

Recognizing RGB Images by Learning from RGB-D Data

Lin Chen¹ Wen Li² Dong Xu²

¹Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore

²School of Computer Engineering, Nanyang Technological University, Singapore

Abstract

In this work, we propose a new framework for recognizing RGB images captured by the conventional cameras by leveraging a set of labeled RGB-D data, in which the depth features can be additionally extracted from the depth images. We formulate this task as a new unsupervised domain adaptation (UDA) problem, in which we aim to take advantage of the additional depth features in the source domain and also cope with the data distribution mismatch between the source and target domains. To effectively utilize the additional depth features, we seek two optimal projection matrices to map the samples from both domains into a common space by preserving as much as possible the correlations between the visual features and depth features. To effectively employ the training samples from the source domain for learning the target classifier, we reduce the data distribution mismatch by minimizing the Maximum Mean Discrepancy (MMD) criterion, which compares the data distributions for each type of feature in the common space. Based on the above two motivations, we propose a new SVM based objective function to simultaneously learn the two projection matrices and the optimal target classifier in order to well separate the source samples from different classes when using each type of feature in the common space. An efficient alternating optimization algorithm is developed to solve our new objective function. Comprehensive experiments for object recognition and gender recognition demonstrate the effectiveness of our proposed approach for recognizing RGB images by learning from RGB-D data.

1. Introduction

With the rapid adoption of affordable equipments (e.g., Kinect sensors) for capturing depth information, there is an increasing research interest in developing new technologies for various visual recognition tasks (e.g., object recognition, face and gender recognition) using depth images. One common assumption in most visual recognition methods including the recent works using both color and depth images [23, 21] is that the training and testing samples come

from the same data distribution. If one dataset is used for training and another dataset is used for testing, the performances of most existing visual recognition methods will degrade significantly [28] because the feature distributions of samples from different datasets may have very different statistical properties. Meanwhile, to cope with the considerable variation in feature distributions, new domain adaptation methods were recently developed for different computer vision applications [15, 22, 14, 16, 7, 6, 10, 8, 13, 1].

In this work, we propose a new framework for recognizing RGB images captured with the conventional cameras by leveraging a set of labeled RGB-D data, in which the depth features can be additionally extracted from the depth images. Our work is based on the observation that several labeled RGB-D datasets were recently released for various vision recognition tasks [23, 21] as well as the recent progress on learning using privileged information [29, 26], which shows the additional features (i.e., privileged information) that are not available at the testing stage are still useful for many classification tasks. We formulate our task as a new unsupervised domain adaptation (UDA) problem, in which we have the single-view visual features extracted from the RGB images in the target domain while we have both the visual features and the depth features in the source domain.

Specifically, to effectively utilize the additional depth features in the source domain, we seek two optimal projection matrices to map the samples from both domains into a common space such that we can preserve as much as possible the correlations between the visual features and depth features. To effectively employ the source samples for learning the target classifier, we reduce the data distribution mismatch between two domains by minimizing the Maximum Mean Discrepancy (MMD) criterion [17] for each type of feature in the common space, which compares the data distributions based on the distance between the means of samples from two domains. Motivated by the above two aspects, we propose a new SVM based objective function to simultaneously learn the two projection matrices and the optimal target classifier, in which we expect the source samples from different classes can be well separated when using each type of feature in the common

space. We also develop an efficient alternating optimization algorithm to solve this non-trivial optimization problem. Our comprehensive experiments for object recognition and gender recognition demonstrate that our approach (referred to as domain adaptation from multi-view to single-view or DA-M2S in short) outperforms several state-of-the-art methods including the existing UDA methods as well as SVM+ [29] and Rank Transfer [26] that use the depth features as privileged information without coping with the data distribution mismatch.

We summarize the main contributions of this paper as follows: 1) we propose a new framework for recognizing RGB images by leveraging a set of labeled RGB-D data. From our framework, we formulate a new domain adaptation problem, where we have the additional features in the source domain that are not available in the target domain; 2) We propose a new UDA method DA-M2S and the extensive experiments demonstrate its effectiveness for recognizing RGB images by learning from RGB-D data.

2. Related Work

Domain Adaptation: Our work is related to domain adaptation, in which the distribution of test data is different from that of training data [20, 16, 15, 14, 13, 1, 7]. The existing unsupervised domain adaptation (UDA) methods assume there are no labeled data in the target domain and these methods can be generally divided into three categories: sample reweighting approaches, feature (transform) based approaches and classifier based approaches. The sample reweighting approaches like kernel mean matching (KMM) [20] aim to reduce the domain distribution mismatch by reweighting the samples in the source domain. The feature (transform) based approaches seek new domain-invariant features or learn new feature transformations for domain adaptation. For example, SGF [16] and GFK [15] were proposed based on the Grassmann manifold assumption, and GFK was further extended in [14] by selecting the landmarks from the source domain. Recently, the domain invariant projection (DIP) [1] method was proposed to learn a domain invariant subspace, while the subspace alignment (SA) method [13] was developed to align two subspaces from two domains to reduce the domain distribution mismatch. The classifier based approaches directly learn the target classifiers (*e.g.*, SVM based classifiers) for domain adaptation. For example, Duan *et al.* [7] proposed a learning method called DAM by using the virtual labels generated from pre-learned classifiers. However, the existing UDA methods assume the samples from the source domain share the same feature representation with those from the target domain. For these methods, it is unclear how to effectively utilize the additional depth features in the source domain.

