2] E. Candes, and B. Recht. Exact matrix completion via convex optimization. In *FCM'09*, 9(6):717-772. 3

[3] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis?. In *J. of ACM'11*, 58(3):11:1-11:37. 3

[4] J.-F. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. In *J. Optimization, SAIM'10*, 20(4):1956-1982. 5

[5] R. Caruana. Multi-task learning. *Machine Learning'97*, 28(1):41-75. 3

[6] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD'11*. 3

[7] W. Dai, Y. Chen, G. Xue, Q. Yang, and Y. Yu. Translated learning: transfer learning across different feature spaces. In *NIPS'08*. 2

[8] H. DamueIII, and D. Marcu. Domain adaptation for statistical classifiers. In *JAIR'06*, 26(1):101-126. 2

[9] H. Daume III. Frustratingly easy domain adaptation. In *ACL'07*. 1, 2

[10] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML'07*. 5, 6

[11] L. Duan, I. Tsang, D. Xu, and S. Maybank. Domain transfer SVM for video concept detection. In *CVPR'09*. 2, 3

[12] L. Duan, D. Xu, I. Tsan, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR'10*. 1, 2

[13] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV'11*. 1, 2, 5, 6, 7

[14] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui. Cross-domain learning methods for high-level visual concept classification. In *ICIP'08*. 1, 2, 3

[15] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In *CVPR'11*. 1, 2

[16] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. In *UIUC Technical Report'09*. 2, 4

[17] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l2,1-norm minimization. In *UAI'09*. 5

[18] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML'10*. 3

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV'04*, 60(2), 91-110. 7

[20] K. Mikolajczyk, and C. Schmid. Scale and affine invariant interest point detectors. In *IJCV'04*, 60(1):63-86. 8

[21] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new Domains. In *ECCV'10*. 1, 2, 3, 5, 6

[22] B.-D. Shai, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. In *Machine Learning'10*, 79(1-2):151-175. 2

[23] Z. Wen, and W. Yin. A feasible method for optimization with orthogonality constraints. In *Rice Univ. Technical Report'10*. 4

[24] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. In *NIPS'09*. 3

[25] J. Yang, R. Yan, and A. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM MM'07*. 2, 5

[26] http://trecvid.nist.gov/. *TRECVID MED 2011*. 7

[27] X. Zhu. Semi-supervised learning literature survey. In *Computer Science, Univ. Wisconsin-Madison'06*. 1