This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

Visual Recognition in RGB Images and Videos by Learning from RGB-D Data

Wen Li, Lin Chen, Dong Xu, *Senior Member, IEEE*, and Luc Van Gool

Abstract—In this work, we propose a framework for recognizing RGB images or videos by learning from RGB-D training data that contains additional depth information. 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 handle the domain distribution mismatch, we propose to learn an optimal projection matrix to map the samples from both domains into a common subspace such that the domain distribution mismatch can be reduced. Such projection matrix can be effectively optimized by exploiting different strategies. Moreover, we also use different ways to utilize the additional depth features. To simultaneously cope with the above two issues, we formulate a unified learning framework called domain adaptation from multi-view to single-view (DAM2S). By defining various forms of regularizers in our DAM2S framework, different strategies can be readily incorporated to learn robust SVM classifiers for classifying the target samples, and three methods are developed under our DAM2S framework. We conduct comprehensive experiments for object recognition, cross-dataset and cross-view action recognition, which demonstrate the effectiveness of our proposed methods for recognizing RGB images and videos by learning from RGB-D data.

Index Terms—domain adaptation, object recognition, human action recognition.

1 Introduction

With the advance of RGB-D equipments (*e.g.*, Kinect sensors) for capturing depth information, there is an increasing research interest in developing new technologies using depth images and videos for various visual recognition tasks (*e.g.*, object recognition, face recognition, and action recognition). While the effectiveness of depth information has been demonstrated in recent works, those techniques cannot be applied to most ordinary visual recognition applications, in which images and videos are captured by conventional RGB cameras (*e.g.*, smartphones).

To this end, we propose a new framework for recognizing RGB images and videos captured with the conventional cameras by leveraging a set of labeled RGB-D data. Our work is based on the observation that several labeled RGB-D datasets [1], [2], [3] were recently released for various vision recognition tasks as well as the recent progress on learning using privileged information [4], [5], which shows the additional features (*i.e.*, privileged information) that are not available at the testing stage are still useful for many classification tasks. In the context of our work, depth information is usually more robust to illumination changes and complex backgrounds in the visual recognition tasks, compared with the RGB images/videos, and thus providing complementary information for recognizing RGB images and videos.

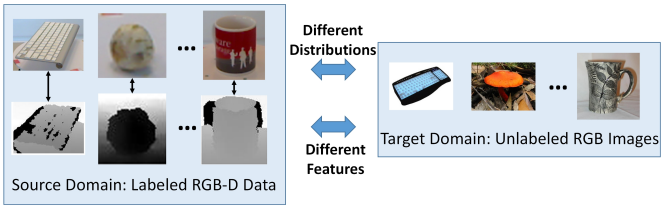

Fig. 1. Image recognition in RGB images by learning from RGB-D data: we have both RGB images and depth images in the source domain, and only RGB images in the target domain.

Another issue is that the RGB testing data and the RGB-D training data are captured with different equipments, which leads to a domain distribution mismatch between the training and test data. This is also known as the dataset bias problem [6]. When one dataset is used for training and another dataset is used for testing, the performance of most existing visual recognition methods will be degraded significantly because the feature distributions of samples from different datasets may have very different statistical properties. To cope with the considerable variation in feature distributions, new domain adaptation methods were recently developed in both machine learning and computer vision communities [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18].

In this work, we formulate our task as a new unsupervised domain adaptation (UDA) problem, in which we have single-view visual features extracted from the RGB images or videos in the target domain (the domain of test data) while we have both visual features and depth features in the source domain (see Fig 1). We

- *W. Li and L. Van Gool are with the Computer Vision Laboratory, ETH Zürich, Switzerland.*
  E-mail: {liwen,vangool}@vision.ee.ethz.ch
- *L. Chen is with the Amazon, 500 Boren Avenue North, Seattle, WA 98109.*
  E-mail: gggchenlin@gmail.com
- *D. Xu is with the School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia.*
  e-mail: dongxudongxu@gmail.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

propose a unified framework called Domain Adaptation from Multi-view to Single-view (DAM2S), in which we simultaneously address the domain distribution mismatch between the source and target domains and also take advantage of the additional depth features in the source domain to learn robust classifiers for classifying the target RGB images or videos.

In our preliminary work [19], we have proposed an approach for recognizing RGB images by exploiting RGB-D images from the source domain, in which we simultaneously reduce the domain distribution mismatch between two domains by minimizing the Maximum Mean Discrepancy (MMD) criterion [20], and map the samples with the visual and depth features into a common subspace by maximizing the correlation of two types of features in the common subspace. In this work, we show that under our newly proposed DAM2S framework, more strategies can be employed to effectively cope with those two issues by defining different forms of regularizers. In particular, to address the domain adaptation issue, we propose to learn an optimal projection matrix to map the samples from both domains into a common subspace such that the domain distribution mismatch can be reduced. Such projection matrices can be effectively optimized by exploiting different strategies, such as reducing Maximum Mean Discrepancy (MMD) or aligning the source and target subspaces. Moreover, to effectively utilize the additional depth features, we can also employ different strategies, by either maximizing the correlation between different types of features in the common subspace, or preserving the consistency of the classifiers from different features. Accordingly, we develop three methods under our DAM2S framework for effectively recognize RGB images and videos in the target domain by exploiting the RGB-D data in the source domain, in which the approach proposed in our previous work [19] can be regarded as an example under our newly proposed framework.

We conduct comprehensive experiments on different visual recognition tasks to evaluate our proposed DAM2S methods. Besides the object recognition and gender recognition tasks in [19], we also conduct additional experiments for cross-dataset human action recognition and cross-view human action recognition in RGB videos by exploiting a set of RGB-D videos. We demonstrate that the proposed methods under our DAM2S framework outperform the state-of-the-art methods including the existing UDA methods, multi-view learning methods and learning using privileged information methods.