Recently, heterogeneous domain adaptation (HDA)

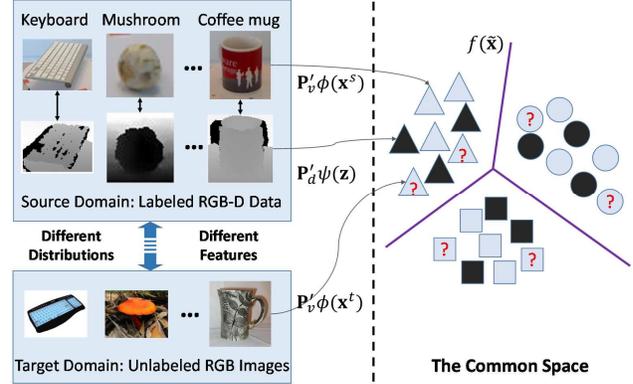


Figure 1. Object recognition in RGB images by using labeled RGB-D data, where we have two views of features (*i.e.*, visual features and depth features) in the source domain, and a single view of visual features in the target domain. Grey color and black color respectively denote the projected nonlinear visual and depth features in the common space. The source domain samples from different categories are represented by different shapes and the unlabeled samples from the target domain are with question marks.

methods [9, 24, 22] were also proposed, in which the samples from different domains are generally represented by different types of features. However, labeled samples in the target domain must be provided in the existing HDA methods [9, 24, 22], while we do not require any labeled target domain samples in this work. Moreover, the samples in the source domain are represented by using only one type of features in the existing UDA and HDA methods. In contrast, in this work we have both visual and depth features in the source domain, while the depth features are not available at the testing stage.

Our work is also different from the existing multi-view domain adaptation methods [31] and the recent work called multi-domain adaptation from heterogeneous sources (MDA-HS) [3]. In [31], all the samples in the source and target domains have multiple types of features, while in [3] the samples from the target domain have all types of features from all the source domains. In contrast, in our work we only have single-view features in the target domain. Our work is different from existing multi-domain adaptation methods [7, 4], because we have additional depth features in the source domain.

Learning Using Privileged Information: Our work is also related to the recent progress on learning using privileged information [29, 26], in which training data contains additional features (*i.e.*, privileged information) that are not available at the testing stage. However, these works [29, 26] assume that the training and testing samples come from the same data distribution. In contrast, our work explicitly copes with the data distribution mismatch between two domains.

3. The Proposed Approach

In this paper, we denote a vector/matrix by a lower-case/uppercase letter in bold. The transpose of a vector or a matrix is denoted by the superscript $'$. We define \mathbf{I}_n as the $n \times n$ identity matrix and $\mathbf{O}_{n \times m}$ as the $n \times m$ matrix of all zeros. We also define $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ as the $n \times 1$ column vectors of all zeros and all ones, respectively. For simplicity, we use $\mathbf{I}, \mathbf{O}, \mathbf{0}$ and $\mathbf{1}$ when the dimension is obvious.

3.1. Problem Setup

In our problem, given a set of labeled RGB-D data in the source domain and unlabeled RGB images in the target domain, we aim to learn a robust classifier to predict the class labels of RGB images in the target domain.

We extract visual features and depth features from the RGB images and depth images, respectively. The source domain samples can be represented as $\{(\mathbf{z}_i, \mathbf{x}_i^s, y_i)\}_{i=1}^{n_s}$ where \mathbf{z}_i and \mathbf{x}_i^s are respectively the depth feature and the visual feature and $y_i \in \mathcal{Y}$ is the label for the i -th source domain sample with $\mathcal{Y} = \{1, \dots, K\}$ being the set of all possible labels and n_s is the total number of samples in the source domain. We also denote $\mathbf{y} = [y_1, \dots, y_{n_s}]'$ as the label vector for the source domain samples. Similarly, the target domain samples can be represented as $\{\mathbf{x}_i^t\}_{i=1}^{n_t}$ where \mathbf{x}_i^t is the visual feature for the i -th target domain sample and n_t is the total number of samples in the target domain.

As shown in Fig. 1, the major challenges are from different features and the data distribution mismatch between two domains. To handle the first challenge, we propose to project different features into a common space by learning two projection matrices \mathbf{P}_d and \mathbf{P}_v for the depth features (*i.e.*, \mathbf{z}) and visual features (*i.e.*, \mathbf{x}^s and \mathbf{x}^t), respectively. The two types of features are transformed to the same representation in this common space by using the learnt projection matrices. Moreover, to handle the data distribution mismatch between two domains, we also minimize the Maximum Mean Discrepancy (MMD) [17] criterion for each type of feature in the common space. Then we can learn a robust classifier which aims to well separate all the source domain labeled samples in this common space. Intuitively, a suitable common space will be beneficial for learning a more robust classifier; and the robust classifier can also help us find a more discriminative common space. Therefore, we propose to simultaneously seek the projection matrices and learn the optimal classifier. Based on the empirical risk minimization (ERM) principle, we formulate our learning problem as follows:

$$\min_{f, \mathbf{P}_v, \mathbf{P}_d} \mu \Omega(\mathbf{P}_d, \mathbf{P}_v) + r(f) + C \ell(f, \mathbf{P}_d, \mathbf{P}_v), \quad (1)$$

where f is the target classifier, $r(f)$ is the regularizer term to control the complexity of the classifier, $\Omega(\cdot)$ is the regularizer term on the projection matrices \mathbf{P}_v and \mathbf{P}_d , $\ell(\cdot)$ is

the loss on the training samples, and μ and C are the trade-off parameters. We will introduce the details of these terms below.

3.2. The Formulation

For the sake of generality, we consider the kernelized case in this work, *i.e.*, the projection matrices \mathbf{P}_d and \mathbf{P}_v are learnt based on the nonlinear features induced by the kernels of the depth features and visual features. The linear case can be easily derived in a similar manner.

Formally, let us denote the nonlinear depth feature as $\psi(\mathbf{z}) \in \mathbb{R}^{m_d}$ and the nonlinear visual feature as $\phi(\mathbf{x}) \in \mathbb{R}^{m_v}$ where m_d and m_v are respectively the dimensions of the nonlinear depth feature and the nonlinear visual feature, which are usually unknown. Then the projection matrices can be defined as $\mathbf{P}_d \in \mathbb{R}^{m_d \times m}$ and $\mathbf{P}_v \in \mathbb{R}^{m_v \times m}$, respectively, where m is the dimension of the common space.

