2013 IEEE International Conference on Computer Vision

Cross-view Action Recognition over Heterogeneous Feature Spaces

Xinxiao Wu Han Wang Cuiwei Liu Yunde Jia
Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology
Beijing 100081, P.R. China
{wuxinxiao,wanghan,liucuiwei,jiayunde}@bit.edu.cn

Abstract

In cross-view action recognition, "what you saw" in one view is different from "what you recognize" in another view. The data distribution even the feature space can change from one view to another due to the appearance and motion of actions drastically vary across different views. 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 learn a discriminative common feature space for linking source and target views to transfer knowledge between them. Two projection matrices that respectively map data from source and target views into the common space are optimized via simultaneously minimizing the canonical correlations of inter-class samples and maximizing the intra-class canonical correlations. Our model is neither restricted to corresponding action instances in the two views nor restricted to the same type of feature, and can handle only a few or even no labeled samples available in the target view. To reduce the data distribution mismatch between the source and target views in the common feature space, a non-parametric criterion is included in the objective function. We additionally propose a joint weight learning method to fuse multiple source-view action classifiers for recognition in the target view. Different combination weights are assigned to different source views, with each weight presenting how contributive the corresponding source view is to the target view. The proposed method is evaluated on the IXMAS multi-view dataset and achieves promising results.

1. 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 performance in another view. One possible solution [14, 18, 19, 11] 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 [4] and the view-style independent manifold representation [7]. Wu *et al*. [15] proposed a latent kernelized structural SVM for view-invariant action recognition where the view is modeled as a latent variable and inferred during both training and testing stage. Some other methods [17, 3] learn a separate model for each action class in each view, however, it is difficult to collect sufficient labeled samples for each view to cover all the action classes. Recently, transfer learning based methods [2, 9, 20] 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.

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 source view and target view, and can handle the heterogeneous action representations in the two views. Two projection matrices are learned to respectively map the source and target views to a common space, by simultaneously minimizing the canonical correlations of inter-class samples, maximizing the canonical correlations of intra-class samples, and minimizing the canonical correlation between the means of source-view and target-view samples. Instead of requiring the corresponding observation of the same action instance from source and target views, our method explores how to take advantage of label information to learn a common feature space with discrimination.

In order to adapt multiple source views to the target view,

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.81

we additionally present a joint weight learning method to effectively combine multiple transferred source-view classifiers to generate the target-view classifiers. Since different source views perform different relations with the target view, for each source view, a specific weight is adopted to represent its closeness to the target view.

2. Related work

From the perspective of cross-view action recognition, some work [2, 9, 20] is closely related to our approach. Farhadi *et al*. [2] 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*. [9] 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*. [20] presented a transferable dictionary pair consisting of two dictionaries that correspond to the source and target 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 *et al*. [8] proposed "virtual views" to connect action descriptors between 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 [8], our method can handle the cross-view action recognition when the actions are represented by heterogeneous features in source and target views.

From the perspective of transfer learning, our work is also related to the methods [10, 12, 13, 6] which find a "good" common feature space for source and target domains. Taylor and Cristianini [10] learned a common feature space by maximizing the correlation between the source and target training data without any label information. Shi *et al*. [12] 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 [10] and [12], our method does not require the sample correspondence between source and target domains. Moreover, our method utilizes the label information to discover a common feature space with more discrimination. 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 and minimizing the inter-domain similarity. Their method assumes the manifold structure on the dataset. Kulis *et al*. [6] proposed to learn an asymmetric kernel transformation to transfer feature knowledge between source and target domains.

3. Heterogeneous transfer discriminant-analysis of canonical correlations

3.1. Problem statement

In this work, each action sample is represented by an orthogonal linear subspace of sequential image features. Denote $\mathrm{X} = [x_1, x_2, ..., x_M] \in \mathbb{R}^{D\times M}$ as the sequential image features of an action sample, where $x_i \in \mathbb{R}^D$ represents the i-th image feature. The orthogonal linear subspace of X is denoted by $\mathrm{P} \in \mathbb{R}^{D\times m}$ s.t. $\mathrm{X}\mathrm{X}^T = \mathrm{P}\Lambda\mathrm{P}^T$, where Λ is the m largest eigenvalues and P is the corresponding eigenvectors. Given a large number of labeled training samples from the source view $\{\mathrm{X}_i^s|_{i=1}^{N_s}\}$ with $\mathrm{X}_i^s \in \mathbb{R}^{D_s\times M_i}$, a limited (even no) number of labeled training samples from the target view $\{\mathrm{X}_i^t|_{i=1}^{N_t}\}$ with $\mathrm{X}_i^t \in \mathbb{R}^{D_t\times M_i}$, and some unlabeled samples from the target view $\{\mathrm{X}_i^u|_{i=1}^{N_u}\}$ with $\mathrm{X}_i^u \in \mathbb{R}^{D_t\times M_i}$, where 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 for respectively mapping the source and target views to the common space.

