Cross-View Action Recognition Over Heterogeneous Feature Spaces

Xinxiao Wu, *Member, IEEE*, Han Wang, Cuiwei Liu, and Yunde Jia, *Member, IEEE*

***Abstract*— In cross-view action recognition, what you saw in one view is different from what you recognize in another view, since the data distribution even the feature space can change from one view to another. In this paper, we address the problem of transferring action models learned in one view (source view) to another different view (target view), where action instances from these two views are represented by heterogeneous features. A novel learning method, called heterogeneous transfer discriminant-analysis of canonical correlations (HTDCC), is proposed to discover a discriminative common feature space for linking source view and target view to transfer knowledge between them. Two projection matrices are learned to, respectively, map data from the source view and the target view into a common feature space via simultaneously minimizing the canonical correlations of interclass training data, maximizing the canonical correlations of intraclass training data, and reducing the data distribution mismatch between the source and target views in the common feature space. In our method, the source view and the target view neither share any common features nor have any corresponding action instances. Moreover, our HTDCC method is capable of handling only a few or even no labeled samples available in the target view, and can also be easily extended to the situation of multiple source views. We additionally propose a weighting learning framework for multiple source views adaptation to effectively leverage action knowledge learned from multiple source views for the recognition task in the target view. Under this framework, different source views are assigned different weights according to their different relevances to the target view. Each weight represents how contributive the corresponding source view is to the target view. Extensive experiments on the IXMAS data set demonstrate the effectiveness of HTDCC on learning the common feature space for heterogeneous cross-view action recognition. In addition, the weighting learning framework can achieve promising results on automatically adapting multiple transferred source-view knowledge to the target view.**

***Index Terms*— Cross-view action recognition, transfer learning, heterogeneous features, multiple views adaptation.**

Manuscript received August 5, 2013; revised January 19, 2014, July 22, 2014, and May 2, 2015; accepted May 27, 2015. Date of publication June 12, 2015; date of current version August 10, 2015. This work was supported in part by the Natural Science Foundation of China under Grant 61203274, in part by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20121101120029, and in part by the Excellent Young Scholars Research Fund through the Beijing Institute of Technology, Beijing, China, under Grant 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Andrea Cavallaro.

The authors are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wuxinxiao@bit.edu.cn; wanghan@bit.edu.cn; liucuiwei@bit.edu.cn; jiayunde@bit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2445293

I. Introduction

CROSS-VIEW human action recognition has posed substantial challenges for computer vision algorithms due to the large variations from one view to another. Since the same action appears quite differently when observed from different views, action models learned from one view may degrade the performances in another view. One possible solution [1]–[4] is building a view-independent 3D model of human body via the 3D reconstruction from multiple calibrated cameras or epipolar geometry reasoning based on point correspondences. Another strategy resorts to exploiting action representations that are insensitive to the changes of views, such as temporal self-similarity descriptors [5] and the view-style independent manifold representation [6]. Some other methods [7], [8] learn a separate model for each action class in each view, however, it is difficult to collect sufficient labeled samples covering all the action classes from all the views. Recently, transfer learning based methods [9]–[11] have emerged to adapt the action knowledge learned on one or more views (source views) to another different view (target view) by exploring the statistical connections between them. All these methods assume that the data from different views are represented by the same type of features with the same dimension.

In this work, we propose a new transfer learning approach, namely Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC), for cross-view action recognition over heterogeneous feature spaces. Our method is not restricted to action features of the same type between the source view and the target view, and can handle the heterogeneous action representations in the two views. Instead of requiring the corresponding observation of the same action instance from the source view and the target view, our method explores how to take advantage of label information of training data to learn a shared common feature space with more discriminations. Specifically, two projection matrices are learned to respectively map the source view and the target view to the common feature space by simultaneously minimizing the canonical correlations of inter-class training data and maximizing the canonical correlations of intra-class training data. In order to reduce the data distribution mismatch between the source view and the target view in the common feature space, a nonparametric criterion is incorporated in the objective function for minimizing the canonical correlation between the means of source-view and target-view samples in the optimization problem. Using the learned common feature space, action models learned in the source view can be

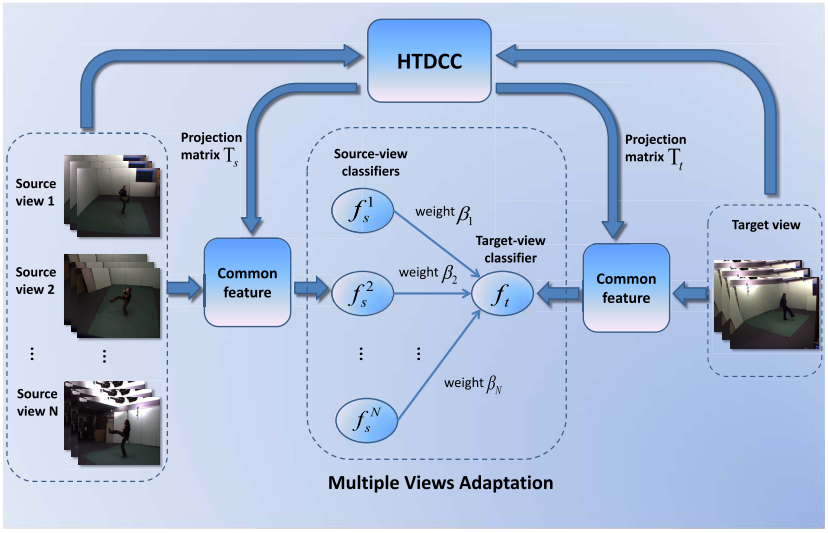


Fig. 1. Illustration of our framework. At the training stage, two projection matrices (i.e., Ts and Tt) are learned to find the common feature space between the source and target views using the HTDCC method; the source-view classifiers and their corresponding weights are learned to generate the target-view classifier using Multiple Views Adaptation Method. In testing, the input is an action video from target view, and the output is the class label of the input video.

easily adapted to the target view for classification. Thus, our method can successfully deal with the situation when there are limited or even no labeled data for training the action classifiers in the target view. Furthermore, our method can be readily generalized to the situation of multiple source views, in which multiple projection matrices are learned to map multiple source views and the target view to the common feature space.

Considering that one single source view can provide partial action knowledge, we additionally propose a weighting learning framework for multiple source views adaptation to adapt multiple transferred source-view classifiers to generate the target-view classifier. Since different source views perform different relations with the target view, a specific weight is assigned to each source view to measure its relevance to the target view. Consequently, our learning method can first automatically discover which source view is helpful to the target view, and then effectively transfer the beneficial source-view knowledge to the target view. For each source view, the Multiple Kernel Learning (MKL) method is employed to learn a robust classifier by effectively fusing multiple features. Multiple MKL source-view classifiers are combined to generate the target-view classifier according to their corresponding weights. The basic framework of our method is illustrated in Figure 1.

The main contributions of this work are summarized as follows: (1) a Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC) method is proposed for cross-view action recognition over heterogeneous features by discriminatively learning a common feature space. It is worth mentioning that our method can be readily applied to other recognition tasks (e.g., object recognition and face recognition) when the training and test data come from different domains with different features; (2) a weighting learning framework for multiple source views adaptation is proposed to fuse the pre-learned action models from multiple source views for building the target-view classifier. This learning scheme can automatically select the most relevant source views to the target view and meanwhile alleviate the negative transfer of less relevant source views. It can incorporate different common-feature learning methods (e.g., KCCA [12] and DAMA [13]) to make these methods applicable for multiple domains adaptation task.

A preliminary version of this paper appeared in ICCV 2013 [14]. The differences between this paper and [14] are: (1) this paper extends the HTDCC method to multiple source views in theory, making HTDCC applicative in both single view and multiple views adaptation; (2) in combining multiple source-view classifiers, this paper adopts MKL method for pre-learning the source-view classifiers by fusing multiple features, while [14] uses SVM to learn the source-view classifiers on single feature. In the experiments, for the related methods, this paper employs multiple features in the source view, learns multiple common feature spaces, and trains the source-view classifiers using the MKL method. In [14], all methods only use single feature in the source view.

II. Related Work

A. Cross-View Action Recognition Using Transfer Learning

From the perspective of cross-view action recognition, some work [9]–[11] is closely related to our approach. Farhadi and Tabrizi [9] used maximum margin clustering to generate the splits in the source view and then transferred the split values to the target view to learn the split-based features in the target view. Their work requires feature-to-feature correspondence at the frame-level to train a classifier. Liu *et al.* [10] proposed a bipartite graph-based approach to learn bilingual-words from source-view and target-view vocabularies, and then transferred action models between two views via the bag-of-bilingual-words model. Zheng *et al.* [11] presented a transferable dictionary pair consisting of two dictionaries that correspond to the source and target

TABLE I
SUMMARIZATION OF DIFFERENCE BETWEEN OUR METHOD AND STATE-OF-THE-ART METHODS ON CROSS-VIEW ACTION RECOGNITION USING TRANSFER LEARNING

Methods	Difference between the method and our method
[9]	[9] requires feature-to-feature correspondence while our method does not require feature-to-feature correspondence.
[10], [11]	[10], [11] require video-to-video correspondence while our method does not require video-to-video correspondence.
[15]	[15] uses the same feature type in all the views while our method allows heterogeneous feature in source and target views.

views respectively, and learned the same sparse representation of each video in the pair views. These two methods rely on simultaneous observations of the same action instance from multiple views. In contrast, our method requires neither the feature-to-feature correspondence nor the video-to-video correspondence, which significantly relaxes the requirements on the training data. Li and Zickler [15] proposed "virtual views" to connect action descriptors between the source and target views. Each virtual view is associated with a linear transformation of the action descriptor, and the sequence of transformed descriptors can be used to compare actions from different views. Different from [15], our method can handle the cross-view action recognition when the actions are represented by heterogeneous features in the source and target views. Table I summarizes the difference between our method and the state-of-the-art methods on cross-view action recognition using transfer learning.