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 [7], [21], [9], [11], [15], [16], [17], [18], [12], [22]. In particular, the unsupervised domain adaptation (UDA) methods assume there is no labeled data in the target domain. Different strategies have been proposed to reduce domain distribution mismatch including sample reweighting, feature transformation, classifier adaptation, *etc*. However, those UDA methods assume the source domain samples share the same feature representation with the target domain samples, so it is unclear how to effectively utilize the additional depth features in the source domain.

Recently, heterogeneous domain adaptation (HDA) methods [23], [10] 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 [23], [10], 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 [24] and the recent work called multi-domain adaptation from heterogeneous sources (MDA-HS) [25]. In [24], all the samples in the source and target domains have multiple types of features, while in [25] the samples from the target domain have all types of features from all source domains. In contrast, we only have single-view features in the target domain. Our work is different from existing multi-domain adaptation methods [12], [26], because we have additional depth features in the source domain, and we focus on the single source domain setting.

Learning using Multiple Features and Privileged Information: Our work is also related to multiple-view learning, where the training data consists of multiple views of features. One of the representative works is canonical correlation analysis (CCA) as well as its kernel variant kernel CCA (KCCA) [27], in which a common subspace is learnt to maximize the correlations between different views of features. When the labels of training samples are available, the consistency criterion [28] is commonly used in the multiple view learning methods, which assumes the classifiers from different views should have consistent predictions. Most multi-view learning works were proposed for the semi-supervised learning scenario, such as co-training [29], co-labeling [30] and so on. However, those multi-view learning works assume that the test data also consists of multiple views of features, which is different from our problem, where we only have one type of features in the target domain.

Recently, several works [4], [5], [31], [32], [33], [34], [35] were proposed for learning using privileged information, in which the training data contains additional features (*i.e.*, privileged information) that are not available at the testing stage. However, these works [4], [5], [31], [32] assume that the training and test samples come from the same data distribution. Recently, Hoffman *et al.* proposed to detect object in depth testing data by using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

RGB-D training data in [36], while a network hallucination approach was proposed for object detection in RGB images by leveraging depth information in the training data in [37]. While their work specifically focuses on the object detection task, it would be an interesting research topic to combine the features extracted from those deep learning methods and the newly proposed classification methods DAM2S to improve the classification results.

3 Domain Adaptation with Additional Features

For ease of presentation, we denote a vector/matrix by a lowercase/uppercase letter in bold. The transpose of a vector/matrix is denoted by the superscript $'$. We define $\mathbf{I}_n$ as the $n \times n$ identity matrix. We also define $\mathbf{1}_n \in \mathbb{R}^n$ as the $n \times 1$ column vectors of all ones. For simplicity, we also use $\mathbf{I}$ and $\mathbf{1}$ when the dimension is obvious.

3.1 Framework

We extract the visual features and depth features from the RGB images/videos and depth images/videos, respectively. The source domain samples can be represented as $\{(\mathbf{z}_i, \mathbf{x}_i^s)|_{i=1}^{n_s}\}$ where $\mathbf{z}_i$ and $\mathbf{x}_i^s$ are respectively the depth feature and the visual feature, and n_s is the total number of samples in the source domain. We also define $l_{i,k} \in \{+1, -1\}$ as the label of the i-th sample corresponding to the k-th class, where $k = 1, \ldots, K$, and K is the number of classes. Namely, $l_{i,k} = 1$ means the i-th sample belongs to the k-th class, and $l_{i,k} = -1$, otherwise. 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.

Subspace learning has been shown to be robust to various visual variations [38]. To handle the distribution mismatch, we learn a common subspace (parameterized by a matrix $\mathbf{P}$) for the visual features from two domains, such that the domain distribution mismatch can be reduced in this common subspace. A classifier f^v is learnt in this common subspace for predicting the test samples. Moreover, to incorporate the depth features of training samples, we also learn an auxiliary classifier f^d, which is used to help the learning of f^v. Based on the empirical risk minimization (ERM) principle, we formulate a unified learning scheme as follows,

$$\min r(f^v, f^d) + C\ell(f^v, f^d) + \mu\,\Omega(\mathbf{P}) + \lambda\Delta(\cdot,\cdot), \qquad (1)$$

where $r(f^v, f^d)$ is the regularizer to control the complexity of the classifiers f^v and f^d, $\Omega(\mathbf{P})$ is the regularizer term on the parameter matrix $\mathbf{P}$ that decides the common subspace, $\ell(f^v, f^d)$ is the loss term on the training samples, and C, μ and λ are the tradeoff parameters. The last term $\Delta(\cdot,\cdot)$ is a regularizer term to associate the depth features and visual features, such that f^d can be used to help the learning of f^v. For example, we can enforce the decision values of two classifiers f^v and f^d to be consistent on the training samples, so the last term can be written as $\Delta(f^v, f^d)$. We use a general form $\Delta(\cdot,\cdot)$ in (1) in order to exploit other strategies.

With the above framework, we employ different strategies to cope with the domain distribution mismatch, which leads to different regularizers $\Omega(\mathbf{P})$. To further utilize the depth features, we also define different $\Delta(\cdot,\cdot)$ for associating the visual and depth features. We will discuss the detailed forms of those terms, and develop the corresponding algorithms below.

3.2 Learning Projection for Domain Adaptation

In the following subsections, we investigate two strategies for learning the projection matrix $\mathbf{P}$ such that the data distribution mismatch between the source and target domains based on the visual features can be reduced.

3.2.1 Reducing the Maximum Mean Discrepancy

Our first strategy is to minimize the Maximum Mean Discrepancy (MMD) [20] criterion, which is widely used to measure the distribution difference between the data sampled from two datasets.