3.2. Background

Discriminant-Analysis of Canonical Correlations (DCC) [5] learns a projection matrix by maximizing canonical correlations of within-class samples and minimizing canonical correlations of between-class samples. Assume N training samples are given as $\{\mathrm{X}_i|_{i=1}^N\}$, where X_i belongs to one action class denoted by C_i. The discriminative projection matrix $\mathrm{T} = [t_1, t_2, ..., t_m] \in \mathbb{R}^{D\times m}$ defined by $\mathrm{Y} = \mathrm{T}^T\mathrm{X}$, where $m \leq D$ and $|t_i| = 1$, to make the projected samples more discriminative using canonical correlations. Orthonormal subspaces of the projected data are given by $\mathrm{Y}\mathrm{Y}^T = (\mathrm{T}^T\mathrm{X})(\mathrm{T}^T\mathrm{X})^T = (\mathrm{T}^T\mathrm{P})\Lambda(\mathrm{T}^T\mathrm{P})^T$. The matrix P is normalized to P' so that the columns of $\mathrm{T}^T\mathrm{P}'$ are orthonormal. The similarity of two projected samples is defined as the sum of canonical correlations $\mathrm{F}_{ij} = \max_{\mathrm{Q}_{ij},\mathrm{Q}_{ji}} \mathrm{Tr}(\mathrm{T}^T\mathrm{P}'_j\mathrm{Q}_{ji}\mathrm{Q}_{ij}^T\mathrm{P}_i'^T\mathrm{T})$, where the solution of Q_{ij} and Q_{ji} is given by the SVD computation $(\mathrm{T}^T\mathrm{P}'_i)^T(\mathrm{T}^T\mathrm{P}'_j) = \mathrm{Q}_{ij}\Lambda\mathrm{Q}_{ij}^T$. 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

$$\mathrm{T} = \arg\max_{\mathrm{T}} \frac{E_w(\mathrm{T})}{E_b(\mathrm{T})}, \tag{1}$$

where $E_w(\mathrm{T}) = \sum_{i=1}^{N}\sum_{k\in W_i}\mathrm{F}_{ik}$ and $E_b(\mathrm{T}) = \sum_{i=1}^{N}\sum_{l\in B_i}\mathrm{F}_{il}$. The two index sets $W_i = \{j|C_j = C_i\}$ and $B_i = \{j|C_j \neq C_i\}$, respectively, denote the intra-class and inter-class samples for a given sample of class C_i.

Transfer Discriminant-Analysis of Canonical Correlations (TDCC) [16] is the extension of DCC for handling the situation when the training and testing 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.

3.3. Learning on heterogeneous feature spaces

Our goal is to extend [16] to a more general case when the training data and testing data are drawn from different views with heterogeneous features. Two projection matrices are learned to respectively map the source view and 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 data distributions of 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 $\mathrm{T}_s = [t_{s,1}, t_{s,2}, ..., t_{s,d}] \in \mathbb{R}^{D_s\times d}$ 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 projected P_i^s is $\mathrm{T}_s^T\mathrm{P}_i^{s\prime}$ 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 $\mathrm{T}_t = [t_{t,1}, t_{t,2}, ..., t_{t,d}] \in \mathbb{R}^{D_t\times d}$ is defined by $\mathrm{Y}_i^t = \mathrm{T}_t^T\mathrm{X}_i^t$. Let P_i^t be the orthonormal subspace of X_i^t; and the projected representation of P_i^t is $\mathrm{T}_t^T\mathrm{P}_i^{t\prime}$ where $\mathrm{P}_i^{t\prime}$ indicates the normalization of P_i^t.