B. Transfer Learning on Heterogeneous Features

From the perspective of transfer learning, our work is also related to the methods [12], [13], [16], [17] which find a "good" common feature space for the source and target domains. Shawe-Taylor and Cristianini [12] learned a common feature space by maximizing the correlation between the source and target data without any label information. Shi *et al.* [16] proposed a heterogeneous spectral mapping to discover a common feature subspace by learning two feature mapping matrices as well as the optimal projection of the data from both domains. The label information of training data from both domains is not used. Different from [12] and [16], our method does not require the sample correspondence between the source and target domains, and instead utilizes the label information to discover a common feature space with more discriminations. Wang and Mahadevan [13] proposed a manifold alignment based method to learn a common feature space for all heterogeneous domains by simultaneously maximizing the intra-domain similarity, minimizing the inter-domain similarity and preserving the topology of each domain. Although Wang and Mahadevan [13] used the class labels of data which is very similar to our method, they assumed that the data should have a manifold structure while we dose not require the manifold assumption of dataset. Kulis *et al.* [17] proposed a nonlinear metric learning method to learn an asymmetric feature transformation for the source and target data. Different from [17] which models a direct transformation from the source domain to the target domain, our method discovers a common feature space to connect the source domain and the target domain by learning two projection matrices. Duan *et al.* [18] proposed a heterogeneous feature augmentation method for heterogeneous domain adaptation, in which firstly the heterogeneous features are augmented using two feature mapping functions and then two projection matrices for the source and target data are learned by the standard SVM with the hinge loss in both linear and nonlinear cases. In their method, the learning of projection matrices and classifiers is jointly formulated in a standard SVM framework. While in our method, we first learn the projection matrices to find a common feature space and then can employ any classifiers based on the common space for cross-domain recognition. Table II summarizes the difference between our method and the state-of-the-art methods on heterogeneous transfer learning.

TABLE II
SUMMARIZATION OF DIFFERENCE BETWEEN OUR METHOD AND STATE-OF-THE-ART METHODS ON HETEROGENEOUS TRANSFER LEARNING

Methods	Difference between the method and our method
[12],[16]	[12],[16] require sample correspondence between source and target domains while our method does not require the correspondence.
[13]	[13] assumes the manifold structure of data while our method does not require the manifold assumption of dataset.
[17]	[17] models one transformation from source to target views while our method models two transformations to learn a common space.
[18]	[18] jointly learns the projection matrices and classifiers while our method first learn projection matrices and then train the classifiers.

C. Canonical Correlation Analysis for Action Recognition

Several existing methods [19]–[21] are related to our work in terms of canonical correlation analysis for action recognition. Kim *et al.* [19] proposed a tensor canonical correlation analysis method for human action classification, which extends classical canonical correlation analysis to multidimensional data arrays by taking into account the joint space-time domain of the video data. In [21], the authors extended the discriminant-analysis of canonical correlations method [22] to an incremental version for action recognition, in which the discriminative model is incrementally updated to capture the changes of human appearance. Different from these methods, our method focuses on cross-view action recognition over heterogeneous feature spaces using discriminant-analysis of canonical correlations, where the training data come from different domains with different features. Recently, Wu *et al.* [20] proposed a transfer discriminant-analysis canonical correlations for cross-view action recognition by minimizing the mismatch between data distributions of source and target views. The main difference between [20] and our method is that we focus on the heterogeneous transferring learning where the features from the source view and the target view are different, while [20] can only handle the same type of feature between the source view and the target view.

TABLE III
SUMMARIZATION OF DIFFERENCE BETWEEN OUR METHOD AND STATE-OF-THE-ART METHODS ON CANONICAL CORRELATION ANALYSIS FOR ACTION RECOGNITION

Methods	Difference between the method and our method
[19], [21]	[19], [21] use canonical correlation analysis for single-view action recognition while our method focuses on cross-view recognition.
[20]	[20] can only handle the same feature type between source and target views while our method can handle heterogeneous features.

Table III summarizes the difference between our method and the state-of-the-art methods on canonical correlation analysis for action recognition.

III. HETEROGENEOUS TRANSFER DISCRIMINANT-ANALYSIS OF CANONICAL CORRELATIONS

A. Problem Statement

In this work, each action sample is represented by an linear subspace of sequential image features. We do not take into account the temporal dynamics of an action and in some cases several principle images are sufficient to recognize what a person is doing. Denote $\mathrm{X} = [x_1, x_2, \ldots, x_M] \in \mathbb{R}^{D\times M}$ as the sequential image features of an action sample, where $x_i \in \mathbb{R}^D$ represents the i-th image feature. Suppose we have a large number of labeled training samples from the source view $\{\mathrm{X}_i^s|_{i=1}^{N_s}\}$ with $\mathrm{X}_i^s \in \mathbb{R}^{D_s\times M_i^s}$ where D_s is the dimension of source-view image feature and M_i^s is the number of images of the i-th source-view video, a limited number of labeled training samples from the target view $\{\mathrm{X}_i^t|_{i=1}^{N_t}\}$ with $\mathrm{X}_i^t \in \mathbb{R}^{D_t\times M_i^t}$ where D_t is the dimension of target-view image feature and M_i^t is the number of images of the i-th labeled target-view video, and some unlabeled samples from the target view $\{\mathrm{X}_i^u|_{i=1}^{N_u}\}$ with $\mathrm{X}_i^u \in \mathbb{R}^{D_t\times M_i^u}$ where M_i^u is the number of images of the i-th unlabeled target-view video. Since the source and target samples are represented by heterogeneous image features i.e., $D_s \neq D_t$, we aim to find a common feature space of the two views as well as two projection matrices T_s and T_t which respectively map the source and target views to the common space.

B. Background

Discriminant-analysis of Canonical Correlations (DCC) [22] learns a projection matrix by maximizing canonical correlations of within-class samples and minimizing canonical correlations of between-class samples. Assume that N training samples are given as $\{\mathrm{X}_i|_{i=1}^N\}$ where $\mathrm{X}_i \in \mathbb{R}^{D\times M_i}$ belongs to one action class denoted by C_i. A m-dimensional linear subspace of X_i is represented by an orthonormal basis matrix $\mathrm{P}_i \in \mathbb{R}^{D\times m}$ s.t. $\mathrm{X}_i\mathrm{X}_i{}^T = \mathrm{P}_i\Lambda_i\mathrm{P}_i^T$, where Λ_i and P_i are the eigenvalue and eigenvector matrices of the m largest eigenvalues, respectively. Suppose that a projection matrix $\mathrm{T} = [t_1, t_2, \ldots, t_d] \in \mathbb{R}^{D\times d}$ is defined by $\mathrm{Y}_i = \mathrm{T}^T\mathrm{X}_i$ to make the projected samples more discriminative using canonical correlations, where $d \leq D$ and $|t_i| = 1$. Then the orthonormal basis matrices of the subspaces of projected data are given by $\mathrm{Y}_i\mathrm{Y}_i^T = (\mathrm{T}^T\mathrm{X}_i)(\mathrm{T}^T\mathrm{X}_i)^T = (\mathrm{T}^T\mathrm{P}_i)\Lambda_i(\mathrm{T}^T\mathrm{P}_i)^T$. Except when T is an orthonormal matrix, $\mathrm{T}^T\mathrm{P}_i$ is not generally an orthonormal basis matrix. According to [22] where the canonical correlations are only defined for orthonormal basis matrices of subspaces, the matrix P_i should be normalized to P'_i for a fixed T so that any orthonormal components of $\mathrm{T}^T\mathrm{P}'_i$ can represent an orthonormal basis matrix of the projected data. Specifically, the normalization is given as follows: (1) QR-decomposition of $\mathrm{T}^T\mathrm{P}_i$ is performed s.t. $\mathrm{T}^T\mathrm{P}_i = \Phi_i\Delta_i$, where $\Phi_i \in \mathbb{R}^{d\times m}$ is the orthonormal matrix composed by the first m columns and $\Delta_i \in \mathbb{R}^{m\times m}$ is the $m\times m$ invertible uppertriangular matrix; (2) Given $\Phi_i = \mathrm{T}^T(\mathrm{P}_i\Delta_i^{-1})$, P'_i is computed by $\mathrm{P}'_i = \mathrm{P}_i\Delta_i^{-1}$.

The similarity of any two projected samples is defined as the sum of canonical correlations $\mathrm{F}_{ij} = \max_{\mathrm{Q}_{ij},\mathrm{Q}_{ji}} \mathrm{Tr}(\mathrm{T}^T\mathrm{P}'_j\mathrm{Q}_{ji}\mathrm{Q}_{ij}^T\mathrm{P}_i'^T\mathrm{T})$ s.t. $\mathrm{Q}_{ij}^T\mathrm{Q}_{ij} = \mathrm{Q}_{ij}\mathrm{Q}_{ij}^T = \mathrm{Q}_{ji}^T\mathrm{Q}_{ji} = \mathrm{Q}_{ji}\mathrm{Q}_{ji}^T = \mathrm{I}$, where the solutions of Q_{ij} and Q_{ji} are given by the SVD computation $(\mathrm{T}^T\mathrm{P}'_i)^T(\mathrm{T}^T\mathrm{P}'_j) = \mathrm{Q}_{ij}\Lambda\mathrm{Q}_{ji}^T$. T is determined to maximize the similarities of any pair of intra-class samples and minimize the similarities of any pair of inter-class samples, defined by

$$\mathrm{T} = \underset{\mathrm{T}}{\operatorname{argmax}} \frac{E_w(\mathrm{T})}{E_b(\mathrm{T})}, \tag{1}$$

where $E_w(\mathrm{T})$ and $E_b(\mathrm{T})$ represent the sums of similarities of any pair of intra-class and inter-class samples, respectively, defined by $E_w(\mathrm{T}) = \sum_{i=1}^N\sum_{k\in W_i}\mathrm{F}_{ik}$ and $E_b(\mathrm{T}) = \sum_{i=1}^N\sum_{l\in B_i}\mathrm{F}_{il}$ where the indices are defined as $W_i = \{j|C_j = C_i\}$ and $B_i = \{j|C_j \neq C_i\}$. That is, the two index sets W_i and B_i respectively denote the intra-class and inter-class samples for a given sample of class C_i.

Transfer Discriminant-analysis of Canonical Correlations (TDCC) [20] is an extension of DCC for handling the situation when the training and test samples have different data distribution properties. In order to reduce the mismatch between data distributions of different domains, an effective nonparametric criterion is integrated into the discriminative function in Eqn.1, formulated as

$$\mathrm{T} = \arg\max_{\mathrm{T}} \frac{E_w(\mathrm{T}) + \alpha E_r(\mathrm{T})}{E_b(\mathrm{T})}, \tag{2}$$

where $E_r(\mathrm{T})$ is the canonical correlation of between-view mean samples from source and target domains and α is the tradeoff parameter.

C. Learning on Heterogeneous Feature Spaces

Table IV lists the important notations used throughout the paper.