Let us denote $\phi(\cdot) : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^{m_v}$ as the nonlinear feature mapping induced by a characteristic kernel $\mathbf{K}_v^s = \mathbf{\Phi}_s'\mathbf{\Phi}_s$, where $\mathbf{\Phi}_s = [\phi(\mathbf{x}_1^s), \ldots, \phi(\mathbf{x}_{n_s}^s)] \in \mathbb{R}^{m_v \times n_s}$ is the data matrix of source samples in the nonlinear feature space, D_v is the dimensionality of the original features, m_v is the dimensionality of the nonlinear features that is usually unknown and can be infinite. The MMD between two domains can be written as, $\text{MMD} \doteq \|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_i^s) - \frac{1}{n_t}\sum_{i=1}^{n_t}\phi(\mathbf{x}_i^t)\|^2$.

In this work, we aim to learn a projection matrix $\mathbf{P} \in \mathbb{R}^{m_v \times m}$ such that the MMD between the source and target domains is reduced after projecting the samples from two domains into a m-dimensional common subspace with this projection matrix [1]. The MMD criterion has been used in the literature for minimizing the distribution mismatch when learning latent features in [39], [40]. Motivated by those works, we define the regularizer $\Omega(\mathbf{P})$ in (1) as follows,

$$\Omega^{mmd}(\mathbf{P}) = \frac{1}{2}\|\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{P}'\phi(\mathbf{x}_i^s) - \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{P}'\phi(\mathbf{x}_i^t)\|^2. \qquad (2)$$

3.2.2 Aligning two Subspaces

Recently, another subspace based approach called subspace alignment (SA) was proposed in [17] for domain adaptation, in which they aim to align the subspaces of the source and target domains. Specifically, let us denote $\mathbf{S}_s \in \mathbb{R}^{D_v \times m}$ as the projection matrix, which projects the source domain samples into the source subspace, where m is the dimension of the source subspace. Such a projection matrix can be obtained by using the subspace learning methods such as principle components analysis (PCA). Similarly, we denote the projection matrix of

1. While m_v is unknown and can be infinite, in the implementation, we solve the matrix $\mathbf{P}$ in the kernel space without explicitly knowing the dimension m_v (see [19]).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

target subspace as $\mathbf{S}_t \in \mathbb{R}^{D_v \times m}$. Then, the objective of SA can be formulated as,

$$\min_{\mathbf{T}} \|\mathbf{S}_s\mathbf{T} - \mathbf{S}_t\|_F^2, \tag{3}$$

where $\mathbf{T} \in \mathbb{R}^{m\times m}$ is a transformation matrix aligning the source subspace to the target subspace. The above problem has the closed form solution, namely, $\mathbf{T} = \mathbf{S}_s'\mathbf{S}_t$. After obtaining the transformation matrix, the source samples can be transformed into the target subspace for learning the classifier for predicting the samples in the target domain.

However, the SA method was originally developed for the linear case. Moreover, it does not consider the labels of training samples for learning a discriminative transformation. Finally, the additional depth features cannot be used, either. In the following, we first extend SA into kernel SA (KSA), and propose a new regularizer based on kernel SA, which can be readily incorporated into our DAM2S framework for learning robust classifiers by additionally considering the depth features.

Given a kernel matrix $\mathbf{K} = \mathbf{\Phi}'\mathbf{\Phi}$ with $\mathbf{\Phi}$ being the nonlinear features induced by $\mathbf{K}$, we can obtain the projection matrix $\mathbf{S} = \mathbf{\Phi}\mathbf{A}$ by using kernel PCA (KPCA), where $\mathbf{A}$ is a coefficient matrix that satisfies $\mathbf{A}'\mathbf{\Phi}'\mathbf{\Phi}\mathbf{A} = \mathbf{I}$. Recall the kernel matrix based on the source domain visual features is denoted as $\mathbf{K}_v^s = \mathbf{\Phi}_s'\mathbf{\Phi}_s$, and we also denote $\mathbf{K}_v^t = \mathbf{\Phi}_t'\mathbf{\Phi}_t$ as the kernel matrix on the target domain visual features, where $\mathbf{\Phi}_t = [\phi(\mathbf{x}_1^t), \ldots, \phi(\mathbf{x}_{n_t}^t)] \in \mathbb{R}^{m_v\times n_t}$. Then, we can obtain two projection matrices by using KPCA, $\mathbf{S}_s = \mathbf{\Phi}_s\tilde{\mathbf{A}}_s$ and $\mathbf{S}_t = \mathbf{\Phi}_t\tilde{\mathbf{A}}_t$, where $\tilde{\mathbf{A}}_s \in \mathbb{R}^{n_s\times m}$ and $\tilde{\mathbf{A}}_t \in \mathbb{R}^{n_t\times m}$ are two coefficient matrices. Then we formulate the kernel SA problem as,

$$\min_{\mathbf{T}} \|\mathbf{\Phi}_s\tilde{\mathbf{A}}_s\mathbf{T} - \mathbf{\Phi}_t\tilde{\mathbf{A}}_t\|_F^2, \tag{4}$$

which has the closed form solution that $\mathbf{T} = \tilde{\mathbf{A}}_s'\mathbf{\Phi}_s'\mathbf{\Phi_t}\tilde{\mathbf{A}}_t$. Finally, we introduce a new regularizer for the objective in (1) as follows,

$$\Omega^{ksa}(\mathbf{P}) = \frac{1}{2}\|\mathbf{P} - \mathbf{T}\|_F^2 \tag{5}$$

where $\mathbf{T}$ is pre-learnt by using kernel SA as described above. Note we use the same symbol $\mathbf{P}$ in (5) as the one in the MMD based regularizer $\Omega^{mmd}(\mathbf{P})$ for ease of presentation, their physical meanings are different. The matrix $\mathbf{P}$ in the MMD based regularizer in (2) is the projection matrix, which is used to project the visual features of source and target domains into the common subspace, while the matrix $\mathbf{P}$ in the kernel SA based regularizer is the transformation matrix, which is used to transform the visual features in the source subspace to the target subspace.