The learning framework of Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC) is formulated as:

$$\max_{\mathrm{T}_s,\mathrm{T}_t} \frac{E_w(\mathrm{T}_s,\mathrm{T}_t) + \alpha E_r(\mathrm{T}_s,\mathrm{T}_t)}{E_b(\mathrm{T}_s,\mathrm{T}_t)}. \tag{3}$$

$E_w(\mathrm{T}_s,\mathrm{T}_t) = \sum_{i=1}^{N_s}\sum_{j\in W_i^s}\mathrm{F}_{ij}^s + \sum_{i=1}^{N_l}\sum_{j\in W_i^t}\mathrm{F}_{ij}^t + \sum_{i=1}^{N_s}\sum_{j\in W_i^{st}}\mathrm{F}_{ij}^{st} + \sum_{i=1}^{N_l}\sum_{j\in W_i^{ts}}\mathrm{F}_{ij}^{ts}$ represents the similarities of intra-class training samples from both source and target views. $E_b(\mathrm{T}_s,\mathrm{T}_t) = \sum_{i=1}^{N_s}\sum_{j\in B_i^s}\mathrm{F}_{ij}^s + \sum_{i=1}^{N_l}\sum_{j\in B_i^t}\mathrm{F}_{ij}^t + \sum_{i=1}^{N_s}\sum_{j\in B_i^{st}}\mathrm{F}_{ij}^{st} + \sum_{i=1}^{N_l}\sum_{j\in B_i^{ts}}\mathrm{F}_{ij}^{ts}$ represents the similarities of inter-class training samples from both source and target views. F_{ij}^s represents the canonical correlation of two projected samples from the source view and F_{ij}^t represents the canonical correlation of two projected samples from the target view. Both F_{ij}^{st} and F_{ij}^{ts} represent the canonical correlations of two projected samples of which one sample is from the source view and the other sample is from the target view. They are parameterized as follows:

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

with the solutions:

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

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 source view for a given target-view sample of class C_i^t.

$E_r(\mathrm{T}_s,\mathrm{T}_t) = \mathrm{F}_r^{st} + \mathrm{F}_r^{ts}$ represents the similarity between the projected source-view mean sample and the projected target-view mean sample, where $\mathrm{F}_r^{st} = \max_{\mathrm{Q}_r^{st},\mathrm{Q}_r^{ts}} \mathbf{Tr}(\mathrm{T}_t^T\mathrm{P}_r^{t\prime}\mathrm{Q}_r^{ts}\mathrm{Q}_r^{st\,T}\mathrm{P}_r^{s\prime T}\mathrm{T}_s)$ and $\mathrm{F}_r^{ts} = \max_{\mathrm{Q}_r^{ts},\mathrm{Q}_r^{st}} \mathbf{Tr}(\mathrm{T}_s^T\mathrm{P}_r^{s\prime}\mathrm{Q}_r^{st}\mathrm{Q}_r^{ts\,T}\mathrm{P}_r^{t\prime T}\mathrm{T}_t)$ by $(\mathrm{T}_s^T\mathrm{P}_r^{s\prime})^T(\mathrm{T}_t^T\mathrm{P}_r^{t\prime}) = \mathrm{Q}_r^{st}\Lambda\mathrm{Q}_r^{ts\,T}$. $\mathrm{P}_r^{s\prime}$ is the normalized orthonormal subspace of the mean of source-view training samples $\mathrm{X}_r^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathrm{X}_i^s$, and $\mathrm{P}_r^{t\prime}$ is the normalized orthonormal subspace of the mean of target-view training samples $\mathrm{X}_r^t = \frac{1}{N_t+N_u}(\sum_{i=1}^{N_t}\mathrm{X}_i^t + \sum_{i=1}^{N_u}\mathrm{X}_i^u)$.

By the linear algebra transformation $A^TB = \mathbf{I} - (A - B)^T(A - B)/2$, we can rewrite the objective function in Eqn.3 as

$$\max_{\mathrm{T}_s,\mathrm{T}_t} \frac{\mathbf{Tr}(\begin{bmatrix}\mathrm{T}_s\\ \mathrm{T}_t\end{bmatrix}^T \begin{bmatrix}\mathrm{S}_b^s & \mathrm{S}_b^{ts}\\ \mathrm{S}_b^{st} & \mathrm{S}_b^t\end{bmatrix} \begin{bmatrix}\mathrm{T}_s\\ \mathrm{T}_t\end{bmatrix})}{\mathbf{Tr}(\begin{bmatrix}\mathrm{T}_s\\ \mathrm{T}_t\end{bmatrix}^T \begin{bmatrix}\mathrm{S}_w^s & \mathrm{S}_w^{ts} + \alpha\mathrm{S}_r^{ts}\\ \mathrm{S}_w^{st} + \alpha\mathrm{S}_r^{st} & \mathrm{S}_w^t\end{bmatrix} \begin{bmatrix}\mathrm{T}_s\\ \mathrm{T}_t\end{bmatrix})}, \quad (4)$$