Intuitively, we should preserve as much as possible the useful information from the original features when learning the projection matrices \mathbf{P}_d and \mathbf{P}_v . Inspired by the multi-view learning method KCCA [19], we propose to maximize the correlation between the two types of features based on correspondence information between the two views of features in the source domain. Namely, we minimize the following regularizer:

$$\Omega^{kcca}(\mathbf{P}_d, \mathbf{P}_v) = -\text{tr}(\mathbf{P}_d' \Psi \Phi_s' \mathbf{P}_v), \quad (2)$$

where $\Psi = [\psi(\mathbf{z}_1), \dots, \psi(\mathbf{z}_{n_s})] \in \mathbb{R}^{m_d \times n_s}$ and $\Phi_s = [\phi(\mathbf{x}_1^s), \dots, \phi(\mathbf{x}_{n_s}^s)] \in \mathbb{R}^{m_v \times n_s}$ are the data matrices of the nonlinear depth and visual features in the source domain. To avoid the trivial solutions for \mathbf{P}_d and \mathbf{P}_v , we also introduce the constraint $\mathbf{P}_d' \Psi \Psi' \mathbf{P}_d + \mathbf{P}_v' \Phi_s \Phi_s' \mathbf{P}_v = \mathbf{I}_m$.

Moreover, to handle the distribution mismatch between the source and target domains, we also expect the distributions of samples from different domains are similar in the common space. Specifically, we employ the Maximum Mean Discrepancy (MMD) [17] criterion to measure the distribution mismatch between two domains. Considering we have two types of features in the source domain, we apply the MMD criterion for each type of feature in the common space, and obtain the following regularizer:

$$\begin{aligned} \Omega^{mmd}(\mathbf{P}_d, \mathbf{P}_v) = & \frac{1}{2} \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{P}_v' \phi(\mathbf{x}_i^s) - \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbf{P}_v' \phi(\mathbf{x}_j^t) \right\|^2 \\ & + \frac{1}{2} \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{P}_d' \psi(\mathbf{z}_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbf{P}_d' \psi(\mathbf{z}_j) \right\|^2. \end{aligned}$$

Then, our regularizer $\Omega(\mathbf{P}_d, \mathbf{P}_v)$ in (1) can be obtained by combining the two regularizers:

$$\Omega(\mathbf{P}_d, \mathbf{P}_v) = \Omega^{kcca}(\mathbf{P}_d, \mathbf{P}_v) + \lambda \Omega^{mmd}(\mathbf{P}_d, \mathbf{P}_v), \quad (3)$$

where λ is a tradeoff parameter for balancing the two terms.

Now we develop the detailed form of our DA-M2S method¹ based on multi-class SVM [5]. For ease of presentation, we denote one training sample in the common space as $\tilde{\mathbf{x}}$, which can be $\mathbf{P}'_d\psi(\mathbf{z})$ or $\mathbf{P}'_v\phi(\mathbf{x}^s)$. So in total we have $2n_s$ labeled samples in the source domain, where n_s samples are based on the visual features and the other n_s samples are based on the depth features. We define the decision function as $f(\tilde{\mathbf{x}}) = \operatorname{argmax}_{k=1,\dots,K} \mathbf{w}'_k\tilde{\mathbf{x}}$ with \mathbf{w}_k being the weight vector for the k -th class. By defining a matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_K]$, we write the objective function in (1) as follows:

$$\begin{aligned} \min_{\substack{(\mathbf{P}_d, \mathbf{P}_v) \in \mathcal{P}, \\ \mathbf{W}, \xi_i}} \quad & \mu\Omega(\mathbf{P}_d, \mathbf{P}_v) + \frac{1}{2}\|\mathbf{W}\|_F^2 + C \sum_{i=1}^n \xi_i, \quad (4) \\ \text{s.t.} \quad & \mathbf{w}'_{y_i}\tilde{\mathbf{x}}_i - \mathbf{w}'_k\tilde{\mathbf{x}}_i \geq e_i^k - \xi_i, \quad (5) \\ & \forall i = 1, \dots, 2n_s, k = 1, \dots, K, \end{aligned}$$

where $\|\mathbf{W}\|_F^2$ is the regularizer to control the complexity of the classifier f , e_i^k is an indicator which equals to 0 if $y_i = k$ and 1 otherwise, $\mathcal{P} = \{(\mathbf{P}_d, \mathbf{P}_v) | \mathbf{P}'_d\mathbf{\Psi}\mathbf{\Psi}'\mathbf{P}_d + \mathbf{P}'_v\mathbf{\Phi}_s\mathbf{\Phi}'_s\mathbf{P}_v = \mathbf{I}_m\}$ is the feasible set of $(\mathbf{P}_d, \mathbf{P}_v)$, and μ and C are the tradeoff parameters as defined in (1).

3.3. The Duality

Since the dimensions of the nonlinear features $\psi(\mathbf{z})$ and $\phi(\mathbf{x})$ (i.e., m_d and m_v) are usually unknown, in this section, we derive the kernel form of the problem in (4) using the Lagrangian method. First, similar as in KCCA, we represent the projection matrices as the combination of the nonlinear features, i.e., $\mathbf{P}_d = \mathbf{\Psi}\mathbf{A}_d$, $\mathbf{P}_v = \mathbf{\Phi}_s\mathbf{A}_v$ where $\mathbf{A}_d, \mathbf{A}_v \in \mathbb{R}^{n_s \times m}$ are the combination coefficient matrices. We also define a matrix $\mathbf{A} = [\mathbf{A}'_d, \mathbf{A}'_v]' \in \mathbb{R}^{2n_s \times m}$, and then the regularizer in (2) becomes,

$$\Omega^{kcca}(\mathbf{A}) = -\frac{1}{2}\operatorname{tr}(\mathbf{A}'\mathbf{B}^{kcca}\mathbf{A}), \quad (6)$$

where

$$\mathbf{B}^{kcca} = \begin{bmatrix} \mathbf{O} & \mathbf{K}_d\mathbf{K}_v^s \\ \mathbf{K}_v^s\mathbf{K}_d & \mathbf{O} \end{bmatrix},$$

with $\mathbf{K}_d = \mathbf{\Psi}'\mathbf{\Psi} \in \mathbb{R}^{n_s \times n_s}$ being the kernel matrix for the source domain depth features, and $\mathbf{K}_v^s = \mathbf{\Phi}_s'\mathbf{\Phi}_s \in \mathbb{R}^{n_s \times n_s}$ being the kernel matrix for the source domain visual features.