3.3 Incorporating Depth Information

In this section, we investigate how to effectively incorporate the depth features to learn more robust classifiers. Our methods are motivated by the two-view learning works, in which different views of features help each other to learn robust classifiers. In the follows, we consider two different strategies by maximizing the feature correlation and retaining the classifier consistency.

3.3.1 Maximizing Feature Correlation

Canonical correlation analysis (CCA) is one pioneering work in two-view learning, which aims to learn two projection matrices to map the training samples with different features into a common subspace, such that the feature correlation can be maximized. Kernel canonical correlation analysis (KCCA) is an extension of CCA by applying the kernel trick. Formally, let us denote $\psi(\cdot) : \mathbb{R}^{D_d} \rightarrow \mathbb{R}^{m_d}$ as the nonlinear feature mapping induced by the kernel $\mathbf{K}_d$ based on the source domain depth features, where D_d is the dimension of the original depth features, and m_d is the dimensionality of the nonlinear features. We also define $\mathbf{\Psi} = [\psi(\mathbf{z}_1), \ldots, \psi(\mathbf{z}_{n_s})] \in \mathbb{R}^{m_d\times n_s}$ as the data matrix of nonlinear depth features (*i.e.*, $\mathbf{K}_d = \mathbf{\Psi}'\mathbf{\Psi}$).

Now we consider how to incorporate the KCCA method into the formulation of our DAM2S framework in (1). Recall that in (1), we aim to learn a projection matrix $\mathbf{P}$ for the visual features such that the domain distribution mismatch can be reduced. To employ the additional depth features, we assume that in the common subspace, our framework can not only reduce domain distribution mismatch, but also maximize the feature correlation between two types of features. Therefore, we define the regularizer $\Delta(\cdot,\cdot)$ in (1) as follows,

$$\Delta^{mfc}(\mathbf{P},\mathbf{Q}) = -\text{tr}(\mathbf{Q}'\mathbf{\Psi}\mathbf{\Phi}_s'\mathbf{P}) \tag{6}$$

where $(\mathbf{P},\mathbf{Q}) \in \mathcal{P}_A = \{(\mathbf{P},\mathbf{Q})|\mathbf{P}'\mathbf{\Phi}_s\mathbf{\Phi}_s'\mathbf{P} + \mathbf{Q}'\mathbf{\Psi}\mathbf{\Psi}'\mathbf{Q} = \mathbf{I}_m\}$. By minimizing the above regularizer, the correlation between two types of features will be maximized when learning the classifiers in the common subspace.

Then, we learn the two classifiers f^v and f^d based on the projected visual and depth samples in the common subspace. In particular, we respectively learn two SVM classifiers f^v and f^d for the visual and depth features in the new subspace determined by the projection matrices $\mathbf{P}$ and $\mathbf{Q}$. Let us denote the visual feature based decision function for the k-th class as $f_k^v(\mathbf{x}) = \mathbf{w}_k'\mathbf{P}'\phi(\mathbf{x}) + b_k$, and the depth feature based decision function for the k-th class as $f_k^d(\mathbf{z}) = \tilde{\mathbf{w}}_k'\mathbf{Q}'\psi(\mathbf{z}) + \tilde{b}_k$, where $\mathbf{w}_k$ and $\tilde{\mathbf{w}}_k$ are the weight vectors, and b_k and $\tilde{b}_k$ are bias terms. We propose the following objective function for learning the classifiers from two types of features and K classes and two projection matrices $\mathbf{P}$ and $\mathbf{Q}$ that decide the common subspace,

$$\begin{aligned}\min_{\substack{(\mathbf{Q},\mathbf{P})\in\mathcal{P}_A,\mathbf{w}_k,\tilde{\mathbf{w}}_k\\ b_k,\tilde{b}_k,\xi_{i,k}^v,\xi_{i,k}^d}} &\frac{1}{2}\sum_{k=1}^{K}(\|\mathbf{w}_k\|^2+\|\tilde{\mathbf{w}}_k\|^2)+C\sum_{k=1}^{K}\sum_{i=1}^{n_s}(\xi_{i,k}^v+\xi_{i,k}^d)\\ &+\mu\Omega^{mmd}(\mathbf{P})+\lambda\Delta^{mfc}(\mathbf{P},\mathbf{Q}), \qquad (7)\\ \text{s.t.}\quad & l_{i,k}(\mathbf{w}_k'\mathbf{P}'\phi(\mathbf{x}_i^s)+b_k) \geq 1-\xi_{i,k}^v, \qquad (8)\\ & l_{i,k}(\tilde{\mathbf{w}}_k'\mathbf{Q}'\psi(\mathbf{z}_i)+\tilde{b}_k) \geq 1-\xi_{i,k}^d, \qquad (9)\\ & \xi_{i,k}^v \geq 0, \xi_{i,k}^d \geq 0, \quad \forall i,k,\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

where $\|\mathbf{w}_k\|^2$ (*resp.*, $\|\tilde{\mathbf{w}}_k\|^2$) is the regularizer to control the complexity of the classifier f_k^v (*resp.*, f_k^d), $\xi_{i,k}^v$ (*resp.*, $\xi_{i,k}^d$) is the hinge loss from the k-th classifier for the i-th sample using the visual feature (*resp.*, the depth feature), $\Omega^{mmd}(\mathbf{P})$ is the MMD based regularizer as defined in (2), $\Delta^{mfc}(\mathbf{P},\mathbf{Q})$ is the regularizer for maximizing the feature correlation as defined in (6), $\mathcal{P}_A$ is the feasible set of $(\mathbf{P},\mathbf{Q})$, and C, μ and λ are the tradeoff parameters as defined in (1).