where

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

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

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

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

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

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

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

$$S_w^{st} = \sum_{i=1}^{N_s} \sum_{j \in W_i^{st}} (P_j^{t\prime}Q_{ji}^{st} - P_i^{s\prime}Q_{ij}^{st})(P_j^{t\prime}Q_{ji}^{st} - P_i^{s\prime}Q_{ij}^{st})^T,$$

$$S_r^{ts} = (P_r^{s\prime}Q_r^{ts} - P_r^{t\prime}Q_r^{ts})(P_r^{s\prime}Q_r^{ts} - P_r^{t\prime}Q_r^{ts})^T,$$

$$S_r^{st} = (P_r^{t\prime}Q_r^{st} - P_r^{s\prime}Q_r^{st})(P_r^{t\prime}Q_r^{st} - P_r^{s\prime}Q_r^{st})^T.$$

Finally, by the eigen-decomposition

$$\begin{bmatrix}\mathrm{S}_b^s & \mathrm{S}_b^{ts}\\ \mathrm{S}_b^{st} & \mathrm{S}_b^t\end{bmatrix} t = \lambda \begin{bmatrix}\mathrm{S}_w^s & \mathrm{S}_w^{ts} + \alpha\mathrm{S}_r^{ts}\\ \mathrm{S}_w^{st} + \alpha\mathrm{S}_r^{st} & \mathrm{S}_w^t\end{bmatrix} t, \quad (5)$$

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 $[t_1, t_2, ..., t_d]$.

We use an iterative optimization algorithm to find the optimized projection matrices T_s and T_t. With the identity matrix I as the initial values of T_s and T_t, the detailed algorithm of HTDCC is listed in Algorithm 1. Once the optimal T_s and T_t are found, the similarity of any two action samples is measured by first mapping them to the common space and then computing the canonical correlations between them in the common space. We apply SVM to train a classifier for each action class by using the projected labeled training data from both source and target views. For SVM, we introduce a kernel based on the similarity between any pairwise samples in the learned common space.

4. Multiple source views combination

Since single source view may provide partial action knowledge, it is beneficial to combine multiple source-view classifiers for improving the recognition performance in the target view. Different source views perform different correlations to the target view, and action classifiers from different source views will make different contributions to the target classifiers. Therefore, we aim to increase the chance of selecting 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 paper, a joint weight learning framework is proposed to assign different combination weights to different source views based on their relevances to the target view. The target classifier is actually a combination of transferred multiple source classifiers according to the corresponding weights. Considering the limited number of labeled samples in the target view, we also utilize the unlabeled target data to learn the target-view classifier. Consequently, the weights of multiple source-view classifiers are learned by minimizing the loss function of the target-view classifier on the labeled target-view samples and the loss function based on the smoothness assumption of the unlabeled target-view samples.

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

$$f_t(\mathrm{X}^t) = \sum_{g=1}^{G} \beta_g f_{s,g}(\mathrm{X}^t), \quad (6)$$

where $\beta_g > 0$ is the weight for the g-th source view, constrained by $\sum_{g=1}^{G} \beta_g = 1$. The proposed learning framework is given by

$$\min_{f_t} \Omega(f_t) + \lambda_l \Omega_l(f_t) + \lambda_u \Omega_u(f_t), \quad (7)$$

where $\lambda_l > 0$ and $\lambda_u > 0$ are tradeoff parameters. The details of each term in Eqn.7 are described as follows.

$\Omega(f_t) = \frac{1}{2}\|\boldsymbol{\beta}\|^2$ controls the complexity of the target classifier f_t, where $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_G]^{\mathrm{T}}$.

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

$$\Omega_l(f_t) = \sum_{i=1}^{N_t} \|f_t(\mathrm{X}_i^t) - C_i^t\|^2, \quad (8)$$

