Transfer Discriminant-Analysis of Canonical Correlations for View-Transfer Action Recognition

Xinxiao Wu, Cuiwei Liu, and Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, 100081 ,China
{wuxinxiao,liucuiwei,jiayunde}@bit.edu.cn

Abstract. A novel transfer learning approach, referred to as Transfer Discriminant-Analysis of Canonical Correlations (Transfer DCC), is proposed to recognize human actions from one view (target view) via the discriminative model learned from another view (source view). To cope with the considerable change between feature distributions of source view and target view, Transfer DCC includes an effective nonparametric criterion in the discriminative function to minimize the mismatch between data distributions of these two views. We utilize the canonical correlation between the means of samples from source view and target view to measure the data distribution distance between the two views. Consequently, Transfer DCC learns an optimal projection matrix by simultaneously maximizing the canonical correlation of mean samples from source view and target view, maximizing the canonical correlations of within-class samples and minimizing the canonical correlations of between-class samples. Moreover, we propose a Weighted Canonical Correlations scheme to fuse the multi-class canonical correlations from multiple source views according to their corresponding weights for recognition in the target view. Experiments on the IXMAS multi-view dataset demonstrate the effectiveness of our method.

Keywords: Transfer discriminative learning, canonical correlation analysis, view-transfer action recognition

1 Introduction

Automatic human action recognition from different views has become an essential research topic in computer vision for its wide real-world applications such as video surveillance, human-action interaction and video retrieval. Due to the changing positions of cameras and self-occlusions between different body parts, the appearance of actions may drastically vary from one view to another. Therefore, cross-view action recognition poses substantial challenges for computer vision algorithms.

One strategy [1, 2] is to learn a separate action model for each view, however, it is difficult to collect sufficient labeled samples for each view to cover all the

action categories. Another alternative resorts to 3D reconstruction from multiple views [3, 4] or epipolar geometry reasoning [5, 6]. These methods require calibration setup of multiple cameras or reliable point correspondences, which limits their applicability in practice. Other methods [7, 8] exploit the view-invariant descriptor for recognition. Recently, transfer-learning based approaches [9–11] for cross-view action recognition have received considerable attention in computer vision. This emerging family of methods transfers the feature representation or action models learned on one or more views (source views) to another different view (target view) by exploring the statistical connections between source view and target view without inferring camera geometry or 3D construction from multiple cameras.

In transfer learning, the feature distributions of samples from source view and target view change tremendously, so training with samples from the source view may degrade the performance in the target view. Therefore, it is crucial to reduce the difference between the data distributions of source view and target view. We propose a new transfer learning approach, called Transfer Discriminant-Analysis of Canonical Correlations (Transfer DCC), to handle the considerable inconsistency between data distributions of source view and target view. Our method extends the Discriminant-Analysis of Canonical Correlations (DCC) [12] method to include a nonparametric criterion for comparing data distributions based on the canonical correlation between the means of samples from two views in the projected discriminative data space. Consequently, in Transfer DCC, the optimal discriminative function is proposed to project both source-view and target-view samples, so that the canonical correlation of between-view mean samples is maximized while the canonical correlations of within-class samples and between-class samples are respectively maximized and minimized in the projected data space. In order to further improve the recognition performance, we additionally present a Weighted Canonical Correlations scheme to effectively and flexibly fuse the multi-class canonical correlations from multiple source views according to their corresponding weights to generate the multi-class canonical correlation in the target view. The estimation of the combination weights is formulated under a multi-task learning framework and the learned weight value describes how contributive the canonical correlation of the corresponding class from the corresponding view is to the action prediction in the target view.

2 Related work

The methods most closely related to our approach are that of [9–11]. Farhadi *et al* [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. This method relies 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* [11] proposed "virtual views" to connect action descriptors from source view and target view. 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. From a different perspective, our method simply reduces the inconsistency of data distributions between source view and target view to improve the discrimination in target view, without assuming the continuous transformation of action descriptors between two views.