Note the regularizer $r(f^v,f^d)$ in (1) corresponds to $\frac{1}{2}\sum_{k=1}^K(\|\mathbf{w}_k\|^2+\|\tilde{\mathbf{w}}_k\|^2)$ in the above objective, and the loss term $\ell(f^v,f^d)$ is $\sum_{k=1}^K\sum_{i=1}^{n_s}(\xi_{i,k}^v+\xi_{i,k}^d)$. We refer to the above formulation as DAM2S_A. After optimizing the above problem, the final classifier is obtained by equivalently fusing the decision values from those two classifiers (*i.e.*, $f(\mathbf{x})=\frac{1}{2}(\mathbf{w}_k'\mathbf{P}'\phi(\mathbf{x})+b_k+\tilde{\mathbf{w}}_k'\mathbf{P}'\phi(\mathbf{x})+\tilde{b}_k)$) for predicting any test sample $\mathbf{x}$ from the target domain.

3.3.2 Enforcing Classifier Consistency

Using the MMD based regularizer: Besides the KCCA based approach, another strategy in multi-view learning is to enforce the classifier consistency, *i.e.*, to enforce the predictions from the classifiers learnt on two types of features to be consistent on the training data. Formally, let us define the depth feature based SVM classifier for the k-th category as $f_k^d(\mathbf{z})=\tilde{\mathbf{w}}_k'\psi(\mathbf{z})+\tilde{b}_k$. Correspondingly, we denote $f_k^v(\mathbf{x})=\mathbf{w}_k'\mathbf{P}'\phi(\mathbf{x})+b_k$ as the visual feature based SVM classifier for the k-th category, which is the same as defined in Section 3.3.1 with $\mathbf{P}$ being the projection matrix for the visual features. Inspired by the SVM2K method [28], we define the regularizer $\Delta(\cdot,\cdot)$ in (1) by using the ϵ-insensitive loss of training samples based on the predictions from the visual and depth features based classifiers, namely,

$$\Delta^{ecc}(f^v,f^d)=\sum_{k=1}^K\sum_{i=1}^{n_s}\eta_{i,k} \tag{10}$$

where $\eta_{i,k}=\max(\epsilon,\left|f_k^v(\mathbf{x}_i^s)-f_k^d(\mathbf{z}_i)\right|)$, and ϵ is a constant set as 0.001 in our experiments. By minimizing the above regularizer, the visual and depth features based classifiers are enforced to be consistent for each category. Then we arrive at the following objective function,

$$\min_{\substack{\mathbf{P}\in\mathcal{P}_B,\mathbf{w}_k,\tilde{\mathbf{w}}_k\\ b_k,\tilde{b}_k,\xi_{i,k}^v,\xi_{i,k}^d}}\frac{1}{2}\sum_{k=1}^K(\|\mathbf{w}_k\|^2+\|\tilde{\mathbf{w}}_k\|^2)+C\sum_{k=1}^K\sum_{i=1}^{n_s}(\xi_{i,k}^v+\xi_{i,k}^d)$$
$$+\mu\Omega^{mmd}(\mathbf{P})+\lambda\Delta^{ecc}(f^v,f^d), \tag{11}$$
$$\text{s.t.}\quad l_{i,k}(\mathbf{w}_k'\mathbf{P}'\phi(\mathbf{x}_i^s)+b_k)\geq 1-\xi_{i,k}^v, \tag{12}$$
$$l_{i,k}(\tilde{\mathbf{w}}_k'\psi(\mathbf{z}_i)+\tilde{b}_k)\geq 1-\xi_{i,k}^d, \tag{13}$$
$$\xi_{i,k}^v\geq 0,\xi_{i,k}^d\geq 0,\quad \forall i,k,$$

where the first two terms are used to control the complexity of classifiers, the third and fourth terms are the hinge loss defined as in DAM2S_A, the regularizer $\Omega^{mmd}(\mathbf{P})$ is the MMD based regularizer as defined in (2), and the last term $\Delta^{ecc}(f^v,f^d)$ is the regularizer in (10) for enforcing the classifier consistency. Note we do not learn the projection matrix $\mathbf{Q}$ for the depth features, because we alternatively exploit the classifier consistency between the two types of features. Accordingly, the feasible set is defined as $\mathcal{P}_B=\{\mathbf{P}|\mathbf{P}'\mathbf{\Phi}_s\mathbf{\Phi}_s'\mathbf{P}=\mathbf{I}_m\}$.

We refer to the above formulation as DAM2S_B. After optimizing the above problem, the visual feature based classifiers f_k^v's are used to predict the samples in the target domain in the testing stage.