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 $\mathrm{T}_s \in \mathbb{R}^{D_s \times d}$ and $\mathrm{T}_t \in \mathbb{R}^{D_t \times d}$
Initialize: $\mathrm{T}_s = \mathrm{T}_t = \mathrm{I}$.
1. Compute the mean of source-view samples by $\mathrm{X}_r^s = \frac{1}{N_s}\sum_{i=1}^{N_s} \mathrm{X}_i^s$.
2. Compute the mean of target-view samples by $\mathrm{X}_r^t = \frac{1}{N_t+N_u}(\sum_{i=1}^{N_t} \mathrm{X}_i^t + \sum_{i=1}^{N_u} \mathrm{X}_i^u)$.
3. Compute the orthonormal basis matrices P_i^s, P_i^t, P_r^s, P_r^t of X_i^s, X_i^t, X_r^s, X_r^t, respectively, by $\mathrm{X}\mathrm{X}^T = \mathrm{P}\Lambda\mathrm{P}^T$.
4. **Do iterate the following**:
5. Normalize P_i^s, P_i^t, P_r^s and P_r^t to $\mathrm{P}_i^{s\prime}$, $\mathrm{P}_i^{t\prime}$, $\mathrm{P}_r^{s\prime}$ and $\mathrm{P}_r^{t\prime}$ by QR-decomposition: $\mathrm{T}^T\mathrm{P} = \Phi\Delta$, $\mathrm{P}' = \mathrm{P}\Delta^{-1}$.
6. For pairs $(\mathrm{P}_i^{s\prime}, \mathrm{P}_j^{s\prime})$, $(\mathrm{P}_i^{t\prime}, \mathrm{P}_j^{t\prime})$, $(\mathrm{P}_i^{s\prime}, \mathrm{P}_j^{t\prime})$ and $(\mathrm{P}_i^{t\prime}, \mathrm{P}_j^{s\prime})$, respectively, do SVDs:
$(\mathrm{T}_s^T\mathrm{P}_i^{s\prime})^T(\mathrm{T}_s^T\mathrm{P}_j^{s\prime}) = \mathrm{Q}_{ij}^s\Lambda\mathrm{Q}_{ji}^{s\,T}$, $(\mathrm{T}_t^T\mathrm{P}_i^{t\prime})^T(\mathrm{T}_t^T\mathrm{P}_j^{t\prime}) = \mathrm{Q}_{ij}^t\Lambda\mathrm{Q}_{ji}^{t\,T}$,
$(\mathrm{T}_s^T\mathrm{P}_i^{s\prime})^T(\mathrm{T}_t^T\mathrm{P}_j^{t\prime}) = \mathrm{Q}_{ij}^{st}\Lambda\mathrm{Q}_{ji}^{st\,T}$, $(\mathrm{T}_t^T\mathrm{P}_i^{t\prime})^T(\mathrm{T}_s^T\mathrm{P}_j^{s\prime}) = \mathrm{Q}_{ij}^{ts}\Lambda\mathrm{Q}_{ji}^{ts\,T}$.
7. For $\mathrm{P}_r^{s\prime}$, $\mathrm{P}_r^{t\prime}$, do SVD: $(\mathrm{T}_s^T\mathrm{P}_r^{s\prime})^T(\mathrm{T}_t^T\mathrm{P}_r^{t\prime}) = \mathrm{Q}_r^{st}\Lambda\mathrm{Q}_r^{ts\,T}$.
8. 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}, and S_r^{ts} according to Eqn.4.
9. Compute the top-d eigenvectors $\{t_i\}_{i=1}^d$ according to Eqn.5.
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 $[t_1, t_2, ..., t_d]$.
10. **End**

where X_i^t is the i-th labeled training sample from the target view, C_i^t is the action class label of X_i^t, and N_t is the number of labeled target-view training samples.

$\Omega_u(f_t)$ is a group loss function based on the smoothness assumption of the unlabeled target-view data, parameterized as

$$\Omega_u(f_t) = \sum_{g=1}^{G}\beta_g \sum_{k=1,k\neq g}^{G}\sum_{i=1}^{N_u}\|f_s^g(\mathrm{X}_i^u) - f_s^k(\mathrm{X}_i^u)\|^2, \quad (9)$$

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 sample X_i^u, its decision values of different source view classifiers should be similar to each other.

Putting all the terms together, we have the following optimization problem:

$$\begin{aligned}\min_{\beta} \quad & \frac{1}{2}\|\beta\|^2 + \sum_{i=1}^{N_t}\|f_t(\mathrm{X}_i^t) - C_i^t\|^2 \\ & + \sum_{g=1}^{G}\beta_g \sum_{k=1,k\neq g}^{G}\sum_{i=1}^{N_u}\|f_s^g(\mathrm{X}_i^u) - f_s^k(\mathrm{X}_i^u)\|^2, \quad (10) \\ \mathbf{s.t.} \quad & \sum_{g=1}^{G}\beta_g = 1, \beta_g > 0, \forall g.\end{aligned}$$

The optimization problem of Eqn.10 can be solved by a standard Quadratic Programming.

5. Experiments

5.1. Dataset

We evaluate the performance of our method on the IXMAS multi-view dataset [14] which 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.