Our work is also relevant to transfer learning [13] which retains previous knowledge learned from one or multiple existing domains to improve learning in the new domains of interest. In recent years, transfer learning has been successfully applied in many real-world applications such as image and video classification [14–18]. In this paper, we apply the proposed transfer learning method to the view-transfer action recognition. Based on different definitions of the knowledge to be transferred, our method can be categorized into the feature-representation-transfer, which finds a "good" projected feature space that reduces difference between source and target domains. Based on different situations between source and target domains, our method belongs to the transductive transfer, in which the source and target tasks are the same while the source and target domains are different.

3 Transfer DCC learning framework

Each action sample is represented by an orthogonal linear subspace of an image set (i.e., sequential images) and the similarity between two actions is defined by the canonical correlation of the corresponding two subspaces. We do not take into account the temporal dynamics of an action and in some cases several principal images even a single image is sufficient to recognize what a person is doing.

Given a large number of labeled samples from the source view, a small (even no) number of labeled samples from the target view, and some unlabeled samples from the target view, we propose Transfer DCC to find a discriminative space (represented by a projection matrix T) where the canonical correlation of mean samples from source view and target view is maximized while the canonical correlations of within-class samples and between-class samples from all the labeled training samples are maximized and minimized, respectively. Then, the similarity between two actions is measured by the canonical correlation between the two image sets projected by the projection matrix T.

3.1 Brief review of DCC

Discriminant-Analysis of Canonical Correlations (DCC) [12] introduces a linear discriminative function to maximize canonical correlations of within-class

sets and minimize canonical correlations of between-class sets. Assume m image sets are given as $\{\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_m\}$, where X_i represents a matrix with each column describing an image. X_i belongs to one action class denoted by C_i. A d-dimensional linear subspace of X_i is represented by an orthonormal basis matrix $\mathrm{P}_i \in \mathrm{R}^{D\times d}$ s.t. $\mathrm{X}_i\mathrm{X}_i^T = \mathrm{P}_i\Lambda_i\mathrm{P}_i^T$, where Λ_i is the d largest eigenvalues, P_i is the corresponding eigenvectors, and D is the dimension of image descriptor. The discriminative projection matrix $\mathrm{T} = [\mathrm{t}_1, \mathrm{t}_2, ..., \mathrm{t}_n] \in \mathrm{R}^{D\times n}$ is defined by $\mathrm{T} : \mathrm{X}_i \rightarrow \mathrm{Y}_i = \mathrm{T}^T\mathrm{X}_i$, where $n \leq D$ and $|\mathrm{t}_i| = 1$, to make the projected image sets more discriminative using canonical correlations. Orthonormal basis matrices of the subspaces of the projected data are given by $\mathrm{Y}_i\mathrm{Y}_i^T = (\mathrm{T}^T\mathrm{X}_i)(\mathrm{T}^T\mathrm{X}_i)^T = (\mathrm{T}^T\mathrm{P}_i)\Lambda_i(\mathrm{T}^T\mathrm{P}_i)^T$. The matrix P_i is normalized to P'_i so that the columns of $\mathrm{T}^T\mathrm{P}'_i$ are orthonormal. By the SVD computation $(\mathrm{T}^T\mathrm{P}'_i)^T(\mathrm{T}^T\mathrm{P}'_j) = \mathrm{Q}_{ij}\Lambda\mathrm{Q}_{ji}^T$, the similarity of two projected data sets is defined as the sum of canonical correlations $\mathrm{F}_{ij} = \max_{\mathrm{Q}_{ij},\mathrm{Q}_{ji}} \mathrm{tr}\{\mathrm{T}^T\mathrm{P}'_j\mathrm{Q}_{ji}\mathrm{Q}_{ij}^T\mathrm{P}_i'^T\mathrm{T}\}$. T is determined to maximize the similarities of any pair of within-class sets and minimize the similarities of pair-wise sets of different classes by

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

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

3.2 View transfer via minimizing data distribution mismatch

The conventional DCC method assumes that the training and test data are drawn from the same data distribution. However, for view-transfer action recognition, the training and test samples from different views have different data distribution properties (such as mean, intra-class and inter-class variance). So we extend DCC to handle the problem of reducing the mismatch between the data distributions of source view and target view. An effective nonparametric criterion is integrated into the discriminative function in Eq.(1) to compare data distributions by the canonical correlation between the means of image sets from source view and target view in the projected data space. Then the learning framework of Transfer DCC is 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 view and target view, and defined on all the training samples from source view and target view in the projected data space. α is the tradeoff parameter to balance the data distribution difference between these two views and the discriminative criterion defined on all the labeled training data.