Different from [20] and [22], our proposed Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC) deals with the situation when the training data and test data are drawn from different views with heterogeneous features. In HTDCC, two projection matrices are learned to respectively map the source view and the target view to a common space where the samples from the same class are closely-related to each other, the samples from different classes are well-separated from each other, and the

TABLE IV
NOTATION

X_i^s	i-th source-view training sample
X_i^t	i-th labeled target-view training sample
X_i^u	i-th unlabeled target-view training sample
P_i^s	orthonormal basis matrix of subspace of X_i^s
P_i^t	orthonormal basis matrix of subspace of X_i^t
P_i^u	orthonormal basis matrix of subspace of X_i^u
$\mathrm{P}_i^{s\prime}$	normalization of P_i^s
$\mathrm{P}_i^{t\prime}$	normalization of P_i^t
$\mathrm{P}_i^{u\prime}$	normalization of P_i^u
T_s	projection matrix from source view to the common space
T_t	projection matrix from target view to the common space

data distributions of the source and target views are matched to each other. Given the source-view training data $\{\mathrm{X}_i^s\}_{i=1}^{N_s}$ with the corresponding labels $\{C_i^s\}_{i=1}^{N_s}$ where X_i^s denotes the i-th training sample from the source view and C_i^s is the action class label of X_i^s, the source-view projection matrix $\mathrm{T}_s = [t_{s,1}, t_{s,2}, \ldots, t_{s,d}] \in \mathbb{R}^{D_s \times d}$ is defined by $\mathrm{Y}_i^s = \mathrm{T}_s^T \mathrm{X}_i^s$. Let $\mathrm{P}_i^s \in \mathbb{R}^{D_s \times m}$ be the orthonormal basis matrix of the m-dimensional linear subspace of X_i^s, the projection of P_i^s is represented by $\mathrm{T}_s^T \mathrm{P}_i^{s\prime}$ where $\mathrm{P}_i^{s\prime}$ indicates the normalization of P_i^s. Given the labeled target-view training data $\{\mathrm{X}_i^t\}_{i=1}^{N_t}$ with the corresponding labels $\{C_i^t\}_{i=1}^{N_t}$ and the unlabeled target-view training data $\{\mathrm{X}_i^u\}_{i=1}^{N_u}$, the target-view projection matrix $\mathrm{T}_t = [t_{t,1}, t_{t,2}, \ldots, t_{t,d}] \in \mathbb{R}^{D_t \times d}$ is defined by $\mathrm{Y}_i^t = \mathrm{T}_t^T \mathrm{X}_i^t$. Let $\mathrm{P}_i^t \in \mathbb{R}^{D_t \times m}$ and $\mathrm{P}_i^u \in \mathbb{R}^{D_t \times m}$ be respectively the orthonormal subspaces of X_i^t and X_i^u, then the projected representations of P_i^t and P_i^u are $\mathrm{T}_t^T \mathrm{P}_i^{t\prime}$ and $\mathrm{T}_t^T \mathrm{P}_i^{u\prime}$ where $\mathrm{P}_i^{t\prime}$ and $\mathrm{P}_i^{u\prime}$ indicate the normalizations of P_i^t and P_i^u, respectively.

The projection matrices T_s and T_t are defined with objective function J by

$$\underset{\mathrm{T}_s,\mathrm{T}_t}{\operatorname{argmax}} J = \underset{\mathrm{T}_s,\mathrm{T}_t}{\operatorname{argmax}} \frac{E_w(\mathrm{T}_s, \mathrm{T}_t) + \alpha E_r(\mathrm{T}_s, \mathrm{T}_t)}{E_b(\mathrm{T}_s, \mathrm{T}_t)}, \tag{3}$$

where $E_w(\mathrm{T}_s, \mathrm{T}_t)$ and $E_b(\mathrm{T}_s, \mathrm{T}_t)$ respectively represent the sums of similarities of all the pairs of intra-class and inter-class training samples from both source and target views. $E_r(\mathrm{T}_s, \mathrm{T}_t)$ represents the similarity between the source-view mean sample and the target-view mean sample. The detailed formulations are given by

$$\begin{aligned}
E_w(\mathrm{T}_s, \mathrm{T}_t) &= \sum_{i=1}^{N_s} \sum_{j \in W_i^s} \mathrm{F}_{ij}^s + \sum_{i=1}^{N_t} \sum_{j \in W_i^t} \mathrm{F}_{ij}^t \\
&\quad + \sum_{i=1}^{N_s} \sum_{j \in W_i^{st}} \mathrm{F}_{ij}^{st} + \sum_{i=1}^{N_t} \sum_{j \in W_i^{ts}} \mathrm{F}_{ij}^{ts}, \\
E_b(\mathrm{T}_s, \mathrm{T}_t) &= \sum_{i=1}^{N_s} \sum_{j \in B_i^s} \mathrm{F}_{ij}^s + \sum_{i=1}^{N_t} \sum_{j \in B_i^t} \mathrm{F}_{ij}^t \\
&\quad + \sum_{i=1}^{N_s} \sum_{j \in B_i^{st}} \mathrm{F}_{ij}^{st} + \sum_{i=1}^{N_t} \sum_{j \in B_i^{ts}} \mathrm{F}_{ij}^{ts}, \\
E_r(\mathrm{T}_s, \mathrm{T}_t) &= \mathrm{F}_r^{st} + \mathrm{F}_r^{ts},
\end{aligned} \tag{4}$$

where the index sets $W_i^s = \{j | C_j^s = C_i^s\}$ and $B_i^s = \{j | C_j^s \neq C_i^s\}$ respectively indicate the intra-class and inter-class data from the source view for a given source-view sample of class C_i^s. $W_i^t = \{j | C_j^t = C_i^t\}$ and $B_i^t = \{j | C_j^t \neq C_i^t\}$ respectively indicate the intra-class and inter-class data from the target view for a given target-view sample of class C_i^t. $W_i^{st} = \{j | C_j^t = C_i^s\}$ and $B_i^{st} = \{j | C_j^t \neq C_i^s\}$ respectively indicate the intra-class and inter-class data from the target view for a given source-view sample of class C_i^s. $W_i^{ts} = \{j | C_j^s = C_i^t\}$ and $B_i^{ts} = \{j | C_j^s \neq C_i^t\}$ respectively indicate the intra-class and inter-class data from the source view for a given target-view sample of class C_i^t. F_{ij}^s and F_{ij}^t represent the canonical correlations of any two projected samples from the source view and the target view, respectively. Both F_{ij}^{st} and F_{ij}^{ts} represent the canonical correlations between two projected samples of which one sample is from the source view and the other is from the target view. Both F_r^{st} and F_r^{ts} represent the canonical correlation between the mean of source-view samples and the mean of target-view samples in the common feature space. They are parameterized as follows:

$$\begin{aligned}
\mathrm{F}_{ij}^s &= \max_{\mathrm{Q}_{ij}^s, \mathrm{Q}_{ji}^s} \mathbf{Tr}(\mathrm{T}_s^T \mathrm{P}_j^{s\prime} \mathrm{Q}_{ji}^s \mathrm{Q}_{ij}^{s\,T} \mathrm{P}_i^{s\prime T} \mathrm{T}_s), \\
\mathrm{F}_{ij}^t &= \max_{\mathrm{Q}_{ij}^t, \mathrm{Q}_{ji}^t} \mathbf{Tr}(\mathrm{T}_t^T \mathrm{P}_j^{t\prime} \mathrm{Q}_{ji}^t \mathrm{Q}_{ij}^{t\,T} \mathrm{P}_i^{t\prime T} \mathrm{T}_t), \\
\mathrm{F}_{ij}^{st} &= \max_{\mathrm{Q}_{ij}^{st}, \mathrm{Q}_{ji}^{st}} \mathbf{Tr}(\mathrm{T}_t^T \mathrm{P}_j^{s\prime} \mathrm{Q}_{ji}^{st} \mathrm{Q}_{ij}^{st\,T} \mathrm{P}_i^{s\prime T} \mathrm{T}_s), \\
\mathrm{F}_{ij}^{ts} &= \max_{\mathrm{Q}_{ij}^{ts}, \mathrm{Q}_{ji}^{ts}} \mathbf{Tr}(\mathrm{T}_s^T \mathrm{P}_j^{s\prime} \mathrm{Q}_{ji}^{ts} \mathrm{Q}_{ij}^{ts\,T} \mathrm{P}_i^{t\prime T} \mathrm{T}_t), \\
\mathrm{F}_r^{st} &= \max_{\mathrm{Q}_r^{st}, \mathrm{Q}_r^{ts}} \mathbf{Tr}(\mathrm{T}_t^T \mathrm{P}_r^{t\prime} \mathrm{Q}_r^{ts} \mathrm{Q}_r^{st\,T} \mathrm{P}_r^{s\prime T} \mathrm{T}_s), \\
\mathrm{F}_r^{ts} &= \max_{\mathrm{Q}_r^{ts}, \mathrm{Q}_r^{st}} \mathbf{Tr}(\mathrm{T}_t^T \mathrm{P}_r^{s\prime} \mathrm{Q}_r^{st} \mathrm{Q}_r^{ts\,T} \mathrm{P}_r^{t\prime T} \mathrm{T}_s),
\end{aligned} \tag{5}$$

where $\mathrm{P}_r^{s\prime}$ is the normalization of the mean of orthonormal subspaces of source-view training samples $\mathrm{P}_r^s = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathrm{P}_i^s$, and $\mathrm{P}_r^{t\prime}$ is the normalization of the mean of orthonormal subspaces of target-view training samples $\mathrm{P}_r^t = \frac{1}{N_t + N_u} (\sum_{i=1}^{N_t} \mathrm{P}_i^t + \sum_{i=1}^{N_u} \mathrm{P}_i^u)$. The Q_{ij}^s, Q_{ji}^s, Q_{ij}^t, Q_{ji}^t, Q_{ij}^{st}, Q_{ji}^{st}, Q_{ij}^{ts}, Q_{ji}^{ts}, Q_r^{st} and Q_r^{ts} are constrained to be orthonormal matrices, solved by

$$\begin{aligned}
(\mathrm{T}_s^T \mathrm{P}_i^{s\prime})^T (\mathrm{T}_s^T \mathrm{P}_j^{s\prime}) &= \mathrm{Q}_{ij}^s \Lambda \mathrm{Q}_{ji}^{s\,T}, \\
(\mathrm{T}_t^T \mathrm{P}_i^{t\prime})^T (\mathrm{T}_t^T \mathrm{P}_j^{t\prime}) &= \mathrm{Q}_{ij}^t \Lambda \mathrm{Q}_{ji}^{t\,T}, \\
(\mathrm{T}_s^T \mathrm{P}_i^{s\prime})^T (\mathrm{T}_t^T \mathrm{P}_j^{t\prime}) &= \mathrm{Q}_{ij}^{st} \Lambda \mathrm{Q}_{ji}^{st\,T}, \\
(\mathrm{T}_t^T \mathrm{P}_i^{t\prime})^T (\mathrm{T}_s^T \mathrm{P}_j^{s\prime}) &= \mathrm{Q}_{ij}^{ts} \Lambda \mathrm{Q}_{ji}^{ts\,T}, \\
(\mathrm{T}_s^T \mathrm{P}_r^{s\prime})^T (\mathrm{T}_t^T \mathrm{P}_r^{t\prime}) &= \mathrm{Q}_r^{st} \Lambda \mathrm{Q}_r^{ts\,T}.
\end{aligned}$$