Using the kernel SA based regularizer: As discussed in Section 3.2.2, we can also employ the kernel SA based regularizer to reduce the domain distribution mismatch. In particular, by replacing the regularizer $\Omega^{mmd}(\mathbf{P})$ in (11) with the kernel SA based regularizer defined in (5), we arrive at the following objective function,

$$\min_{\substack{\mathbf{P},\mathbf{w}_k,\tilde{\mathbf{w}}_k,\\ b_k,\tilde{b}_k,\xi_{i,k}^v,\xi_{i,k}^d}}\frac{1}{2}\sum_{k=1}^K(\|\mathbf{w}_k\|^2+\|\tilde{\mathbf{w}}_k\|^2)+C\sum_{k=1}^K\sum_{i=1}^{n_s}(\xi_{i,k}^v+\xi_{i,k}^d)$$
$$+\mu\Omega^{ksa}(\mathbf{P})+\lambda\Delta^{ecc}(f^v,f^d), \tag{14}$$
$$\text{s.t.}\quad l_{i,k}(\mathbf{w}_k'\mathbf{P}'\mathbf{S}_s'\phi(\mathbf{x}_i^s)+b_k)\geq 1-\xi_{i,k}^v, \tag{15}$$
$$l_{i,k}(\tilde{\mathbf{w}}_k'\psi(\mathbf{z}_i)+\tilde{b}_k)\geq 1-\xi_{i,k}^d, \tag{16}$$
$$\xi_{i,k}^v\geq 0,\xi_{i,k}^d\geq 0,\quad \forall i,k,$$

which is referred to as DAM2S_C. As in DAM2S_B, the learnt visual feature based classifiers f_k^v's are used to predict the samples in the target domain in the test stage.

TABLE 1
Summary of the objective functions from our three DAM2S algorithms.

Method	DAM2S_A	DAM2S_B	DAM2S_C
$r(f^v,f^d)$	$\frac{1}{2}\sum_{k=1}^K(\|\mathbf{w}_k\|^2+\|\tilde{\mathbf{w}}_k\|^2)$		
$\ell(f^v,f^d)$	$\sum_{k=1}^K\sum_{i=1}^{n_s}(\xi_{i,k}^v+\xi_{i,k}^d)$		
$\Omega(\mathbf{P})$	$\Omega^{mmd}(\mathbf{P})$		$\Omega^{ksa}(\mathbf{P})$
$\Delta(\cdot,\cdot)$	$\Delta^{mfc}(\mathbf{P},\mathbf{Q})$	$\Delta^{ecc}(f^v,f^d)$	

3.4 Summary and Optimization

Summary: By using different strategies, we have proposed two regularizers to reduce the domain distribution mismatch and two forms of $\Delta(\cdot,\cdot)$ for incorporating the depth features., which leads to three DAM2S algorithms, DAM2S_A, DAM2S_B and DAM2S_C, as summarized in Table 1. We use the SVM classifiers for both types of features in all three algorithms, so the regularizer for controlling the classifier complexity and the loss term are the same. Then, we employ the MMD based regularizer $\Omega^{mmd}(\mathbf{P})$ in DAM2S_A and DAM2S_B, and the kernel SA based regularizer $\Omega^{ksa}(\mathbf{P})$ in DAM2S_C. For incorporating the depth information, we utilize the feature correlation maximization based regularizer in DAM2S_A, and the classifier consistency based regularizer in DAM2S_B and DAM2S_C. Note we do not combine the kernel SA based regularizer $\Omega^{ksa}(\mathbf{P})$ and the feature correlation maximization based regularizer $\Delta^{mfc}(\mathbf{P},\mathbf{Q})$, because the transformation matrix $\mathbf{P}$ in kernel SA has a different physical meaning with the projection matrix $\mathbf{P}$ in KCCA.

Optimization: We employ the similar alternating optimization strategy as in our preliminary work [19]. The above three problems can be optimized in a unified form

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

by deriving their dual forms, in which the parameter matrix $\mathbf{P}$ (or parameter matrices $(\mathbf{P}, \mathbf{Q})$) can be written in a single matrix in the dual form. Then, we alternatingly optimizing this matrix, and the SVM dual problems. This procedure is repeated until convergence. We leave the details in the Supplementary.

4 Experiments

4.1 Baseline Approaches

To the best of our knowledge, most existing works are not specifically designed for recognizing RGB images/videos in one domain by learning from the RGB-D data from another domain. Thus, we extend a broad range of existing works as the baselines for fair comparison, which includes:

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

Multi-view Learning: The multi-view learning approaches include KCCA [27] and SVM2K [28], in which the two-view data in the source domain are used for training. We use the classifier trained on the visual features for prediction. For SVM2K, two classifiers are trained by using the two-view data in the source domain, and we use visual feature base classifier to predict the target domain data. For KCCA, we train two SVM classifiers by using the projected depth and visual features in the common subspace. For the target samples, the decision values from the two classifiers based on the projected visual features are equally fused for prediction.

Learning Using Privileged Information: We compare our methods with the SVM+ approach proposed in [4], in which 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 [21], DAM [12], SGF [11], TCA [40], Landmark (LMK) [16], Subspace Alignment (SA) [17], and Domain Invariant Projection (DIP) [18]. For these methods, the samples from both domains based on the visual features are used for training the classifiers, and we predict the target domain data based on the visual features.

Note that the semi-supervised multi-view learning methods [41] and the multi-view domain adaptation approaches [24] 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 [10], [23] also cannot be used because the labeled samples in the target domain are required in these HDA methods.

4.2 Experimental Setup

We evaluate the effectiveness of our proposed three algorithms for different visual recognition tasks, including object recognition, cross-dataset human action recognition, and cross-view human action recognition.

Object Recognition: We evaluate our proposed three DAM2S algorithms for object recognition by using the RGB-D Object dataset [1] as the source domain and the Caltech-256 dataset [42] as the target domain. The RGB-D Object dataset contains the color and depth images of different objects from 51 categories. The Catech-256 dataset contains only the color images. In this work, we use the 10 common categories between the two datasets, which leads to a total number of 2059 training images. Moreover, all the target domain samples are also used as unlabeled data in the training stage for the baseline domain adaptation methods and our DAM2S algorithms.