An action video is represented by sequential images/descriptors. We extract two heterogeneous representations: sequential optical flows and sequential silhouettes, to respectively describe source-view actions and target-view actions. A silhouette descriptor is extracted from the body region and fixed to the size of $40 \times 80 = 3200$. An optical flow descriptor is constructed by the concatenation of four flow components with the size of $40 \times 80 \times 4 = 12800$. The dimension of the linear subspace for either silhouette image set or optical flow descriptor set is around 10.

5.2. Pairwise cross-view recognition

In this experiment, we take one view as the source view and take another different view as the target view. The optical flow feature is adopted in the source view and the silhouette feature is used in the target view. To verify the effectiveness of Heterogeneous Transfer Discriminant-analysis of Canonical Correlations (HTDCC) across pairwise views, we look into the recognition performances of all possible pairwise combinations. The leave-one-subject-

out cross validation strategy (i.e., 12-fold cross validation) is employed. Specifically, for each time, we use videos of one subject from the target view for testing, and use the remaining videos (i.e., videos of the rest 11 subjects) from the target view as well as all the videos from the source view as training data. For the training data, only a small number of samples from the target view and all the samples from the source view are labeled.

We compare HTDCC with the baseline method, called Heterogeneous Discriminant-analysis of Canonical Correlations (HDCC), which excludes the minimization of data distribution mismatch between source and target views in the objective function, i.e. $\alpha = 0$ in Eqn.3. For these two methods, SVM is employed for classification and the regularization parameter is set to $C = 1$ by choosing from $\{1, 10, 100, 1000\}$ on the best performance. The canonical correlations based kernel is used in SVM.

Table 1 demonstrates the recognition results of HTDCC and HDCC with the fraction of labeled samples from the target view of $3/11$. From Table 1, we observe that HTDCC is generally better than HDCC in terms of mean recognition accuracy for all the target views, which clearly demonstrates that HTDCC can successfully deal with the cross-view recognition over heterogeneous feature spaces by minimizing the data distribution mismatch difference between source view and target view.

Our method is also compared with other state-of-the-art methods [6, 12, 10, 13, 1] of transfer learning on heterogeneous feature spaces. For KCCA[10], HeMap[12] and DAMA[13], after learning the projection matrices, we apply SVM to train their final classifiers by using the projected training data from both views. For ARC-t[6], we construct the kernel matrix based on the learned asymmetric transformation metric, and then SVM is also applied to train its final classifier. For HFA[1], the two projection matrices for the source and target data are found by using the standard SVM with the hingeloss. For all methods, the regularization parameter C in SVM is chosen from $\{1, 10, 100, 1000\}$ according to the best performance and the linear kernel is employed.

As shown in Table 2, 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 the target training data. HTDCC outperforms DAMA, possibly due to the lack of the strong manifold strcuture on this dataset. The explanation for the better performance of HTDCC than ARC-t may be that HTDCC utilizes unlabeled target-view training data and incorporates the minimization of the distribution mismatch between source and target views in the objective function.

5.3. Multiple source views fusion

We select one view as the target view and use the other four views as source views to exploit the benefits of combining multiple source views for target recognition. 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 the combination weights of classifiers from multiple source views, we try a fusion method that uses equal combination weights $\beta_g = 1/G$, i.e., $\lambda_l = \lambda_u = 0$ for comparison. To evaluate the contribution of the unlabeled target-view samples for learning the target classifier, we also report the results when excluding the loss function term defined on the unlabeled target-view training data in Eqn.7, i.e., $\lambda_l = 0.1, \lambda_u = 0$. 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.7, i.e., $\lambda_l = 0, \lambda_u = 0.1$.

From the results shown in Table 3, it is interesting that: (1) the fusion of multiple source views achieves better results than each single source view because one single view has limited discriminative ability compared with multiple views; (2) assigning different combination weights to different source views can improve the recognition performance due to the selection of more related source-view classifiers transferred to the target-view classifier. Fig.1 shows some examples of learned weights of multiple source views. We can notice that the more related the source view is to the target view, the higher the learned combination weight becomes. For example, the "Target view 2" is more related to the third source view, and the weight of the third source view is higher than that of other source views.

We also report the recognition accuracy of each action class in Fig.2 which shows that the task of source-view classifier transfer is very hard for some actions and some views. For example, the recognition accuracies of "get up" and "pick up" 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.

6. 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 source and target views to a common space, by simultaneously minimizing the canonical correlations of inter-class samples, maximizing the intra-class canonical correlations, and reducing the data distribution mismatch between source and target views. Moreover, a joint weight learning method is presented to flexi-

Table 1. Pairwise cross-view recognition accuracies using HDCC and HTDCC. Each row is a source view and each column is a target view. The two numbers in a tuple are the recognition accuracies of HDCC and HTDCC, respectively.