By the linear algebra transformation

$$\mathrm{AB}^T = \mathbf{I} - \frac{1}{2}(\mathrm{A} - \mathrm{B})(\mathrm{A} - \mathrm{B})^T \,^1 \tag{6}$$

where $\mathrm{A} = \mathrm{T}^T \mathrm{P}_j' \mathrm{Q}_{ji}$ and $\mathrm{B} = \mathrm{T}^T \mathrm{P}_i' \mathrm{Q}_{ij}$, we can rewrite the objective function in Eqn.3 as

$$\max_{\mathrm{T}_s,\mathrm{T}_t} \frac{\mathbf{Tr}(\begin{bmatrix}\mathrm{T}_s \\ \mathrm{T}_t\end{bmatrix}^T \begin{bmatrix}\mathrm{S}_b^s & \mathrm{S}_b^{ts} \\ \mathrm{S}_b^{st} & \mathrm{S}_b^t\end{bmatrix} \begin{bmatrix}\mathrm{T}_s \\ \mathrm{T}_t\end{bmatrix})}{\mathbf{Tr}(\begin{bmatrix}\mathrm{T}_s \\ \mathrm{T}_t\end{bmatrix}^T \begin{bmatrix}\mathrm{S}_w^s & \mathrm{S}_w^{ts} + \alpha \mathrm{S}_r^{ts} \\ \mathrm{S}_w^{st} + \alpha \mathrm{S}_r^{st} & \mathrm{S}_w^t\end{bmatrix} \begin{bmatrix}\mathrm{T}_s \\ \mathrm{T}_t\end{bmatrix})}, \tag{7}$$

[1]When A and B are orthonormal matrices, we have $\mathrm{AB}^T = (\mathrm{AB}^T)^T = \mathrm{BA}^T$. So $\mathrm{I} - \frac{1}{2}(\mathrm{A} - \mathrm{B})(\mathrm{A} - \mathrm{B})^T = I - \frac{1}{2}(\mathrm{AA}^T - \mathrm{AB}^T - \mathrm{BA}^T + \mathrm{BB}^T) = \mathrm{I} - \frac{1}{2}(2\mathrm{I} - \mathrm{AB}^T - \mathrm{BA}^T) = \frac{1}{2}(\mathrm{AB}^T + \mathrm{BA}^T) = \mathrm{AB}^T$.

where

$$\mathrm{S}_b^s = \sum_{i=1}^{N_s} \sum_{j \in B_i^s} (\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^s - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^s)(\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^s - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^s)^T,$$

$$\mathrm{S}_b^t = \sum_{i=1}^{N_t} \sum_{j \in B_i^t} (\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^t - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^t)(\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^t - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^t)^T,$$

$$\mathrm{S}_b^{ts} = \sum_{i=1}^{N_t} \sum_{j \in B_i^{ts}} (\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^{ts} - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^{ts})(\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^{ts} - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^{ts})^T,$$

$$\mathrm{S}_b^{st} = \sum_{i=1}^{N_s} \sum_{j \in B_i^{st}} (\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^{st} - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^{st})(\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^{st} - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^{st})^T,$$

$$\mathrm{S}_w^s = \sum_{i=1}^{N_s} \sum_{j \in W_i^s} (\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^s - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^s)(\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^s - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^s)^T,$$

$$\mathrm{S}_w^t = \sum_{i=1}^{N_t} \sum_{j \in W_i^t} (\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^t - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^t)(\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^t - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^t)^T,$$

$$\mathrm{S}_w^{ts} = \sum_{i=1}^{N_t} \sum_{j \in W_i^{ts}} (\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^{ts} - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^{ts})(\mathrm{P}_j^{s\prime}\mathrm{Q}_{ji}^{ts} - \mathrm{P}_i^{t\prime}\mathrm{Q}_{ij}^{ts})^T,$$

$$\mathrm{S}_w^{st} = \sum_{i=1}^{N_s} \sum_{j \in W_i^{st}} (\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^{st} - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^{st})(\mathrm{P}_j^{t\prime}\mathrm{Q}_{ji}^{st} - \mathrm{P}_i^{s\prime}\mathrm{Q}_{ij}^{st})^T,$$

$$\mathrm{S}_r^{ts} = (\mathrm{P}_r^{s\prime}\mathrm{Q}_r^{ts} - \mathrm{P}_r^{t\prime}\mathrm{Q}_r^{ts})(\mathrm{P}_r^{s\prime}\mathrm{Q}_r^{ts} - \mathrm{P}_r^{t\prime}\mathrm{Q}_r^{ts})^T,$$

$$\mathrm{S}_r^{st} = (\mathrm{P}_r^{t\prime}\mathrm{Q}_r^{st} - \mathrm{P}_r^{s\prime}\mathrm{Q}_r^{st})(\mathrm{P}_r^{t\prime}\mathrm{Q}_r^{st} - \mathrm{P}_r^{s\prime}\mathrm{Q}_r^{st})^T.$$

Finally, by the eigen-decomposition

$$\begin{bmatrix} \mathrm{S}_b^s & \mathrm{S}_b^{ts} \\ \mathrm{S}_b^{st} & \mathrm{S}_b^t \end{bmatrix} t = \lambda \begin{bmatrix} \mathrm{S}_w^s & \mathrm{S}_w^{ts} + \alpha \mathrm{S}_r^{ts} \\ \mathrm{S}_w^{st} + \alpha \mathrm{S}_r^{st} & \mathrm{S}_w^t \end{bmatrix} t, \tag{8}$$

the optimal T_s and T_t are respectively constructed by the first-D_s rows and the last-D_t rows of the top-d eigenvectors $[t_1, t_2, \ldots, t_d]$.

We use an iterative optimization algorithm to find the optimal projection matrices T_s and T_t. A pseudocode for the learning is given in Algorithm 1. With the identity matrix I as the initial values of T_s and T_t, the algorithm is iterated until it converges to a stable point. The value of the objective function J for all cases becomes stable after first few iterations, starting with the initial value. For all of the experiments in Section V, the number of iterations was fixed to five. The proposed learning took about 95 seconds on a PC with Intel Core 2.83GHz CPU and 8 GM of RAM using non-optimized Matlab code. Once the optimal T_s and T_t are found, the similarity of any two action samples is measured by mapping them to the common space and computing the canonical correlations of them.

D. Extension to Multiple Source Views

Our method can easily be generalized to the situation of multiple source views, in which multiple projection matrices are learned to map multiple source views and the target view to the common feature space. Given K source views, let $\mathrm{T}_{s_k} \in \mathbb{R}^{D_{s_k} \times d}$ $(k = 1, 2, \ldots, K)$ be the k-th source-view projection matrix and $\mathrm{T}_t \in \mathbb{R}^{D_t \times d}$ be the target-view projection matrix. Then the objective function can be formulated by

$$\max_{\mathrm{T}_{s_k}, \mathrm{T}_t} \frac{\sum_k (E_w(\mathrm{T}_{s_k}, \mathrm{T}_t) + \alpha_k E_r(\mathrm{T}_{s_k}, \mathrm{T}_t)) + \sum_k \sum_l E_w(\mathrm{T}_{s_k}, \mathrm{T}_{s_l})}{\sum_k E_b(\mathrm{T}_{s_k}, \mathrm{T}_t) + \sum_k \sum_l E_b(\mathrm{T}_{s_k}, \mathrm{T}_{s_l})}, \tag{9}$$

where $E_w(\mathrm{T}_{s_k}, \mathrm{T}_t)$ and $E_b(\mathrm{T}_{s_k}, \mathrm{T}_t)$ respectively represent the sums of similarities of all the pairs of intra-class and inter-class training samples from both the k-th source view and the target view. $E_r(\mathrm{T}_{s_k}, \mathrm{T}_t)$ denotes the similarity between the mean sample of the k-th source view and the mean sample of the target view. $E_w(\mathrm{T}_{s_k}, \mathrm{T}_t)$, $E_b(\mathrm{T}_{s_k}, \mathrm{T}_t)$ and $E_r(\mathrm{T}_{s_k}, \mathrm{T}_t)$ have been defined in Eqn.4. $E_w(\mathrm{T}_{s_k}, \mathrm{T}_{s_l})$ and $E_b(\mathrm{T}_{s_k}, \mathrm{T}_{s_l})$ indicate the similarities of intra-class and inter-class training samples from both the k-th and l-th source views, respectively, given by

$$E_w(\mathrm{T}_{s_k}, \mathrm{T}_{s_l}) = \sum_{i=1}^{N_{s_k}} \sum_{j \in W_i^{s_l}} \mathrm{F}_{ij}^{s_k s_l} + \sum_{i=1}^{N_{s_l}} \sum_{j \in W_i^{s_k}} \mathrm{F}_{ij}^{s_l s_k},$$

$$E_b(\mathrm{T}_{s_k}, \mathrm{T}_{s_l}) = \sum_{i=1}^{N_{s_k}} \sum_{j \in B_i^{s_l}} \mathrm{F}_{ij}^{s_k s_l} + \sum_{i=1}^{N_{s_l}} \sum_{j \in B_i^{s_k}} \mathrm{F}_{ij}^{s_l s_k},$$

where $\mathrm{F}_{ij}^{s_k s_l}$ represent the canonical correlations between two projected samples of which one sample is from the k-th source view and the other is from the l-th source view, defined by $\mathrm{F}_{ij}^{s_k s_l} = \max_{\mathrm{Q}_{ij}^{s_k s_l}, \mathrm{Q}_{ji}^{s_k s_l}} \mathbf{Tr}(\mathrm{T}_{s_l}^T \mathrm{P}_j^{s_l\prime} \mathrm{Q}_{ji}^{s_k s_l} \mathrm{Q}_{ij}^{s_k s_l T} \mathrm{P}_i^{s_k\prime T} \mathrm{T}_{s_k})$ with the solutions of $\mathrm{Q}_{ij}^{s_k s_l}$ and $\mathrm{Q}_{ji}^{s_k s_l}$ by $(\mathrm{T}_{s_k}^T \mathrm{P}_i^{s_k\prime})^T (\mathrm{T}_{s_l}^T \mathrm{P}_j^{s_l\prime}) = \mathrm{Q}_{ij}^{s_k s_l} \Lambda \mathrm{Q}_{ji}^{s_k s_l T}$. The two sets $W_i^{s_k}$ and $B_i^{s_k}$ respectively index the intra-class and inter-class samples of the k-th source view for a given samples of class C_i. By the linear algebra transformation in Eqn.6, we can derive the following objective function:

$$\max_{\mathrm{T}_{s_k}, \mathrm{T}_t} \frac{\mathbf{Tr}(\mathbb{T}^T \mathbb{G} \mathbb{T})}{\mathbf{Tr}(\mathbb{T}^T \mathbb{H} \mathbb{T})}, \tag{10}$$

where

$$\mathbb{T} = \begin{bmatrix} \mathrm{T}_{s_1} \\ \vdots \\ \mathrm{T}_{s_K} \\ \mathrm{T}_t \end{bmatrix}, \quad \mathbb{G} = \begin{bmatrix} \mathrm{S}_b^{s_1} & \cdots & \mathrm{S}_b^{s_K s_1} & \mathrm{S}_b^{ts_1} \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{S}_b^{s_1 s_K} & \cdots & \mathrm{S}_b^{s_K} & \mathrm{S}_b^{ts_K} \\ \mathrm{S}_b^{s_1 t} & \cdots & \mathrm{S}_b^{s_K t} & \mathrm{S}_b^t \end{bmatrix},$$

$$\mathbb{H} = \begin{bmatrix} \mathrm{S}_w^{s_1} & \cdots & \mathrm{S}_w^{s_K s_1} & \mathrm{S}_w^{ts_1} + \alpha_{s_1} \mathrm{S}_r^{ts_1} \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{S}_w^{s_1 s_K} & \cdots & \mathrm{S}_w^{s_K} & \mathrm{S}_w^{ts_K} + \alpha_{s_K} \mathrm{S}_r^{ts_K} \\ \mathrm{S}_w^{s_1 t} + \alpha_{s_1} \mathrm{S}_r^{s_1 t} & \cdots & \mathrm{S}_w^{s_K t} + \alpha_{s_K} \mathrm{S}_r^{s_K t} & \mathrm{S}_w^t \end{bmatrix},$$

where $\mathrm{S}_b^{s_k}$, $\mathrm{S}_b^{ts_k}$, $\mathrm{S}_b^{s_k t}$, S_b^t, $\mathrm{S}_w^{s_k}$, $\mathrm{S}_w^{ts_k}$, $\mathrm{S}_w^{s_k t}$, S_w^t, $\mathrm{S}_r^{s_k t}$ and $\mathrm{S}_r^{ts_k}$ have been formulated in Eqn.7. $S_b^{s_k s_l}$ and $S_w^{s_k s_l}$ are defined by

$$\mathrm{S}_b^{s_k s_l} = \sum_{i=1}^{N_{s_k}} \sum_{j \in B_i^{s_l}} (\mathrm{P}_j^{s_l\prime} \mathrm{Q}_{ji}^{s_k s_l} - \mathrm{P}_i^{s_k\prime} \mathrm{Q}_{ij}^{s_k s_l})(\mathrm{P}_j^{s_l\prime} \mathrm{Q}_{ji}^{s_k s_l} - \mathrm{P}_i^{s_k\prime} \mathrm{Q}_{ij}^{s_k s_l})^T,$$

$$\mathrm{S}_w^{s_k s_l} = \sum_{i=1}^{N_{s_k}} \sum_{j \in W_i^{s_l}} (\mathrm{P}_j^{s_l\prime} \mathrm{Q}_{ji}^{s_k s_l} - \mathrm{P}_i^{s_k\prime} \mathrm{Q}_{ij}^{s_k s_l})(\mathrm{P}_j^{s_l\prime} \mathrm{Q}_{ji}^{s_k s_l} - \mathrm{P}_i^{s_k\prime} \mathrm{Q}_{ij}^{s_k s_l})^T.$$

Algorithm 1 Heterogeneous Transfer Discriminant-Analysis of Canonical Correlations (HTDCC)

Input:

- N_s labeled training samples $\{\mathrm{X}_i^s\}_{i=1}^{N_s}$ from the source view
- N_t labeled training samples $\{\mathrm{X}_i^t\}_{i=1}^{N_t}$ from the target view
- N_u unlabeled training samples $\{\mathrm{X}_i^u\}_{i=1}^{N_u}$ from the target view

Output:

- Projection matrices T_s and T_t.

Initialize: $\mathrm{T}_s = \mathrm{T}_t = \mathrm{I}$.
Compute the orthonormal subspaces P_i^s, P_i^t, P_i^u of X_i^s, X_i^t, X_i^u, respectively, by $\mathrm{X}\mathrm{X}^T = \mathrm{P}\Lambda\mathrm{P}^T$.
Compute the mean of orthonormal subspaces of the source-view samples by $\mathrm{P}_r^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathrm{P}_i^s$.
Compute the mean of orthonormal subspaces of the source-view samples by $\mathrm{P}_r^t = \frac{1}{N_t+N_u}(\sum_{i=1}^{N_t}\mathrm{P}_i^t + \sum_{i=1}^{N_u}\mathrm{P}_i^u)$.
Do iterate the following:
Normalize P_i^s, P_i^t, P_r^s, P_r^t to $\mathrm{P}_i^{s\prime}$, $\mathrm{P}_i^{t\prime}$, $\mathrm{P}_r^{s\prime}$, $\mathrm{P}_r^{t\prime}$ by QR-decomposition: $\mathrm{T}^T\mathrm{P} = \Phi\Delta$, $\mathrm{P}' = \mathrm{P}\Delta^{-1}$.
Do SVDs for pairs $(\mathrm{P}_i^{s\prime}, \mathrm{P}_j^{s\prime})$, $(\mathrm{P}_i^{t\prime}, \mathrm{P}_j^{t\prime})$, $(\mathrm{P}_i^{s\prime}, \mathrm{P}_j^{t\prime})$, $(\mathrm{P}_i^{t\prime}, \mathrm{P}_j^{s\prime})$, $(\mathrm{P}_r^{s\prime}, \mathrm{P}_r^{t\prime})$, respectively:
$(\mathrm{T}_s^T\mathrm{P}_i^{s\prime})^T(\mathrm{T}_s^T\mathrm{P}_j^{s\prime}) = \mathrm{Q}_{ij}^s\Lambda\mathrm{Q}_{ji}^{s\,T}$, $(\mathrm{T}_t^T\mathrm{P}_i^{t\prime})^T(\mathrm{T}_t^T\mathrm{P}_j^{t\prime}) = \mathrm{Q}_{ij}^t\Lambda\mathrm{Q}_{ji}^{t\,T}$,
$(\mathrm{T}_s^T\mathrm{P}_i^{s\prime})^T(\mathrm{T}_t^T\mathrm{P}_j^{t\prime}) = \mathrm{Q}_{ij}^{st}\Lambda\mathrm{Q}_{ji}^{st\,T}$, $(\mathrm{T}_t^T\mathrm{P}_i^{t\prime})^T(\mathrm{T}_s^T\mathrm{P}_j^{s\prime}) = \mathrm{Q}_{ij}^{ts}\Lambda\mathrm{Q}_{ji}^{ts\,T}$,
$(\mathrm{T}_s^T\mathrm{P}_r^{s\prime})^T(\mathrm{T}_t^T\mathrm{P}_r^{t\prime})) = \mathrm{Q}_r^{st}\Lambda\mathrm{Q}_r^{ts\,T}$.
Compute S_b^s, S_b^t, S_b^{st}, S_b^{ts}, S_w^s, S_w^t, S_w^{st}, S_w^{ts}, S_r^{st}, S_r^{ts} according to Eqn.7.
Compute the top-d eigenvectors $\{t_i\}_{i=1}^d$ according to Eqn.8. T_s is the first-D_s rows of $[t_1, t_2, ..., t_d]$ and T_t is the last-D_t rows of $[t_1, t_2, ..., t_d]$.
End

Then, the final solution is given by the eigen-decomposition:

$$\begin{bmatrix} \mathrm{S}_b^{s_1} & \cdots & \mathrm{S}_b^{s_K s_1} & \mathrm{S}_b^{ts_1} \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{S}_b^{s_1 s_K} & \cdots & \mathrm{S}_b^{s_K} & \mathrm{S}_b^{ts_K} \\ \mathrm{S}_b^{s_1 t} & \cdots & \mathrm{S}_b^{s_K t} & \mathrm{S}_b^{t} \end{bmatrix} t$$
$$= \lambda \begin{bmatrix} \mathrm{S}_w^{s_1} & \cdots & \mathrm{S}_w^{s_K s_1} & \mathrm{S}_w^{ts_1} + \alpha_{s_1}\mathrm{S}_r^{ts_1} \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{S}_w^{s_1 s_K} & \cdots & \mathrm{S}_w^{s_K} & \mathrm{S}_w^{ts_K} + \alpha_{s_K}\mathrm{S}_r^{ts_K} \\ \mathrm{S}_w^{s_1 t} + \alpha_{s_1}\mathrm{S}_r^{s_1 t} & \cdots & \mathrm{S}_w^{s_K t} + \alpha_{s_K}\mathrm{S}_r^{s_K t} & \mathrm{S}_w^{t} \end{bmatrix} t,$$

the optimal $\mathrm{T}_{s_1}, \ldots, \mathrm{T}_{s_K}, \mathrm{T}_t$ are constructed by the rows of the top-d eigenvectors $[t_1, t_2, \ldots, t_d]$.

IV. Multiple Source Views Adaptation for Target View

Owing to the learned common feature space between heterogeneous source and target views, the classifiers pre-trained on the source view can be effectively adapted to the target videos. Since one single source view may provide partial action knowledge, it is beneficial to leverage multiple source-view classifiers to the target-view classifier. Considering that different source views perform different correlations to the target view and different source-view classifiers make different contributions to the target-view classifier, we aim to increase the chance of adapting more related source views (i.e., positive source views) and simultaneously decrease the risk of transferring less related source views (i.e., negative source views).

In this section, a weighting learning framework is presented to assign different weights to different source views based on their relevances to the target view. The target-view classifier is actually a combination of transferred multiple source-view classifiers according to their corresponding weights. Due to the limited number of labeled data in the target view, we also utilize the unlabeled target-view data to learn the target-view classifier. Consequently, the weights of multiple source-view classifiers are learned by minimizing the prediction error of the target-view classifier on the labeled target-view training data and the loss function based on the smoothness assumption of the unlabeled target-view training data.