Similarly, the regularizer $\Omega^{mmd}(\mathbf{P}_d, \mathbf{P}_v)$ can be represented as

$$\Omega^{mmd}(\mathbf{A}) = \frac{1}{2}\operatorname{tr}(\mathbf{A}'\mathbf{B}^{mmd}\mathbf{A}), \quad (7)$$

where

$$\mathbf{B}^{mmd} = \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{K}_v^s & 2\mathbf{K}_v^{st} \end{bmatrix} \text{ss}' \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{K}_v^s & 2\mathbf{K}_v^{st} \end{bmatrix}',$$

¹Note our method can be readily extended to other SVM based formulations.

with $\mathbf{K}_v^{st} = \mathbf{\Phi}'_s\mathbf{\Phi}_t \in \mathbb{R}^{n_s \times n_t}$ being the kernel matrix between the source data and target data based on the visual features, $\mathbf{\Phi}_t = [\phi(\mathbf{x}_1^t), \dots, \phi(\mathbf{x}_{n_t}^t)] \in \mathbb{R}^{m_v \times n_t}$ being the data matrix of nonlinear visual features in the target domain, and $\mathbf{s} = [\frac{1}{n_s}\mathbf{1}'_{n_s}, -\frac{1}{n_t}\mathbf{1}'_{n_t}]' \in \mathbb{R}^{(n_s+n_t)}$.

Let us define $\mathbf{B} = \mathbf{B}^{kcca} - \lambda\mathbf{B}^{mmd} \in \mathbb{R}^{2n_s \times 2n_s}$. Combining (6) and (7), we represent our regularizer $\Omega(\mathbf{P}_v, \mathbf{P}_d)$ in (3) w.r.t. \mathbf{A} as follows:

$$\Omega(\mathbf{A}) = -\frac{1}{2}\operatorname{tr}(\mathbf{A}'\mathbf{B}\mathbf{A}).$$

Moreover, the feasible set \mathcal{P} in (4) becomes $\mathcal{A} = \{\mathbf{A} = [\mathbf{A}'_d, \mathbf{A}'_v]' | \mathbf{A}'_d\mathbf{K}_d\mathbf{K}_d\mathbf{A}_d + \mathbf{A}'_v\mathbf{K}_v^s\mathbf{K}_v^s\mathbf{A}_v = \mathbf{I}_m\}$.

By introducing one dual variable α_i^k for each constraint in (5) and defining a matrix $\mathbf{\Gamma} \in \mathbb{R}^{2n_s \times K}$ with its (i, k) -th entry as $\gamma_i^k = 1 - e_i^k - \alpha_i^k$, we write the dual form of (4) as follows:

$$\min_{\mathbf{A} \in \mathcal{A}} \left\{ \mu\Omega(\mathbf{A}) + \max_{\mathbf{\Gamma} \in \mathcal{M}} J(\mathbf{A}, \mathbf{\Gamma}) \right\}, \quad (8)$$

where $J(\mathbf{A}, \mathbf{\Gamma}) = -\frac{1}{2}\operatorname{tr}(\mathbf{\Gamma}'\mathbf{K}_\mathbf{A}\mathbf{\Gamma}) - \operatorname{tr}(\mathbf{E}'\mathbf{\Gamma})$, and $\mathbf{E} \in \mathbb{R}^{2n_s \times K}$ is a matrix with its (i, k) -th entry as e_i^k , $\mathcal{M} = \{\mathbf{\Gamma} | \mathbf{\Gamma}\mathbf{1}_K = \mathbf{0}_{2n_s}, \gamma_i^k \leq C(1 - e_i^k)\}$ is the feasible set of $\mathbf{\Gamma}$, and $\mathbf{K}_\mathbf{A} \in \mathbb{R}^{2n_s \times 2n_s}$ is the kernel matrix for the samples in the common space, which is defined as follows:

$$\mathbf{K}_\mathbf{A} = \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{O} & \mathbf{K}_v^s \end{bmatrix} \mathbf{A}\mathbf{A}' \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{O} & \mathbf{K}_v^s \end{bmatrix}.$$

Note that the similarities between the depth features and visual features are also integrated in $\mathbf{K}_\mathbf{A}$ by associating with the combination coefficient matrix \mathbf{A} .

4. Solution

The problem in (8) is a non-convex problem w.r.t. \mathbf{A} and $\mathbf{\Gamma}$. Therefore, we propose an alternating optimization algorithm to solve it, in which we use line search when solving for \mathbf{A} at each iteration to ensure the decrease of the objective in (8).

Specifically, given the combination coefficient matrix \mathbf{A} , the optimization problem in (8) becomes

$$\max_{\mathbf{\Gamma} \in \mathcal{M}} -\frac{1}{2}\operatorname{tr}(\mathbf{\Gamma}'\mathbf{K}_\mathbf{A}\mathbf{\Gamma}) - \operatorname{tr}(\mathbf{E}'\mathbf{\Gamma}), \quad (9)$$

which is a multi-class SVM problem, and can be solved efficiently by using the existing solver² in LIBLINEAR [11].