Let $\mathrm{D}^s = \{\mathrm{X}_i^s\}_{i=1}^N$ be the source-view training dataset, where X_i^s is i-th training sample (i.e., an image set) from the source view. Let $\mathrm{D}^t = \mathrm{D}_l^t \bigcup \mathrm{D}_u^t$

be the target-view training dataset, where $\mathrm{D}_l^t = \{\mathrm{X}_{l,i}^t\}_{i=1}^M$ and $\mathrm{D}_u^t = \{\mathrm{X}_{u,i}^t\}_{i=1}^K$ denote the labeled and unlabeled training data from the target view, respectively. $E_r(\mathrm{T})$ in Eq.(2) can be rewritten as:

$$E_r(\mathrm{T}) = \max_{\mathrm{Q}_{st},\mathrm{Q}_{ts}} \mathrm{tr}\{\mathrm{T}^T\mathrm{P}'_s\mathrm{Q}_{st}\mathrm{Q}_{ts}^T\mathrm{P}_t'^T\mathrm{T}\}, \tag{3}$$

where P'_s is the normalized orthonormal basis matrix of the mean of source-view training data $\overline{\mathrm{X}}^s = \frac{1}{N}\sum_{i=1}^N \mathrm{X}_i^s$. P'_t is the normalized orthonormal basis matrix of the mean of target-view training data $\overline{\mathrm{X}}^t = \frac{1}{M+K}(\sum_{i=1}^M \mathrm{X}_{l,i}^t + \sum_{i=1}^K \mathrm{X}_{u,i}^t)$. Q_{st} and Q_{ts} are defined by the SVD computation $(\mathrm{T}^T\mathrm{P}'_s)^T(\mathrm{T}^T\mathrm{P}'_t) = \mathrm{Q}_{st}\Lambda\mathrm{Q}_{ts}^T$. $E_r(\mathrm{T})$ evaluates the data distribution variations between source view and target view, thus we attempt to maximize $E_r(\mathrm{T})$ to reduce the between-view difference.

By the linear algebra $\mathrm{T}^T\mathrm{P}'_j\mathrm{Q}_{ji}\mathrm{Q}_{ij}^T\mathrm{P}_i'^T\mathrm{T} = \mathrm{I} - \mathrm{T}^T(\mathrm{P}'_j\mathrm{Q}_{ji} - \mathrm{P}'_i\mathrm{Q}_{ij})(\mathrm{P}'_j\mathrm{Q}_{ji} - \mathrm{P}'_i\mathrm{Q}_{ij})^T\mathrm{T}/2$, we can rewrite the objective function in Eq.(2) as

$$\mathrm{T} = \arg\max_{\mathrm{T}} \frac{\mathrm{tr}(\mathrm{T}^T\mathrm{S}_b\mathrm{T})}{\mathrm{tr}(\mathrm{T}^T(\mathrm{S}_w + \alpha\mathrm{S}_r)\mathrm{T})}, \tag{4}$$

where $\mathrm{S}_r = (\mathrm{P}'_s\mathrm{Q}_{st} - \mathrm{P}'_t\mathrm{Q}_{ts})(\mathrm{P}'_s\mathrm{Q}_{st} - \mathrm{P}'_t\mathrm{Q}_{ts})^T$, $\mathrm{S}_b = \sum_{i=1}^{N+M}\sum_{l\in B_i}(\mathrm{P}'_l\mathrm{Q}_{li} - \mathrm{P}'_i\mathrm{Q}_{il})(\mathrm{P}'_l\mathrm{Q}_{li} - \mathrm{P}'_i\mathrm{Q}_{il})^T$, $\mathrm{S}_w = \sum_{i=1}^{N+M}\sum_{k\in W_i}(\mathrm{P}'_k\mathrm{Q}_{ki} - \mathrm{P}'_i\mathrm{Q}_{ik})(\mathrm{P}'_k\mathrm{Q}_{ki} - \mathrm{P}'_i\mathrm{Q}_{ik})^T$, $B_i = \{j|C_j \neq C_i\}$ and $W_i = \{j|C_j = C_i\}$. Finally, the optimal T is computed by the eigen-decomposition of $(\mathrm{S}_w + \alpha\mathrm{S}_r)^{-1}\mathrm{S}_b$.