A. Pre-Learned Classifiers of Source Views

Because of limited labeled training samples in the target view, we resort to leveraging the pre-learned source-view classifiers to the target view. For each action class from each source view, the Multiple Kernel Learning (MKL) [23] method is adopted to train a robust classifier on the learned common feature space. We use M base kernel functions $k_m(\mathrm{X}_i, \mathrm{X}_j) = \varphi_m(\phi_m(\mathrm{X}_i))^T\varphi_m(\phi_m(\mathrm{X}_j))$, $m = 1, 2, \ldots, M$, where $\phi_m(\mathrm{X}_i)$ and $\phi_m(\mathrm{X}_j)$ represent the common features extracted from video i and video j, respectively. Given an input video X with its common feature $\phi_m(\mathrm{X})$, the final decision function of X is defined as follows:

$$f(\mathrm{X}) = \sum_{m=1}^{M} d_m \mathbf{w}_m^T \varphi_m(\phi_m(\mathrm{X})) + b, \tag{11}$$

where d_m is the linear combination coefficient of the m-th mixing kernel, with the constraints of $\sum_{m=1}^{M} d_m = 1$ and $d_m \geq 0$. $\mathbf{w}_m$ and b are the parameters of the standard SVM.

In this paper, two different types of visual features (i.e., sequence of optical-flow descriptors and sequence of SIFT descriptors) are used in the source view, and the sequence of silhouette descriptors is adopted to describe an action video in the target view. The proposed HTDCC method is used

to learn two different types of common feature spaces: one links the optical-flow based source-view feature space to the silhouette based target-view feature space; the other connects the SIFT based source-view feature space to the silhouette based target-view feature space. We use two base kernels (i.e., $M = 2$), of which each base kernel corresponds to one type of common feature. On the m-th type of common feature space between the source view and the target view, we introduce a new kernel defined by the canonical correlation between any pairwise projected samples:

$$k_m(\mathrm{X}_i, \mathrm{X}_j) = \max_{Q_{ij}, Q_{ji}} \mathbf{Tr}(\phi_m(\mathrm{X}_j) Q_{ji} Q_{ij}^T \phi_m(\mathrm{X}_i)^T). \quad (12)$$

B. Combination of Multiple Source-View Classifiers

Suppose we have G source views and one target view, the target-view classifier for an input test video X^t is defined by

$$f_t(\mathrm{X}^t) = \sum_{g=1}^{G} \beta_g f_s^g(\mathrm{X}^t), \quad (13)$$

where $\beta_g > 0$ is the weight of pre-learned classifier from the g-th source view, constrained by $\sum_{g=1}^{G} \beta_g = 1$.

The proposed weighting learning framework for multiple source views adaptation to f_t is given by

$$\min_{f_t} \Omega_r(f_t) + \lambda_l \Omega_l(f_t) + \lambda_u \Omega_u(f_t), \quad (14)$$

where $\lambda_l > 0$ and $\lambda_u > 0$ are tradeoff parameters. The details of each term in Eqn.14 are described as follows.

$\Omega_r(f_t) = \frac{1}{2}\|\boldsymbol{\beta}\|^2$ controls the complexity of the target classifier f_t, where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_G]^{\mathrm{T}}$ are the weights of all the source-view pre-learned classifiers.

$\Omega_l(f_t)$ is a loss function of the target-view classifier f_t on the labeled training data from the target view, defined as

$$\Omega_l(f_t) = \sum_{i=1}^{N_t} \| f_t(\mathrm{X}_i^t) - C_i^t \|^2, \quad (15)$$

where X_i^t is the i-th labeled target-view training sample, C_i^t is the action class label of X_i^t, and N_t is the number of labeled target-view training samples. This term enforces the decision value of f_t similar to the ground-truth label.

Since the number of labeled training data from the target view is very limited, the learning of the target classifier f_t may overfit, and the generalization ability of f_t may be degraded. Fortunately, as shown in the traditional transductive learning methods [24], unlabeled data from the target view can be employed to improve the classification performance. Thus we introduce a group loss function $\Omega_u(f_t)$ to ensure the smoothness on the unlabeled target-view data, parameterized as

$$\Omega_u(f_t) = \sum_{g=1}^{G} \beta_g \sum_{k=1, k\neq g}^{G} \sum_{i=1}^{N_u} \| f_s^g(\mathrm{X}_i^u) - f_s^k(\mathrm{X}_i^u) \|^2, \quad (16)$$

where X_i^u represents the i-th unlabeled target-view training sample and f_s^k indicates the k-th source-view classifier. This loss function guarantees that for each unlabeled target-view sample X_i^u, its decision values of different source-view classifiers should be similar to each other.

Putting all the terms together, the optimization problem in Eqn.14 can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 &+ \lambda_l \sum_{i=1}^{N_t} \| f_t(\mathrm{X}_i^t) - C_i^t \|^2 \\ &+ \lambda_u \sum_{g=1}^{G} \beta_g \sum_{k=1, k\neq g}^{G} \sum_{i=1}^{N_u} \| f_s^g(\mathrm{X}_i^u) - f_s^k(\mathrm{X}_i^u) \|^2, \\ \mathbf{s.t.} \sum_{g=1}^{G} \beta_g &= 1, \quad \beta_g > 0, \forall g. \end{aligned} \quad (17)$$

The optimization problem of Eqn.17 can be solved by a standard Quadratic Programming.

Discussion: The most related method to our fusion strategy is Weighted Canonical Correlations (WCC) method [20] and our previous work in [14]. WCC is proposed for linearly combining multi-class canonical correlations from multiple source views to generate the target-view canonical correlations for classification. It is based on the canonical correlations and has limitations to some context. In contrast, our learning framework fuses multiple source-view classifiers to build the target-view classifier. It is more general and flexible to readily incorporate different classifiers, not limited to the MKL classifiers used in this paper. In addition, the learning process in [20] only uses the limited number of labeled data in the target view, while our method also uses the unlabeled target-view data to effectively learn the target-view classifier. Different from [14] which uses SVM to learn the source-view classifiers on single feature, this paper adopts MKL method for pre-learning the source-view classifiers by fusing multiple features which benefits further improving the recognition performance.

V. Experiments

A. Dataset

We evaluate the performance of our method on the IXMAS multi-view dataset [1] which is the most popular dataset with the provided silhouette of human body for recognizing actions across different views. It consists of 11 complete action classes. Each action is executed three times by 12 subjects and recorded by 5 cameras observing the subjects from very different perspectives with the frame rate of 23fps and the frame size of 390×291 pixels. The body position and orientation are freely decided by different subjects. Figure 2 shows some action examples from five views.

In an action video, each frame is described by a feature vector and the whole video is represented by a set of sequential feature vectors. For the source view, we use two different representations: a set of sequential optical flows and a set of sequential bag-of-SIFTs, where each frame is represented by the optical flow of body region between itself and its previous frame, and is also represented by a bag of SIFTs. For the target view, we adopt another heterogeneous feature: a set of sequential silhouette images, where each frame is represented by the silhouette of human body. Both optical flow and

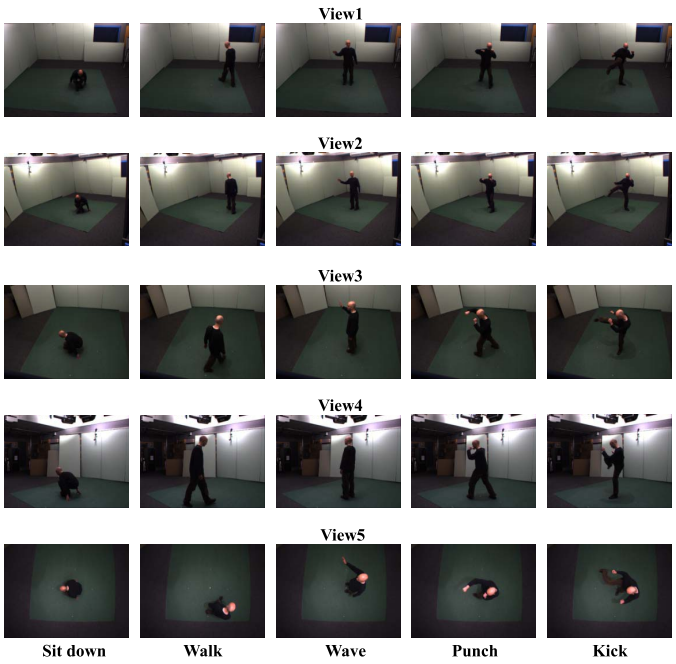


Fig. 2. Samples of frames from action videos on the IXMAS multi-view dataset.

silhouette are extracted from the body region which is obtained using the background subtraction algorithm. Each silhouette is normalized to the size of 40×80 and then converted into a 3200 dimensional vector in a raster-scan manner. The optical flow descriptor is constructed by the concatenation of four flow components with the size of $40 \times 80 \times 4$ and then converted into a 12800 dimensional vector. We extract the local SIFTs from each frame and utilize the bag-of-words model to generate the final bag-of-SIFTs vector. Since the size of codebook is 1000, the dimension of bag-of-SIFTs vector is 1000. For each set of sequential feature vectors (e.g., a set of optical flows, a set of silhouettes, and a set of bag-of-SIFTs), we fix the dimension of its linear subspace to 10.

B. Pairwise Cross-View Recognition

In this experiment, we take one view as the source view and take another different view as the target view. Both the set of sequential optical flows and the set of sequential bag-of-SIFTs are employed in the source view, and the set of sequential silhouettes is extracted in the target view. So two different types of common feature spaces between the source view and the target view should be separately learned: one connects the source-view optical flows with the target-view silhouettes ("opticalflow-silhouette"); the other links the source-view bag-of-SIFTs to the target-view silhouettes ("sift-silhouette").

To verify the effectiveness of Heterogeneous Transfer Discriminant-analysis Canonical Correlations (HTDCC) across pairwise views, we look into the recognition performances of all possible pairwise view combinations. The leave-one-subject-out cross validation strategy is employed. Specifically, for each time, we use the videos of one subject from the target view for testing, and use the videos of the remaining 11 subjects from the target view as well as all the videos from the source view as training data. For the training data, all the source-view samples and a small number of target-view samples are labeled.