On the other hand, when we fix $\mathbf{\Gamma}$, the optimization problem w.r.t. \mathbf{A} can be written as

$$\min_{\mathbf{A} \in \mathcal{A}} -\frac{1}{2}\operatorname{tr}(\mathbf{A}'\tilde{\mathbf{B}}\mathbf{A}), \quad (10)$$

²Note $\mathbf{K}_\mathbf{A}$ can be treated as a linear kernel with the data matrix as $\mathbf{A}' \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{O} & \mathbf{K}_v^s \end{bmatrix} \in \mathbb{R}^{m \times 2n_s}$.

where $\tilde{\mathbf{B}} = \mu\mathbf{B} + \mathbf{G}\mathbf{G}'$ with $\mathbf{G} = \begin{bmatrix} \mathbf{K}_d & \mathbf{O} \\ \mathbf{O} & \mathbf{K}_v^s \end{bmatrix}$ $\Gamma \in \mathbb{R}^{2n_s \times K}$. It shares the similar formulation with KCCA which can be solved by using the generalized eigen-decomposition. It is worth mentioning that the matrix \mathbf{B} in our problem integrates the unlabeled samples from the target domain [see (7)], and the matrix \mathbf{G} also integrates the dual variables of the classifier f in Γ , which indicate that the target domain unlabeled samples and the classifier learnt at the previous iteration also contribute to the learning of the common space in (10).

4.1. Line Search when Solving for \mathbf{A}

Due to the non-convexity of (8), the optimal solution \mathbf{A}^* from (10) cannot ensure the objective of (8) decreases. Therefore, at the t -th iteration, we need to search for a feasible $\mathbf{A}_t \in \mathcal{A}$ between the optimal solution \mathbf{A}^* and the solution \mathbf{A}_{t-1} at the previous iteration. In the following, we first briefly introduce how to solve for the optimal solution \mathbf{A}^* to the problem in (10), then we present our line search method.

4.1.1 Solving for \mathbf{A}^*

The problem in (10) can be reformulated as a generalized eigen-decomposition problem [19] as follows:

$$\tilde{\mathbf{B}}\mathbf{v} = \sigma\mathbf{D}\mathbf{v}, \quad (11)$$

where $\mathbf{D} = \begin{bmatrix} \mathbf{K}_d\mathbf{K}_d & \mathbf{O} \\ \mathbf{O} & \mathbf{K}_v^s\mathbf{K}_v^s \end{bmatrix} \in \mathbb{R}^{2n_s \times 2n_s}$, \mathbf{v} is the eigenvector and σ is the corresponding eigenvalue. The optimal solution to (10) is obtained by combining the m leading eigenvectors corresponding to the largest eigenvalues.

To solve (11), we first perform the incomplete Cholesky decomposition on \mathbf{D} as $\mathbf{D} = \mathbf{C}'\mathbf{C}$ as suggested in [19]. Then, we can rewrite (11) as a standard eigen-decomposition problem, $(\mathbf{C}^{-1})'\tilde{\mathbf{B}}\mathbf{C}^{-1}\mathbf{u} = \sigma\mathbf{u}$, where $\mathbf{u} = \mathbf{C}\mathbf{v}$ is the eigenvector of $(\mathbf{C}^{-1})'\tilde{\mathbf{B}}\mathbf{C}^{-1}$.

Let us denote the eigen-decomposition of $(\mathbf{C}^{-1})'\tilde{\mathbf{B}}\mathbf{C}^{-1}$ as $\mathbf{U}\Sigma\mathbf{U}'$ where \mathbf{U} is the eigenvectors and Σ is a diagonal matrix with the diagonal entries being the eigenvalues. We also define $\tilde{\mathbf{U}}^* \in \mathbb{R}^{2n_s \times m}$ as the matrix containing the m leading eigenvectors in \mathbf{U} corresponding to the largest eigenvalues. Then, the optimal solution to (10) can be obtained as $\mathbf{A}^* = \mathbf{C}^{-1}\tilde{\mathbf{U}}^*$.

4.1.2 Line Search for \mathbf{A}_t

The major challenge in line search is to ensure that the solution satisfies the constraint $\mathbf{A}_t \in \mathcal{A}$. Note that the feasible \mathbf{A} is given in the form of $\mathbf{A} = \mathbf{C}^{-1}\tilde{\mathbf{U}}$ where $\tilde{\mathbf{U}} \in \mathbb{R}^{2n_s \times m}$ is an orthogonal matrix. Let $\mathcal{S} = \text{span}(\tilde{\mathbf{U}})$ be the subspace spanned by $\tilde{\mathbf{U}}$. Obviously, all basis matrices of the subspace \mathcal{S} are feasible, and produce equal objective value in (8).

Algorithm 1 The algorithm for our DA-M2S

Input: The label vector \mathbf{y} , and the kernel matrices \mathbf{K}_d , \mathbf{K}_v^s and \mathbf{K}_v^{st} as defined in Section 3.3.

- 1: Initialize $t = 0$.
- 2: Solve (8) with only the first term [*i.e.*, $\Omega(\mathbf{A})$] to obtain an initial $\mathbf{A}_0 = \mathbf{C}^{-1}\tilde{\mathbf{U}}_0$.
- 3: Solve for Γ_0 in (9) based on \mathbf{A}_0 by using the existing solver in LIBLINEAR [11].
- 4: **repeat**
- 5: Set $t = t + 1$.
- 6: Solve the problem in (11) to obtain $\mathbf{A}^* = \mathbf{C}^{-1}\tilde{\mathbf{U}}^*$.
- 7: Find the optimal basis $\tilde{\mathbf{U}}(\tau)$ in the geodesic path such that the objective in (8) is minimized.
- 8: Solve Γ based on $\mathbf{A}(\tau) = \mathbf{C}^{-1}\tilde{\mathbf{U}}(\tau)$ by using the existing solver in LIBLINEAR [11].
- 9: Set $\mathbf{A}_t = \mathbf{A}(\tau)$, $\Gamma_t = \Gamma$.
- 10: **until** The objective in (8) converges.