3.3 Learning algorithm

Similar to DCC, we use an iterative optimization algorithm to find the optimized projection matrix T. With the identity matrix I as the initial value of T, the detailed algorithm of Transfer DCC is listed in Algorithm 1. Once the optimal T is found, a comparison of any two actions is achieved by projecting them via T and then computing the canonical correlation.

4 Weighted canonical correlations for multiple view fusion

Since single source view may provide partial action knowledge, it is beneficial to combine the action knowledge transferred from multiple source views to improve the recognition performance in the target view. We propose a new Weighted Canonical Correlations method to effectively fuse multi-class canonical correlations from multiple source views into the multi-class canonical correlations of the target view for recognition.

We define the canonical correlation between the target-view sample X^t and the j-th action class from the h-th source view as $a_{jh} = \max_k Similar(\mathrm{X}^t, \mathrm{X}_k)$, where X_k represents the labeled training samples of the j-th action class from both h-th source view and target view. $Similar(\mathrm{X}^t, \mathrm{X}_k)$ denotes the canonical

Algorithm 1 Transfer Discriminant-Analysis of Canonical Correlations

Input: N labeled training samples $\{\mathrm{X}_i^s\}_{i=1}^N$ from the source view
M labeled training samples $\{\mathrm{X}_{l,i}^t\}_{i=1}^M$ from the target view
K unlabeled training samples $\{\mathrm{X}_{u,i}^t\}_{i=1}^K$ from the target view
Output: Projection matrix $\mathrm{T} \in \mathrm{R}^{D\times n}$
Initialize: $\mathrm{T} = \mathrm{I}$.
1. Compute the mean of source-view samples by $\overline{\mathrm{X}}^s = \frac{1}{N}\sum_{i=1}^N \mathrm{X}_i^s$.
2. Compute the mean of target-view samples by $\overline{\mathrm{X}}^t = \frac{1}{M+K}(\sum_{i=1}^M \mathrm{X}_{l,i}^t + \sum_{i=1}^K \mathrm{X}_{u,i}^t)$.
3. Compute the orthonormal basis matrices P_i, P_s, P_t of X_i, $\overline{\mathrm{X}}^s$, $\overline{\mathrm{X}}^t$ by $\mathrm{X}\mathrm{X}^T = \mathrm{P}\Lambda\mathrm{P}^T$, $\mathrm{P} \in \mathrm{R}^{D\times n}$.
4. **Do iterate the following**:
5. Normalize P_i, P_s, P_t to P'_i, P'_s, P'_t by QR-decomposition: $\mathrm{T}^T\mathrm{P} = \Phi\Delta$, $\mathrm{P}' = \mathrm{P}\Delta^{-1}$.
6. For every pair P'_i, P'_j from the labeled training dataset, do SVD: $(\mathrm{T}^T\mathrm{P}'_i)^T(\mathrm{T}^T\mathrm{P}'_j) = \mathrm{Q}_{ij}\Lambda\mathrm{Q}_{ji}^T$.
7. For P'_s, P'_t, do SVD: $(\mathrm{T}^T\mathrm{P}'_s)^T(\mathrm{T}^T\mathrm{P}'_t) = \mathrm{Q}_{st}\Lambda\mathrm{Q}_{ts}^T$.
8. Compute S_r, S_b, S_w.(See Eq.(4))
9. Compute eigenvectors $\{\mathrm{t}_i\}_{i=1}^n$ of $(\mathrm{S}_w + \alpha\mathrm{S}_r)^{-1}\mathrm{S}_b$.
10. **End**
11. $\mathrm{T} = [\mathrm{t}_1, \mathrm{t}_2, ..., \mathrm{t}_n] \in \mathrm{R}^{D\times n}$.

correlation between X^t and X_k by projecting them via the projection matrix T_h learned from the h-th source view. Then the combined canonical correlation between X^t and the j-th action class from V source views is given by $g_j = \mathbf{w}_j^T\mathbf{d}_j$, where $\mathbf{w}_j = [w_{j1}, w_{j2}, ..., w_{jV}]^T \in \mathrm{R}^{V\times 1}$ is the weight vector of the j-th action class canonical correlations from V source views. $\mathbf{d}_j = [a_{j1}, a_{j2}, ..., a_{jV}]^T \in \mathrm{R}^{V\times 1}$ is the j-th class canonical correlation vector from V source views. Actually, g_j can be considered as the final decision value of X^t belonging to the j-th action class. Thus it is essential to estimate the combination weights $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_C] \in \mathrm{R}^{V\times C}$ with C the number of action classes.