	Target view1	Target view2	Target view3	Target view4	Target view5
Source view1		(43.1%, 47.2%)	(42.4%, 41.0%)	(50.7%, 61.8%)	(26.4%, 32.6%)
Source view2	(42.4%, 44.4%)		(43.8%, 44.4%)	(58.3%, 57.6%)	(36.1%, 35.4%)
Source view3	(39.6%, 45.8%)	(45.8%, 48.6%)		(55.6%, 54.2%)	(29.2%, 37.5%)
Source view4	(45.1%, 43.8%)	(43.1%, 41.7%)	(43.8%, 43.1%)		(34.0%, 31.3%)
Source view5	(40.3%, 41.0%)	(40.3%, 45.1%)	(37.5%, 41.0%)	(53.5%, 53.5%)	
Average	(41.8%, 43.8%)	(43.1%, 45.7%)	(41.9%, 42.4%)	(54.1%, 56.8%)	(31.4%, 34.2%)

Table 2. Comparison of different heterogeneous transfer learning methods on the mean recognition accuracy for each target view.

Methods	Target view1	Target view2	Target view3	Target view4	Target view5	Average
KCCA [10]	32.6%	42.9%	26.9%	37.0%	23.6%	32.6%
HeMap [12]	33.7%	39.9%	29.2%	34.7%	22.9%	32.1%
DAMA [13]	33.2%	34.4%	28.1%	31.6%	13.4%	28.1%
ARC-t [6]	29.7%	33.2%	32.8%	33.5%	15.6%	28.9%
HFA [1]	26.6%	33.0%	30.7%	31.8%	13.4%	27.1%
HTDCC	43.8%	45.7%	42.4%	56.8%	34.2%	44.6%

bly combine multiple action classifiers from multiple source views for generating the target-view classifier. Experiments have shown the effectiveness of our method.

7. Acknowledgments

This work was supported in part by the Natural Science Foundations of China (NSFC) under Grant No.61203274, NSFC-Guangdong Joint Fund under Grant No. U1035004 and Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) under Grant No.20121101120029.

References

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

[2] A. Farhadi and M. Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV*, 2008.

[3] J.Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[4] I. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE T-PAMI*, 33(1):172–185, 2011.

[5] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE T-PAMI*, 29(6):1005–1018, 2007.

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

[7] M. Lewandowski, D. Makris, and J. Nebel. View and style-independent action manifolds for human activity recognition. In *ECCV*, 2010.

[8] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, 2012.

[9] J. Liu, M. Shahz, B. Kuipersy, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

[10] J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis. In *Cambridge University Press*, 2004.

[11] Y. Shen and H. Foroosh. View-invariant action recognition using fundamental ratios. In *CVPR*, 2008.

[12] X. Shi, Q. Liu, W. Fan, P. Yu, and R. Zhu. Transfer learning on heterogeneous feature spaces via spectral transformation. In *ICDM*, 2010.

[13] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 2011.

[14] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007.

[15] X. Wu and Y. Jia. View-invariant action recognition using latent kernelized structural svm. In *ECCV*, 2012.

[16] X. Wu, C. Liu, and Y. Jia. Transfer discriminant-analysis of canonical correlations for view-transfer action recognition. In *PCM*, 2012.

[17] X.Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.

[18] P. Yan, S. Khan, and M. Sha. Learning 4d action feature models for arbitrary view action recognition. In *CVPR*, 2008.

[19] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *ICCV*, 2005.

[20] J. Zheng, Z.Jiang, P. Philips, and R. Chellappa. Cross-view action recognition via a transferable dictionary pair. In *BMVC*, 2012.

Table 3. Comparison of different multiple source views fusion methods on the recognition accuracy for each target view.

Methods	Target view1	Target view2	Target view3	Target view4	Target view5	Average
$\lambda_l = \lambda_u = 0$	49.3%	50.7%	46.5%	60.4%	32.6%	47.9%
$\lambda_u = 0$	50.0%	50.0%	46.5%	63.2%	33.3%	48.6%
$\lambda_l = 0$	56.9%	56.2%	56.2%	68.0%	40.3%	55.5%
Our method	57.6%	57.6%	56.9%	68.8%	40.3%	56.3%