Output: $\mathbf{A} = \mathbf{A}_t$ and $\Gamma = \Gamma_t$.

Let us denote $\mathbf{A}_{t-1} = \mathbf{C}^{-1}\tilde{\mathbf{U}}_{t-1}$ as the solution to (10) at the previous iteration, and $\mathbf{A}^* = \mathbf{C}^{-1}\tilde{\mathbf{U}}^*$ as the optimal solution at this iteration, respectively. Recall that all $2n_s$ -by- m subspaces reside on a Grassmann manifold, so our line search problem becomes to find a new subspace \mathcal{S}_t along the geodesic path between two subspaces $\mathcal{S}_{t-1} = \text{span}(\tilde{\mathbf{U}}_{t-1})$ and $\mathcal{S}^* = \text{span}(\tilde{\mathbf{U}}^*)$, whose basis $\tilde{\mathbf{U}}_t$ makes the objective of (8) decrease.

As shown in [16], the geodesic path between \mathcal{S}_{t-1} and \mathcal{S}^* can be represented as $\mathcal{S}(\tau)$ with $0 \leq \tau \leq 1$, and we have $\mathcal{S}(0) = \mathcal{S}_{t-1}$ and $\mathcal{S}(1) = \mathcal{S}^*$. Then, we perform line search using different τ 's to find a subspace $\mathcal{S}(\tau) = \text{span}(\tilde{\mathbf{U}}(\tau))$ according to the method in [16] such that the projection matrix $\mathbf{A}(\tau) = \mathbf{C}^{-1}\tilde{\mathbf{U}}(\tau)$ leads to the minimal objective in (8).

Finally, the details of our algorithm for solving (8) are listed in Algorithm 1. We first initialize the combination coefficient matrix \mathbf{A} by solving (8) with only the first term [*i.e.*, $\Omega(\mathbf{A})$]. Then, we iteratively solve (9) by using the existing solver in [11] and solve the eigenvalue decomposition problem in (11). Then we perform the line search between \mathcal{S}_{t-1} and \mathcal{S}^* to find a better subspace \mathcal{S}_t , such that $\mathbf{A}_t = \mathbf{C}^{-1}\tilde{\mathbf{U}}_t$ leads to the minimal objective in (8). The above procedure is repeated until the objective value no longer decreases. In our experiments, the algorithm converges after about 10 iterations.

By using the learnt $\mathbf{A} = [\mathbf{A}'_d, \mathbf{A}'_v]'$, any test data \mathbf{x}^t from the target domain can be projected into the common space as $\tilde{\mathbf{x}}^t = \mathbf{P}'_v\phi(\mathbf{x}^t) = \mathbf{A}'_v\Phi'_s\phi(\mathbf{x}^t)$. Then we can use the learnt classifier to predict its class label. The final classifier is given by $f(\tilde{\mathbf{x}}^t) = \text{argmax}_{k=1, \dots, K} \mathbf{w}'_k\tilde{\mathbf{x}}^t$, where each $\mathbf{w}_k = \sum_{i=1}^{2n_s} \gamma_i^k \tilde{\mathbf{x}}_i$, in which γ_i^k is the (i, k) -th entry of Γ from Algorithm 1, and $\tilde{\mathbf{x}}_i$ is the i -th training sample from the source domain in the common space.

5. Experiments

In this section, we evaluate the effectiveness of our DA-M2S for object recognition and gender recognition.

5.1. Baseline Approaches

To the best of our knowledge, there is no previous work specifically designed for recognizing RGB images by learning from RGB-D data. Thus, we extend a broad range of existing works as the baselines for fair comparison, which can be divided into four categories as follows:

Naive Approach: The naive approach SVM_A is trained by using the visual features in the source domain without considering the domain distribution mismatch and exploiting the additional depth features.

Multi-view Learning: The multi-view learning approaches include KCCA [19] and SVM2K [12], in which the two-view data in the source domain are used for training. For SVM2K, two classifiers are trained by using the two-view data in the source domain, and we use the one based on visual features to predict the target domain visual features. For KCCA, we train two SVM classifiers by using the projected depth and visual features in the common space and the decision values of target domain samples based on the projected visual features are equally fused for prediction.

Learning Using Privileged Information: For the learning approaches using privileged information such as SVM+ [29] and RankTransfer (RT) [26], we use the additional depth features in the source domain as privileged information for learning the visual feature based classifier.

Unsupervised Domain Adaptation: The domain adaptation approaches include KMM [20], DAM [7], SGF [16], TCA [25], Landmark (LMK) [14], Subspace Alignment (SA) [13], and Domain Invariant Projection (DIP) [1], for which the visual features from both domains are used for training the classifiers and we predict target domain data based on the visual features. We do not compare our DA-M2S with GFK [15], because the subsequent work LMK from the same group is shown to be better (see [14]).

Note that the semi-supervised multi-view learning methods [27] and the multi-view domain adaptation approaches [31] cannot be applied for our problem, because we only have single view of features for the samples in the target domain. Moreover, the heterogeneous domain adaptation (HDA) methods [22, 24] also cannot be used because the labeled samples in the target domain are required in these HDA methods.

5.2. Object Recognition

Experimental Setup: We evaluate our proposed DA-M2S for object recognition by using the RGB-D Object dataset [23] as the source domain and the Caltech-256 dataset [18] as the target domain. The RGB-D Object dataset contains the color and depth images of different objects from 51 categories. The Caltech-256 dataset contains

only color images. In this work, we use the 10 common categories³ between the two datasets. As the RGB-D Object dataset is recorded in the form of video sequences, we uniformly sample the frames with an interval of two seconds, leading to a total number of 2059 training images. All the target domain samples are also used as unlabeled data in the training stage for the baseline domain adaptation methods and our DA-M2S.