Let $\mathrm{Y}_i = [y_{i1}, y_{i2}, ..., y_{iC}]^T \in \mathrm{R}^{C\times 1}$ be the multi-class label of the i-th training sample $\mathrm{X}_{l,i}^t$ from the target view, where y_{ij} represents the j-th class label of $\mathrm{X}_{l,i}^t$. If $\mathrm{X}_{l,i}^t$ belongs to the C-th action class, then $y_{iC} = 1$ and $y_{ij} = 0$, $\forall j \neq C$. Let $\mathbf{d}_{ji} \in \mathrm{R}^{V\times 1}$ be the j-th class canonical correlation vector from V source views of $\mathrm{X}_{l,i}^t$, we aim to find the optimized W via the following formulation:

$$\mathbf{W} = \arg\min_{\mathbf{w}_j} \sum_{j=1}^{C}\sum_{i=1}^{M} L(y_{ij}, \langle \mathbf{w}_j, \mathbf{d}_{ji}\rangle) + \gamma \sum_{j=1}^{C} \langle \mathbf{w}_j, \mathbf{w}_j\rangle, \tag{5}$$

where $\gamma > 0$ is a regularization parameter. The first term $\sum_{j=1}^{C}\sum_{i=1}^{M} L(y_{ij}, \langle \mathbf{w}_j, \mathbf{d}_{ji}\rangle)$ in Eq.(5) is the average of the error cross the classes, measured according to a prescribed loss function L which is a square loss in this paper. The second term $\gamma\sum_{j=1}^{C}\langle \mathbf{w}_j, \mathbf{w}_j\rangle$ is the average of the 2-norm regularization problem cross the classes. This minimization is a convex problem and can be carried out independently cross the classes. In our implement, the Matlab code for the multi-task

feature learning proposed in [19] is utilized to solve the optimization problem in Eq.(5).

5 Experiments

5.1 Dataset and experimental setup

We evaluate the performance of our method on the IXMAS multi-view dataset [3] which consists of 11 complete action classes. Each action is executed three times by 12 subjects and recorded by five 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 an image set of sequential frames and each frame is described by the extracted body region which is normalized to the size of 80×40 pixels. The dimension of the linear subspace of each image set is around 10 to represent 98% data energy of the set. Figure 1 shows some action examples from five views.

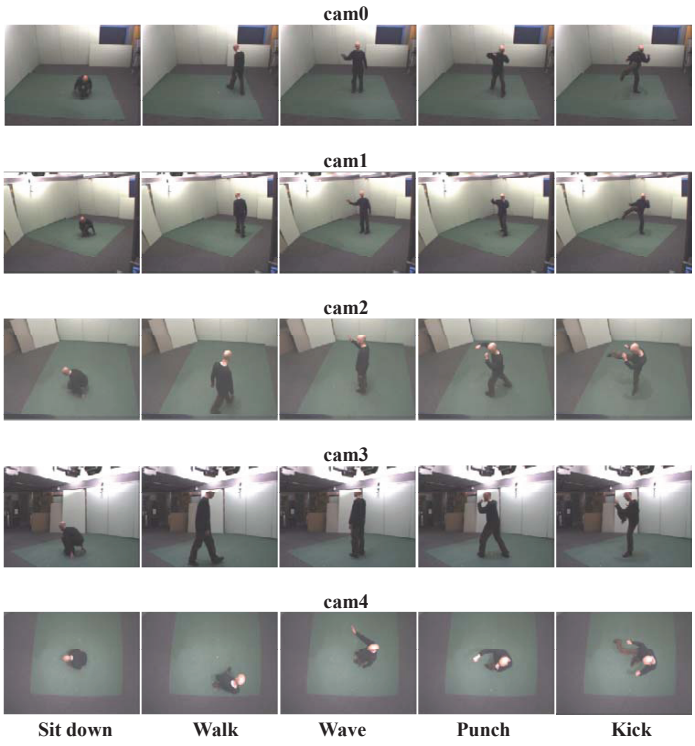

Fig. 1. Samples frames from action videos on the IXMAS multi-view dataset.