For the RGB images from both source and target domains, we extract $4,096$-dim DeCAF$_6$ features [43]. For depth images in the source domain, we follow [44] to extract the kernel descriptors (KDES) features including Gradient KDES and LBP KDES from each depth image. The vocabulary size is set as 1000, and three levels of pyramids are used. Finally, the object level features constructed from the Gradient KDES and LBP KDES features are concatenated into one feature vector for each depth image.

Cross-Dataset Human Action Recognition: For cross-dataset action recogntion, we use the Hollywood 3D dataset [2] as the source domain, and the Hollywood2 dataset [45] as the target domain. The Hollywood2 dataset is a widely used benchmark dataset for human action recognition, which contains $1,707$ (823 in the training set and 884 in the test set) RGB videos from 12 human actions cropped from the Hollywood movies. Similarly, the Hollywood 3D dataset contains 650 RGB-D video clips from 14 human actions cropped from the Hollywood 3D movies. We use the left-eye video clips as the RGB training videos, and the reconstructed depth video clips as the depth training videos. Since the video clips in the two datasets are cropped from different movies captured using different types of equipments in different years, there is considerable domain distribution difference between these two datasets. In our experiments, we use the video from eight common actions between those two datasets for performance evaluation, which leads to 548 RGB-D videos in the source domain, and $1,279$ RGB videos in the target domain. All the target domain samples are also used as unlabeled data in the training stage for the baseline domain adaptation methods and our DAM2S algorithms.

We extract the improved dense trajectory (IDT) features for the RGB video clips from two datasets using the source code provided in [46]. Specifically, three types of space-time (ST) features (*i.e.*, 96-dim Histogram of Oriented Gradient, 108-dim Histogram of Optical Flow and 192-dim Motion Boundary Histogram) are used. We construct the codebook by using k-means clustering on the ST features from all videos in the training dataset to generate $2,000$ clusters, and then use the bag-of-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

words model representation for each type of ST features. Finally, each video is represented as the $6,000$-dim feature by concatenating the $2,000$-dim TF feature from each type of ST features. For the depth video clips, we use the same procedure to extract a $6,000$-dim depth feature. Note that the codebooks for the visual and depth features are different, so they cannot be treated as the same features although their dimensions are the same.

Cross-View Human Action Recognition: We use a recently released multi-view RGB-D action dataset ACT4[2] [3] for our experiments, which contains $2,648$ RGB-D videos from 14 human actions captured by Kinect from four different view points. To evaluate our proposed algorithms, we use the RGB-D videos from the first two views as the source domain, and use only the RGB videos from the remaining two views as the target domain, which leads to $1,324$ RGB-D videos in the source domain, and $1,324$ RGB videos in the target domain. Since the training and testing videos are captured from different views, there also exists considerable domain distribution mismatch. We extract the IDT features, and use the bag-of-words representation for the RGB videos and depth videos with the same procedure as in the cross-dataset setting. The other experimental settings are also the same as in the cross-dataset setting.

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. Moreover, we empirically set $\mu = 10^5$ for our two methods DAM2S_A and DAM2S_B, and $\mu = 10^4$ for our method DAM2S_C. We also empirically set $\lambda = 10^{-2}$ for DAM2S_A, $\lambda = 10^0$ for DAM2S_B, and $\lambda = 10^1$ for DAM2S_C. How to automatically choose the optimal parameters for our methods will be an interesting research issue in the future.

Experimental Results: The results of all methods in three tasks are reported in Table 2. From this table, we observe that our newly proposed DAM2S algorithms outperform all other baseline methods in all three tasks. It demonstrates the effectiveness of our DAM2S methods 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 learning using privileged information approach SVM+ achieve better results when compared with the naive approach SVM_A. Among these methods, SVM2K achieves the best result, as it can more effectively exploit depth information by learning two classifiers for both visual and depth features. Nevertheless, all these methods do not cope with the data distribution mismatch between the source and target domains, thus they are much worse than our three DAM2S algorithms.

The domain adaptation methods KMM, SGF, LMK, TCA, SA and DIP are also better than SVM_A, which shows it is beneficial to reduce domain distribution mismatch between the source and target domains by using these methods. When compared with SVM_A, DAM is slightly worse, possibly because the multi-source domain adaptation method cannot effectively handle the significant domain distribution mismatch with a single source domain in this application. Moreover, our proposed three DAM2S algorithms outperform all those methods by additionally exploiting the depth features in the source domain.

TABLE 3
Comparison of training time (seconds) for our three DAM2S methods for the cross-dataset human action recognition task.

	DAM2S_A	DAM2S_B	DAM2S_C
Training Time	20.7	168.4	398.2

4.3 Comparison among three DAM2S methods

Performance Analysis: By comparing the performance of our three methods in Table 2, we observe that the results of DAM2S_A and DAM2S_B are generally comparable. This indicates that when using the MMD based regularizer to reduce the domain distribution mismatch, the two strategies for incorporating the additional depth features can lead to comparable results. Moreover, it is interesting to observe that DAM2S_C is generally better than DAM2S_A and DAM2S_B in most cases. However, DAM2S_C is worse than the other two methods for the task under the cross-view setting. We conjecture that this is due to the significant domain distribution mismatch in this cross-view setting. It is less effective to use the SA based regularizer to learn a transformation matrix around the prelearnt matrix $\mathbf{T}$, when compared with the MMD based regularizer that optimizes the projection matrix directly based on the MMD criterion.

Time Comparison: We also report the training time of those three methods in Table 3 by using the cross-dataset human action recognition task as an example. From Table 3, we observe that our DAM2S_A is more efficiency than DAM2S_B and DAM2S_C in term of the training time, because DAM2S_B and DAM2S_C need to iteratively solve an SVM2K subproblem, which takes more time than the SVM problem in DAM2S_A. Moreover, DAM2S_C takes more iterations to converge when compared with DAM2S_B, so the training time of DAM2S_C is longer.