We use kernel descriptors (KDES) features [2] in this work, which have shown promising recognition results on this dataset. Specifically, we extract Gradient KDES and LBP KDES features from each RGB/depth image by using the software⁴ provided by the authors. Then, we follow [2] to aggregate the kernel descriptors into object-level features, in which we set the vocabulary size as 1000 and use three level of pyramids (*i.e.*, 1x1, 2x2, 3x3). The object level features respectively constructed from the Gradient KDES and LBP KDES features are concatenated into one feature vector for each RGB/depth image. Note the features for RGB and depth images are different, because we use different vocabularies. We use the same method to extract the visual features for the RGB images in the target domain.

We use the multiclass classification accuracy as the evaluation criterion, which is the average of the accuracies over all the classes. For all the kernel-based approaches, Gaussian kernel is used as the default kernel with the bandwidth parameter set as the mean of the distances between any two samples. We use the default tradeoff parameter $C = 1$ for all methods. Moreover, for our DA-M2S, we empirically fix the parameters as $\mu = 0.1n_s$ and $\lambda = 10^4$. For all other methods, we tune their parameters based on the test dataset and report their best results from the optimal parameters.

Experimental Results: The results of all methods are reported in Table 1. From this table, we observe that our newly proposed DA-M2S outperforms all other baseline methods. It demonstrates the effectiveness of our DA-M2S by employing the additional depth features in the source domain and simultaneously reducing the domain distribution mismatch between the source and target domains.

Specifically, by utilizing the additional depth features, the multi-view learning approaches KCCA and SVM2K as well as the privileged learning approach SVM+ achieve better results when compared with the naive approach SVM_A. RT is worse than SVM_A, possibly because it is based on RankSVM, which is designed for the ranking task rather than the classification task. Among these methods, SVM2K achieves the best result, as it can more effectively exploit depth information by learning two classifiers for both vi-

³The 10 common categories between the two datasets are calculator, cereal_box, coffee_mug, keyboard, flashlight, lightbulb, mushroom, ball, soda_can, tomato.

⁴The code is available at http://www.cs.washington.edu/ai/Mobile_Robots/projects/kdes/.

Table 1. Comparison of recognition accuracies (%) for object recognition. The RGB-D object dataset is used as the source domain and the Caltech-256 dataset is used as the target domain.

SVM_A	SVM+	RT	KCCA	SVM2K	KMM	DAM	SGF	LMK	TCA	SA	DIP	DA-M2S
18.19	18.59	17.16	18.23	20.83	18.10	18.19	19.25	19.45	25.07	21.09	25.47	30.06

Table 2. Recognition accuracies (%) of domain adaptation methods for object recognition using the feature representations in the common space learnt by KCCA.

KMM-C	DAM-C	SGF-C	LMK-C	TCA-C	SA-C	DIP-C
18.47	17.50	19.38	19.72	27.48	21.25	24.76

sual and depth features. Nevertheless, all these methods do not cope with the distribution mismatch between the source and target domains, thus they are much worse than our DA-M2S.

The domain adaptation methods SGF, LMK, TCA, SA and DIP are also better than SVM_A, which shows it is beneficial to reduce the domain distribution mismatch between the source and target domains by using these methods. When compared with SVM_A, KMM and DAM are only comparable or even worse, possibly because both approaches cannot effectively handle the significant domain distribution mismatch in this application. Moreover, our proposed DA-M2S outperforms all those methods by additionally exploiting the depth features in the source domain.

KCCA + UDA Approaches: We additionally report more results for object recognition by using the domain adaptation methods in the common space learnt by using KCCA, which are referred to as KMM-C, DAM-C, SGF-C, LMK-C, TCA-C, SA-C and DIP-C. Specifically, we first learn the projection matrices by using KCCA and project both visual and depth features into the learnt common space. Then, we apply these domain adaptation methods using the projected depth and visual features in the common space to learn two classifiers and equally fuse the decision values of target samples from two classifiers using the projected visual features. The results are shown in Table 2. We observe that most UDA methods are improved by utilizing the additional depth features, when compared with their corresponding results in Table 1. Our DA-M2S still outperforms those baselines, which again demonstrates it is beneficial to simultaneously employ the additional depth features and reduce the domain distribution mismatch.

5.3. Gender Recognition

Experimental Setup: We also evaluate our DA-M2S for gender recognition by using the RGB-D face dataset EURECOM [21] as the source domain, and the RGB dataset Labeled Faces in the Wild-a (LFW-a) [30] as the target domain. The EURECOM dataset [21] contains the RGB and depth images captured by using the Kinect sensor. There are totally 728 pairs of RGB and depth images from 196 females and 532 males. The LFW-a dataset contains a total number of 13, 144 images from 2, 960 females and 10, 184 males, which are collected under the uncontrolled environ-

ment.

All face images in two datasets are aligned and cropped to a fixed size of 120×105 pixels according to the positions of two eyes. The images in the LFW-a dataset have been aligned according to the eye-positions (see [30] for details). For the EURECOM dataset, the manually annotated eye-positions are provided, and the images with only a single eye-position (*i.e.* the profile face images) are not included in our experiments as suggested in [21]. Then, we uniformly divide each face image into 8×7 non-overlapping subregions with the size of each subregion being 15×15 pixels. After that, we extract the Gradient-LBP feature [21] from each subregion for both color and depth images, as it has been shown to be effective for gender recognition [21]. Finally, for each face image, the Gradient-LBP features from all 56 subregions are concatenated into a single feature vector. The same Gradient-LBP features are also extracted for the RGB images in the LFW-a dataset.

Because there are much more male faces than female faces in the EURECOM dataset, we randomly sample 196 male faces from this dataset to balance the positive and negative training samples. We also randomly sample 3,000 samples from a large number of target samples as unlabeled data for the baseline domain adaptation methods and our DA-M2S. The mean recognition accuracy and the standard deviation over ten rounds of experiments are reported for all the methods for gender recognition. The rest of the settings are the same as in object recognition.