Several state-of-the-art methods of transfer learning on heterogeneous features [12], [13], [16]–[18] are compared with our method under the leave-one-subject-out cross validation strategy. Since KCCA [12] and HeMap [16] require the correspondence between the source data and the target data, for each time of the cross-validation we use the label information to align the training samples from two views. The setting of training and test data in DAMA [13], ARC-t [17] and HFA [18] is the same to that in our method. For all these five methods, we convert the linear subspace of a set of sequential feature vectors (i.e., the orthonormal basis matrix) into a long vector by concatenating the columns to represent an action video sample. For KCCA, HeMap and DAMA, two types of common feature spaces are learned between the source view and the target view (i.e., "opticalflow-silhouette" and "sift-silhouette"), which is similar to our method. After learning the projection matrices, we apply the same MKL method to train their final classifiers using the projected training data on the two types of common feature spaces. For ARC-t, we learn two asymmetric transformation metrics: one is between source-view optical flows and target-view silhouettes, and the other is between source-view bag-of-SIFTs and target-view silhouettes. Accordingly, two base kernel matrices are constructed on the two asymmetric transformation metrics, respectively, and then MKL is also applied to learn the final classifier. For HFA, two types of common feature spaces are learned between the source view and the target view (i.e., "opticalflow-silhouette" and "sift-silhouette"). Then the augmented feature for the source view is constructed by concatenating the augmented optical flow and bag-of-SIFTs, and the target-view augmented feature is the augmented silhouette representation. For all these methods, we set the regularization parameter $C = 1$ in SVM and use the linear kernel for fair comparison. As we only have a very limited number of labeled training samples in the target view, the cross-validation technique can not be effectively employed to determine the optimal parameters. Instead, for our HTDCC method, we empirically choose the best parameter α from $\{1, 10, 100\}$ based on their results on the test data. For other methods, we tune their parameters from $\{0.01, 0.1, 1, 10, 100\}$ and report their best results.

Table V demonstrates the recognition results of different related methods using the fraction of labeled target-view training data of $3/11$. It is interesting to notice that HTDCC outperforms other methods, which clearly demonstrates the effectiveness of our method on cross-view action recognition on heterogeneous features. Compared with KCCA and HeMap, HTDCC is able to learn a common feature space with discriminative ability by using the label information of training data. HTDCC outperforms DAMA, possibly due to the lack of the strong manifold structure on this dataset. The explanation for the better performance of HTDCC than ARC-t and HFA may be that HTDCC utilizes unlabeled target-view training data and incorporates the minimization of the distribution mismatch between source and target views in learning the common feature space.

TABLE V

CROSS-VIEW ACTION RECOGNITION ACCURACIES (%) OF DIFFERENT HETEROGENEOUS TRANSFER LEARNING METHODS ON THE IXMAS DATASET WHEN THE FRACTION OF THE LABELED TARGET-VIEW SAMPLES IS 3/11. EACH ROW IS A SOURCE VIEW AND EACH COLUMN IS A TARGET VIEW. THE SEVEN ACCURACY NUMBERS IN A TUPLE ARE THE AVERAGE RECOGNITION ACCURACY OF KCCA [12], HeMap [16], DAMA [13], ARC-t [17], HFA [18], OUR PREVIOUS WORK [14] AND OUR HTDCC METHOD, RESPECTIVELY. DUE TO THE LIMITED SPACE, THE ENTIRE TABLE IS SPLIT INTO THREE SUBTABLES

	Target view1	Target view2
Source view1		(36.11, 34.03, 35.42, 34.03, 30.56,47.2, **52.08**)
Source view2	(40.28, 32.64, 37.50, 31.94, 29.86,44.4, **58.33**)	
Source view3	(36.11, 34.72, 36.11, 33.33, 28.47,45.8, **57.64**)	(37.50, 34.03, 36.11, 34.03, 30.56,48.6, **53.47**)
Source view4	(37.50, 35.42, 35.42, 29.86, 32.64,43.8, **54.17**)	(38.19, 34.03, 34.72, 34.72, 30.56,41.7, **54.86**)
Source view5	(40.28, 35.42, 36.11, 33.33, 29.86,41.0, **54.86**)	(39.58, 34.72, 34.03, 31.94, 31.94,45.1, **60.42**)
Average	(38.54, 34.55, 36.28, 32.12, 30.21,43.8, **56.25**)	(37.85, 34.20, 35.07, 33.68, 30.90,45.7, **55.21**)

	Target view3	Target view4
Source view1	(31.94, 29.17, 27.78, 27.08, 26.39, 41.0, **50.69**)	(34.03, 30.56, 32.64, 30.56, 21.53, 61.8, **61.81**)
Source view2	(29.86, 24.31, 27.78, 27.08, 25.69, 44.4, **50.69**)	(38.19, 30.56, 34.03, 34.03, 24.31, 57.6, **61.11**)
Source view3		(36.11, 31.94, 34.03, 29.17, 25.69, 54.2, **64.58**)
Source view4	(30.56, 29.17, 27.08, 27.78, 27.08, 43.1, **42.36**)	
Source view5	(31.25, 29.17, 29.17, 27.08, 27.08, 41.0, **49.31**)	(37.50, 31.25, 31.25, 30.56, 26.39, 53.5, **59.03**)
Average	(30.90, 27.95, 27.95, 27.26, 26.56, 42.4, **48.26**)	(36.46, 31.08, 32.99, 31.08, 24.48, 56.8, **61.63**)

Methods	Target view5
Source view1	(23.61, 15.97, 18.75, 13.89, 15.28, 32.6, **42.36**)
Source view2	(19.44, 21.53, 18.75, 14.58, 17.36, 35.4, **41.67**)
Source view3	(20.83, 25.00, 18.06, 14.58, 15.28, 37.5, **36.81**)
Source view4	(20.83, 22.22, 17.36, 14.58, 15.28, 31.3, **38.89**)
Source view5	
Average	(21.18, 21.18, 18.23, 14.41, 15.80, 34.2, **39.93**)

TABLE VI

CROSS-VIEW RECOGNITION ACCURACIES (%) USING DIFFERENT FRACTIONS OF LABELED TRAINING SAMPLES FROM THE TARGET VIEW. EACH ROW IS A SOURCE VIEW AND EACH COLUMN IS A TARGET VIEW. THE FOUR ACCURACY NUMBERS IN A TUPLE ARE THE AVERAGE RECOGNITION ACCURACY USING THE FRACTION OF 0, 1/11, 2/11, AND 3/11 RESPECTIVELY. DUE TO THE LIMITED SPACE, THE ENTIRE TABLE IS SPLIT INTO TWO SUBTABLES

	Target view1	Target view2	Target view3
Source view1		(8.33, 25.69, 50.00, 52.08)	(8.33, 25.69, 45.83, 50.69)
Source view2	(8.33, 31.25, 44.44, 58.33)		(9.03, 26.39, 43.06, 50.69)
Source view3	(8.33, 34.03, 36.81, 57.64)	(8.33, 29.86, 44.44, 53.47)	
Source view4	(8.33, 27.78, 42.36, 54.17)	(8.33, 25.00, 46.53, 54.86)	(9.03, 24.31, 41.67, 42.36)
Source view5	(8.33, 27.08, 40.28, 54.86)	(7.64, 24.31, 36.81, 60.42)	(8.33, 21.53, 40.28, 49.31)
Average	(8.33, 30.04, 40.97, 56.25)	(8.16, 26.22, 44.45, 55.21)	(8.68, 24.48, 42.71, 48.26)

	Target view4	Target view5
Source view1	(9.03, 38.89, 56.94, 61.81)	(8.33, 20.14, 31.94, 42.36)
Source view2	(9.72, 34.72, 56.94, 61.11)	(7.64, 18.06, 35.42, 41.67)
Source view3	(9.03, 38.19, 55.56, 64.58)	(8.33, 21.53, 34.72, 36.81)
Source view4		(9.03, 18.75, 29.86, 38.89)
Source view5	(8.33, 35.42, 56.25, 59.03)	
Average	(9.03, 36.81, 56.42, 61.63)	(8.33, 19.62, 32.99 39.93)

We vary the fraction of the labeled samples from the target view in increments of $1/11$ from $0/11$ to $3/11$ and report the results of HTDCC for all the pairwise combinations of the source and target views in Table VI, from which a substantial improvement is observed with the increment of labeled target-view data. Table VII illustrates the mean accuracies of different methods with the increasing number of labeled target-view samples. It is obvious that our HTDCC method generally outperforms all other methods according to the mean recognition accuracy.

C. Multiple Source Views Adaptation

To exploit the benefits of adapting multiple source views for the target recognition, we select one view as the target

TABLE VII

MEAN RECOGNITION ACCURACIES (%) OF DIFFERENT METHODS USING DIFFERENT FRACTIONS OF LABELED TRAINING SAMPLES FROM THE TARGET VIEW

Fraction	KCCA	HeMap	DAMA	ARC-t	HFA	HTDCC
0	-	-	-	-	-	**8.51**
1/11	27.85	22.50	**28.06**	23.47	20.56	27.43
2/11	31.39	28.54	29.86	26.60	24.38	**43.51**
3/11	33.00	29.79	30.10	27.71	25.59	**52.26**

TABLE VIII

COMPARISON OF DIFFERENT MULTIPLE SOURCE VIEWS ADAPTATION METHODS ON THE RECOGNITION ACCURACY (%) FOR EACH TARGET VIEW

Methods	Target view1	Target view2	Target view3	Target view4	Target view5	Average
$\lambda_l = \lambda_u = 0$	58.33	55.56	54.86	66.67	40.97	55.28
$\lambda_l = 0.1, \lambda_u = 0$	59.72	56.25	54.86	67.36	40.97	55.83
$\lambda_l = 0, \lambda_u = 0.1$	65.97	61.81	61.11	71.53	47.22	61.49
Our method	**67.36**	**62.50**	**62.50**	**72.22**	**49.31**	**62.78**

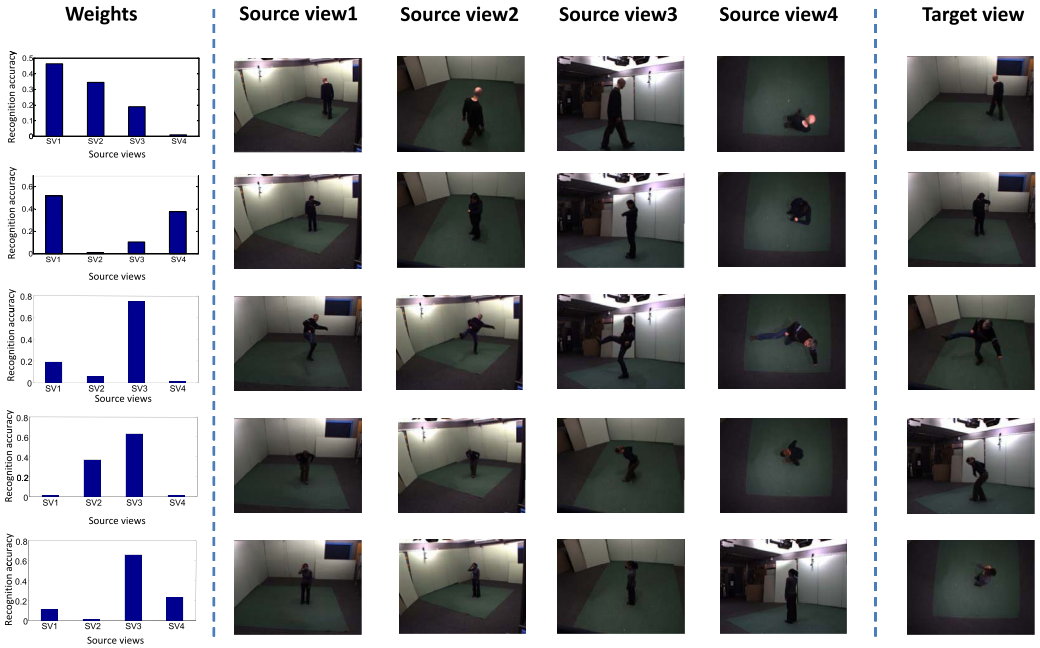


Fig. 3. Examples of the learned combination weights of multiple source views. For each target view, its classifiers are constructed by the combination of transferred four source views based on the weights shown by vertical axis of histograms. In the histograms, the "SV" is short for source view.