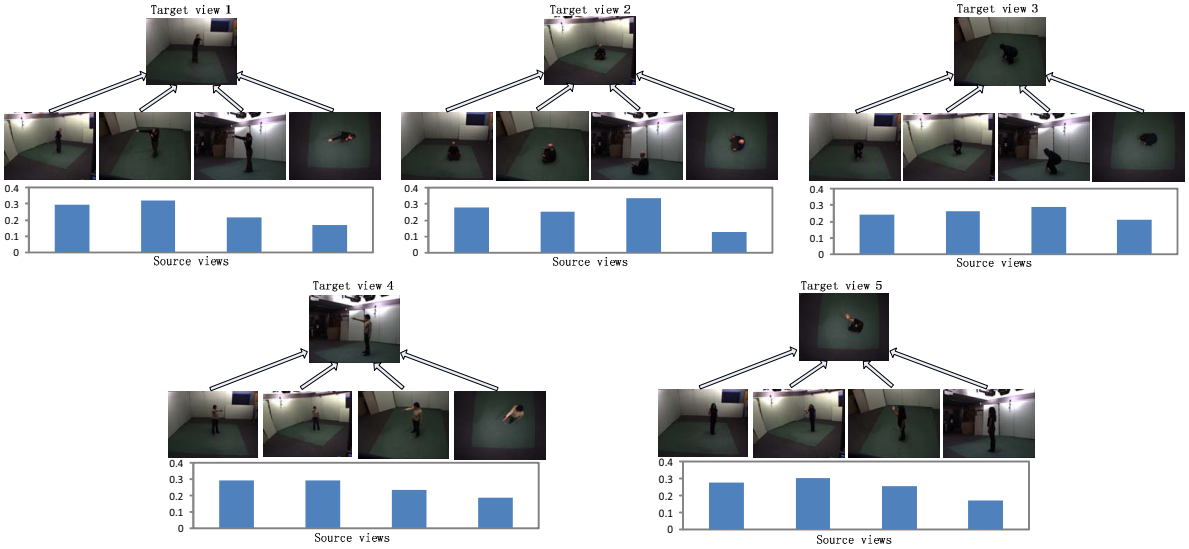

Figure 1. Examples of the learned combination weights of multiple source views. For each target view, its classifiers are constructed by the combination of transferred four source views based on the weights shown by vertical axis of histograms.

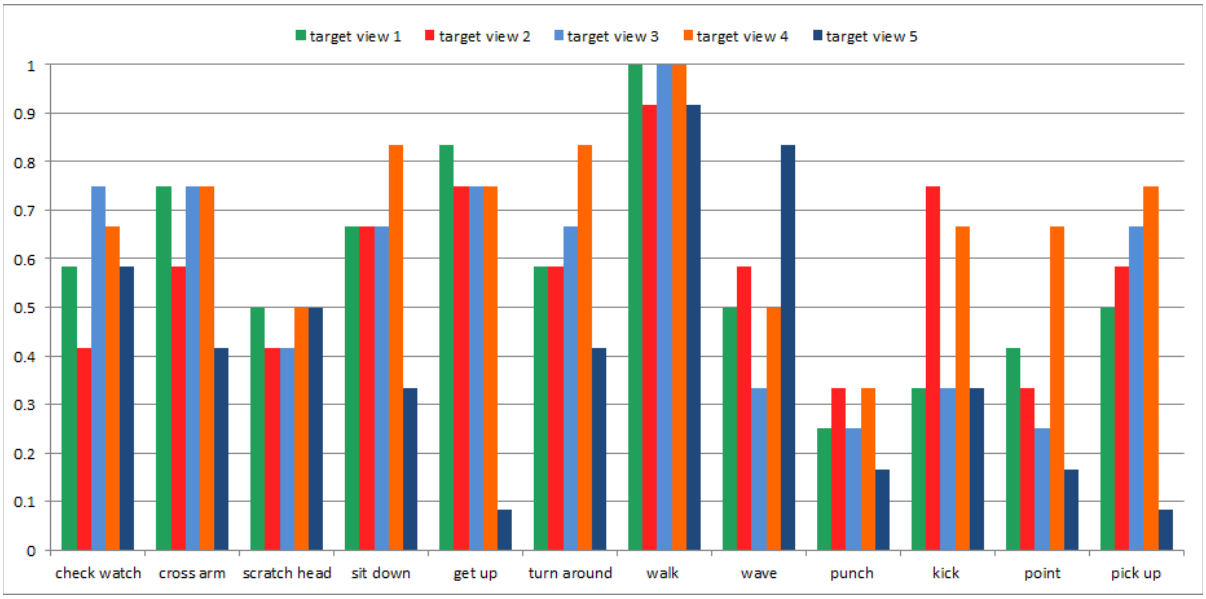

Figure 2. Recognition performance of multiple source views fusion on each action class.