Experimental Results: The results of SVM+ and RT are 64.24 ± 1.66 and 64.22 ± 1.76 respectively, and the results of all other methods are shown in Table 3. We have similar observations as in the object recognition. While other methods generally outperform SVM_A by exploiting the additional depth features or reducing the domain distribution mismatch, our DA-M2S outperforms all the baseline methods by simultaneously considering both aspects in one formulation.

We also observe that SVM2K is much better than SVM_A, demonstrating that it is beneficial to use the additional depth features in the source domain for this application. However, KCCA, SVM+ and RT are not as effective as SVM2K, and they are only comparable to or even worse than SVM_A. Most of the domain adaptation approaches such as SGF, LMK, TCA, SA and DIP are also better than SVM_A. However, KMM and DAM are only comparable or even worse than SVM_A in this application.

Table 3. Comparison of recognition accuracies (mean \pm standard deviation %) for gender recognition. The result in bold is significantly better than the others judged by the significant test with a significance level of 0.05.

SVM_A	KCCA	SVM2K	KMM	DAM	SGF	LMK	TCA	SA	DIP	DA-M2S
64.22 \pm 1.6	63.60 \pm 1.34	67.33 \pm 1.92	64.25 \pm 1.43	63.91 \pm 1.57	67.22 \pm 1.38	65.02 \pm 1.55	65.24 \pm 0.88	67.38 \pm 1.39	64.84 \pm 4.80	68.44\pm1.44

Table 4. Comparison of recognition accuracies (%) between DA-M2S and its special cases.

	Gender	Object
DA-M2S (w/o depth)	67.57 \pm 1.68	28.45
DA-M2S (init)	67.39 \pm 1.02	28.43
DA-M2S	68.44\pm1.44	30.06

5.4. Analysis on DA-M2S

For a better understanding of our DA-M2S, we investigate two special cases of our DA-M2S.

In the first special case, denoted as DA-M2S (w/o depth), we do not consider depth information. Namely, we remove $\Omega^{kcca}(\mathbf{P}_d, \mathbf{P}_v)$ as well as the second term of $\Omega^{mmd}(\mathbf{P}_d, \mathbf{P}_v)$ in the regularizer $\Omega(\mathbf{P}_d, \mathbf{P}_v)$ in (4). Note $\Omega(\mathbf{P}_d, \mathbf{P}_v)$ becomes $\Omega(\mathbf{P}_v)$ with $\mathbf{P}_v \in \mathcal{P}$, where $\mathcal{P} = \{\mathbf{P}_v | \mathbf{P}'_v \Phi_s \Phi'_s \mathbf{P}_v = \mathbf{I}_m\}$. We also remove the n_s constraints related to the depth features in (5). As shown in Table 4, the results are worse than our DA-M2S, which shows it is beneficial to exploit the additional depth features for learning a more robust classifier.

We also report the results of our DA-M2S at the first iteration, which is denoted as DA-M2S (init). As show in Table 4, its performances are also worse than DA-M2S, which demonstrates the effectiveness of our alternating optimization technique for iteratively learning the classifier and the projection matrices.

6. Conclusions

In this paper, we have proposed a new framework for recognizing RGB images by learning from a set of labeled RGB-D data. We formulate this task as a new UDA problem, in which we have both visual and depth features in the source domain, while we only have the visual features in the target domain. An effective method called DA-M2S is proposed to solve this problem by taking advantage of the additional depth features in the source domain and simultaneously reducing the distribution mismatch between the source and target domains. Comprehensive experiments for object recognition and gender recognition have clearly demonstrated the effectiveness of our proposed DA-M2S approach for recognizing RGB images by learning from RGB-D data.

Acknowledgement. This work is supported by the Singapore MoE Tier 2 Grant (ARC42/13).

References

- [1] M. Baktashmotlagh, M. Harandi, and M. S. Brian Lovell. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.
- [2] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *IROS*, 2011.
- [3] L. Chen, L. Duan, and D. Xu. Event recognition in videos by learning from heterogeneous web sources. In *CVPR*, pages 2666–2673, 2013.
- [4] K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *JMLR*, 9:1757–1774, 2008.
- [5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.
- [6] L. Duan, I. W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *T-PAMI*, 34(3):465–479, March 2012.
- [7] L. Duan, I. W. Tsang, D. Xu, and T. Chua. Domain adaptation from multiple sources: A domain-dependent regularization approach. *T-NNLS*, 23(3):504–518, 2012.
- [8] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, pages 1338–1345, 2012.
- [9] L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, pages 711–718, 2012.
- [10] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *T-PAMI*, 34(9):1667–1680, September 2012.
- [11] R. Fan, K. Chang, and C. Hsieh. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.
- [12] J. D. R. Farquhar, H. Meng, S. Szedmak, D. R. Hardoon, and J. Shawe-taylor. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2006.
- [13] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.
- [14] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.
- [15] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.
- [16] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.
- [17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.
- [18] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.
- [19] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computing*, 16(12):2639–2664, 2004.
- [20] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.
- [21] T. Huynh, R. Min, and J.-L. Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *ACCV Workshop*, 2012.
- [22] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.
- [23] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. *ICRA*, 2011.
- [24] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *T-PAMI*, 2013.
- [25] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *T-NN*, 22(2):199–210, 2009.
- [26] V. Sharmanska, I. Austria, N. Quadrianto, and C. Lampert. Learning to rank using privileged information. In *ICCV*, 2013.
- [27] V. Sindhwani, P. Niyogi, and M. Belkin. A coregularization approach to semi-supervised learning with multiple views. In *ICML*, 2005.
- [28] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.
- [29] V. Vapnik and A. Vashist. A new learning paradigm: learning using privileged information. *Neural networks*, 22(5-6):544–57, 2009.
- [30] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *T-PAMI*, 33(10):1978–1990, 2011.
- [31] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence. Multi-view transfer learning with a large margin approach. In *KDD*, 2011.