view and use the rest four views as the source views. The pre-learned MKL classifiers from four source views are fused via the weighting learning framework presented in Section IV. The parameters λ_l and λ_u are empirically set to $\lambda_l = \lambda_u = 0.1$ by choosing from $\{0.1, 1, 10\}$ according to the testing performances. To verify the effectiveness of assigning different weights to different source views, we compare our method with a baseline fusion method that uses equal weights $\beta_g = 1/G$ (i.e., $\lambda_l = \lambda_u = 0$). Further, to evaluate the contribution of the unlabeled target-view samples to learning the target classifier, we compare the results between our method and another baseline method which excludes the loss function term defined on the unlabeled target-view training data in Eqn.14 (i.e., $\lambda_u = 0$, $\lambda_l = 0.1$). To investigate the effect of the labeled target-view data, we also report the results when excluding the loss function of the labeled target-view training data in Eqn.14 (i.e., $\lambda_u = 0.1$, $\lambda_l = 0$). From the results shown in Table VIII, it is interesting to observe that: (1) the adaptation of multiple source views achieves better results than each single source view because one single view has limited discriminative and descriptive abilities compared with multiple source views; (2) assigning different weights to different source views can benefit improving the recognition performance thanks to the selection of more related source-view classifiers transferred to the target-view classifier; (3) employing unlabeled target-view training data is able to significantly improve the performance by capturing the smoothness of source-view classifiers on them.

Fig.3 shows some examples of learned weights of multiple source views. The left column represents the weights,

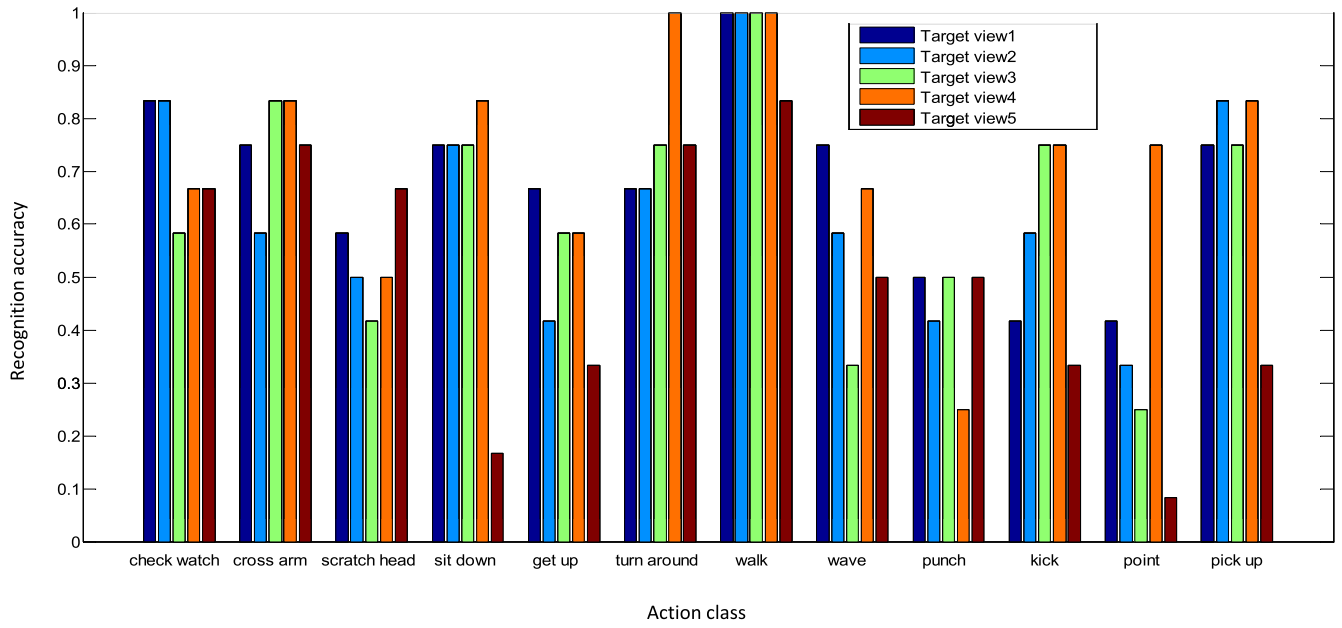


Fig. 4. Recognition performance of multiple source views adaptation on each action class.

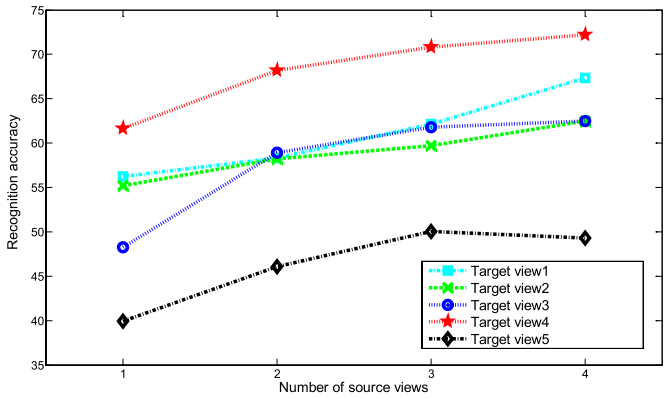


Fig. 5. Results of combining different numbers of source views.

the middle columns indicate the multiple source views, and the right column demonstrates the target view. Each row represents one example. We can notice that the more related the source view is to the target view, the higher the learned weight becomes. Taking the first row for example, the "Source view1" is more related to the "Target view", and its weight is higher than that of other source views. We also report the recognition accuracy of each action class in Fig.4, which shows that the task of transferring source-view classifiers is very hard for some actions and some views. For example, the recognition accuracies of "sit down" and "point" are very low in Target view 5. One of the reasons might be that the majority of the body motions is occluded by the head in this view.

To further evaluate our fusion method using different numbers of source views, we report the recognition accuracies of combining different numbers of source views for each target view in Fig.5. It is obvious that for most cases the performance improves with the increasing number of source views, because more source views can transfer more information to the target view. For Target view 5, the fusion of four source views achieves comparable results with that of three source views. The possible reason is that Target view 5 is less related to the other four source views, so combining more source views may not constantly improve the performance on the target view. It is also interesting to observe that our multiple source-views fusion method performs better than single source view even when there are only two source views.

VI. Conclusions

We have proposed a novel Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC) method for cross-view action recognition. Our method neither requires the same type of feature shared by different views nor limits to any corresponding action instances in different views. Two projection matrices are learned to respectively map the data from the source view and the target view to a common feature space, by simultaneously minimizing the canonical correlations of inter-class samples, maximizing the canonical correlations of intra-class samples, and reducing the data distribution mismatch between source and target views. Moreover, a weighting learning framework for multiple source views adaptation is presented to flexibly combine multiple action classifiers from multiple source views to construct the target-view classifiers. Extensive experiments have shown the effectiveness of our method. In the future, we plan to apply our method to other applications such as cross-view object recognition and face recognition from videos.

References

[1] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.

[2] P. Yan, S. M. Khan, and M. Shah, "Learning 4D action feature models for arbitrary view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[3] A. Yilma and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 150–157.

[4] Y. Shen and H. Foroosh, "View-invariant action recognition using fundamental ratios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–6.

[5] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[6] M. Lewandowski, D. Makris, and J.-C. Nebel, "View and style-independent action manifolds for human activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 547–560.

[7] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 489–496.

[8] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[9] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 154–166.

[10] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3209–3216.

[11] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–2.

[12] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[13] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1541–1546.

[14] X. Wu, H. Wang, C. Liu, and Y. Jia, "Cross-view action recognition over heterogeneous feature spaces," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 609–616.

[15] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2855–2862.

[16] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *Proc. IEEE 10th Int. Conf. Data Mining*, Dec. 2010, pp. 1049–1054.

[17] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1785–1792.

[18] L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2012, pp. 711–718.

[19] T.-K. Kim, K.-Y. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[20] X. Wu, C. Liu, and Y. Jia, "Transfer discriminant-analysis of canonical correlations for view-transfer action recognition," in *Proc. Pacific-Rim Conf. Multimedia*, 2012, pp. 444–454.

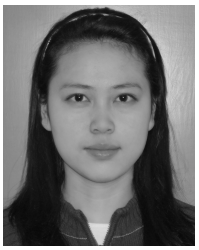
[21] X. Wu, Y. Jia, and W. Liang, "Incremental discriminant-analysis of canonical correlations for action recognition," *Pattern Recognit.*, vol. 43, no. 12, pp. 4190–4197, Dec. 2010.

[22] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[23] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 1065–1072.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

Xinxiao Wu received the B.S. degree from the Nanjing University of Information Science and Technology, in 2005, and the Ph.D. degree from the Beijing Institute of Technology, China, in 2010. She is currently an Associate Professor with the Beijing Institute of Technology. Her research interests include computer vision, machine learning, and video content analysis.



Han Wang received the Ph.D. degree in computer science from the Beijing Institute of Technology. She is currently a Lecturer with the School of Information and Technology, Beijing Forest University. Her research interests include machine learning, computer vision, multimedia retrieval, video analysis, and event recognition.

Cuiwei Liu received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2009, where she is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology.

Yunde Jia received the B.S., M.S., and Ph.D. degrees in mechatronics from the Beijing Institute of Technology (BIT), in 1983, 1986, and 2000, respectively. He is currently a Professor of Computer Science at BIT, and serves as the Director of the Beijing Laboratory of Intelligent Information Technology. He has previously served as the Executive Dean of the School of Computer Science at BIT from 2005 to 2008. He was a Visiting Scientist at Carnegie Mellon University from 1995 to 1997, and a Visiting Fellow with Australian National University, in 2011. His current research interests include computer vision, media computing, and intelligent systems.