Base on the above analysis, DAM2S_A is more suitable for the applications with high requirements in training speed, due to its good tradeoff between performance and efficiency, while DAM2S_C should be a good choice for the performance driven applications where the training and test data are from different datasets.

5 Conclusions

In this paper, we have proposed a new framework for recognizing RGB images/videos by learning from a set of labeled RGB-D data. We formulate this task as a new

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2734890, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2
Comparison of accuracies (%) for objection recognition (OR), cross-dataset human action recognition (CD-HAR), and cross-view human action recognition (CV-HAR).

	SVM_A	SVM+	KCCA	SVM2K	KMM	DAM	SGF	LMK	TCA	SA	DIP	DAM2S_A	DAM2S_B	DAM2S_C
OR	27.52	29.04	31.52	34.21	28.88	27.37	38.93	39.50	37.79	44.55	45.13	**53.17**	**54.33**	**56.89**
CD-HAR	19.17	22.03	22.84	26.57	21.03	19.32	26.25	18.56	26.79	29.56	28.46	**30.64**	**31.75**	**39.17**
CV-HAR	62.33	65.36	64.78	62.81	62.39	61.41	59.18	63.28	52.95	64.90	62.73	**68.86**	**68.85**	**65.56**

unsupervised domain adaptation 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. We propose three DAM2S algorithms to address this new problem by taking advantage of the additional depth features in the source domain and simultaneously reducing data distribution mismatch between the source and target domains. Comprehensive experiments for object recognition, cross-dataset human action recognition and cross-view human action recognition have clearly demonstrated the effectiveness of our proposed three DAM2S approaches for recognizing RGB images and videos by learning from RGB-D data.

REFERENCES

[1] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," *ICRA*, 2011.

[2] S. Hadfield and R. Bowden, "Hollywood 3d: Recognizing actions in 3d natural scenes," in *CVPR*, 2013.

[3] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *ECCV Workshop on Consumer Depth Cameras for Computer Vision*, 2012.

[4] V. Vapnik and A. Vashist, "A new learning paradigm: learning using privileged information." *Neural networks*, vol. 22, no. 5-6, pp. 544–57, 2009.

[5] V. Sharmanska, I. Austria, N. Quadrianto, and C. Lampert, "Learning to rank using privileged information," in *ICCV*, 2013.

[6] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[7] H. Daumé III, "Frustratingly easy domain adaptation," in *ACL*, 2007.

[8] G. Angeli, P. Liang, and D. Klein, "A simple domain-independent probabilistic approach to generation," in *EMNLP*, 2010.

[9] C. Cortes and M. Mohri, "Domain adaptation in regression," in *ALT*, 2011.

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

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

[12] L. Duan, I. W. Tsang, D. Xu, and T. Chua, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *T-NNLS*, vol. 23, no. 3, pp. 504–518, 2012.

[13] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *T-PAMI*, vol. 34, no. 3, pp. 465–479, March 2012.

[14] L. Duan, D. Xu, I. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *T-PAMI*, vol. 34, no. 9, pp. 1667–1680, September 2012.

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

[16] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *ICML*, 2013.

[17] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013.

[18] M. Baktashmotlagh, M. Harandi, and M. S. Brian Lovell, "Unsupervised domain adaptation by domain invariant projection," in *ICCV*, 2013.

[19] L. Chen, W. Li, and D. Xu, "Recognizing RGB images by learning from RGB-D data," in *CVPR*, 2014.

[20] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *JMLR*, vol. 13, pp. 723–773, 2012.

[21] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *NIPS*, 2007.

[22] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[23] W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *T-PAMI*, vol. 36, no. 6, pp. 1134–1148, June 2014.

[24] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence, "Multi-view transfer learning with a large margin approach," in *KDD*, 2011.

[25] L. Chen, L. Duan, and D. Xu, "Event recognition in videos by learning from heterogeneous web sources," in *CVPR*, 2013, pp. 2666–2673.

[26] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *JMLR*, vol. 9, pp. 1757–1774, 2008.

[27] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computing*, vol. 16, no. 12, pp. 2639–2664, 2004.

[28] 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.

[29] A. Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," in *COLT*, 1998, pp. 92–100.

[30] W. Li, L. Duan, I. W.-H. Tsang, and D. Xu, "Co-labeling: A new multi-view learning approach for ambiguous problems," in *ICDM*, 2012, pp. 419–428.

[31] Z. Zhang and Q. Ji, "Classifier learning with hidden information," in *CVPR*, 2015.

[32] Q. Zhang, G. Hua, W. Liu, Z. Liu, and Z. Zhang, "Can visual recognition benefi from auxiliary information in training?" in *ACCV*, 2014.

[33] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *CVPR*, 2016.

[34] Z. Ding, M. Shao, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *AAAI*, 2014.

[35] C. Jia, Y. Kong, Z. Ding, and Y. R. Fu, "Latent tensor transfer learning for RGB-D action recognition," in *ACM MM*, 2014, pp. 87–96.

[36] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *CVPR*, 2016.

[37] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *CVPR*, 2016.

[38] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *IJCV*, vol. 54, no. 1-3, pp. 117–142, Aug. 2003.

[39] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via diemensionality reduction," in *AAAI*, 2008.

[40] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *T-NN*, vol. 22, no. 2, pp. 199–210, 2009.

[41] V. Sindhwani, P. Niyogi, and M. Belkin, "A coregularization approach to semi-supervised learning with multiple views," in *ICML*, 2005.

[42] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep., 2007.

[43] 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," in *ICML*, 2014.

[44] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *IROS*, 2011.

[45] I. L. M. Marszalek and C. Schmid, "Actions in context," in *CVPR*, 2009.

[46] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.