5.2 Pairwise view-transfer recognition

To verify the effectiveness of Transfer DCC 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 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 are utilized 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 Transfer DCC with the baseline DCC [12]. For both these two methods, 1-Nearest Neighbors (1NN) is employed for classification. Table 1 demonstrates the recognition results of Transfer DCC and DCC with the fraction of labeled samples from the target view of 0, 1/11, 2/11 and 3/11. From Table 1, we have the following observations: 1) TDCC is better than DCC in terms of recognition accuracy over almost all the cases, which clearly demonstrates that TDCC can successfully deal with the view-transfer recognition by minimizing the data distribution difference between source view and target view. 2) Only in a few cases (e.g., the source view is C1 and the target view is C4 with the fraction of the labeled target-view samples of 0), TDCC works worse than DCC. The possible explanation is that the data distribution difference between C1 and C4 is so large that the transferred information degrades the performance. In the future, we will investigate how to avoid such negative transfer problem. 3) With the increment of the labeled samples from the target view, the performances of both DCC and TDCC significantly improve because of the increasing supervision information from the target view. It is interesting to note that the proposed TDCC provides more discrimination than DCC even when no target labels are available (i.e., the fraction is 0).

5.3 Multiple view-transfer recognition

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. Multi-class canonical correlations from the four source-target pairs are fused via the proposed Weighted Canonical Correlation (WCC) method. To verify the effectiveness of the combination weights of multi-class canonical correlations from multiple source views, we also try a fusion method that uses equal combination weights of canonical correlations (ECC, i.e.,$\mathbf{W} = \mathrm{I}$) for comparison. Table 2 shows the comparison results between ECC and WCC in the Transfer DCC learning framework. In this experiment, the fraction of the labeled training data from target view is set to 3/11. As shown in Table 2, it is interesting to notice that TDCC-WCC outperforms TDCC-ECC, which obviously demonstrates the effectiveness of our fusion method. By comparing Table 2 to Table 1, it is also interesting to observe that for most target views, the fusion of multiple source views achieves better results than each single source view because of the limited discriminative ability of one single view. Figure 2 demonstrates the recognition accuracy of each action class.

Table 1. Pairwise view-transfer recognition accuracies (%) using Transfer DCC (TDCC) and DCC with the fraction of labeled samples from the target view of 0, 1/11, 2/11 and 3/11. The rows and columns correspond to the training and test views, respectively. "C0", "C1", "C2", "C3" and "C4" represent the five camera views.

(a) fraction=0

	C0	C1	C2	C3	C4
C0		36.4 / 40.2	29.6 / 34.9	22.0 / 28.8	6.8 / 9.1
C1	37.1 / 40.2		42.2 / 44.7	25.0 / 24.2	9.9 / 7.6
C2	30.3 / 37.1	43.2 / 50.8		32.6 / 38.6	10.6 / 15.2
C3	23.5 / 28.8	31.1 / 35.6	35.6 / 44.7		7.6 / 14.4
C4	9.1 / 9.9	16.7 / 13.6	17.4 / 13.6	12.9 / 12.1	
Ave.	25.0 /29.0	31.9 / 35.1	31.2 / 34.5	23.1 / 25.9	8.7 / 11.6

(b) fraction=1/11

	C0	C1	C2	C3	C4
C0		42.4 / 48.5	45.6 / 57.0	53.8 / 47.7	22.7 / 34.1
C1	47.7 / 48.5		50.0 / 50.0	49.2 / 48.5	31.1 / 34.1
C2	39.4 / 47.0	50.0 / 53.8		50.8 / 54.6	28.0 / 31.8
C3	43.2 / 48.5	40.2 / 49.2	37.9 / 50.8		29.6 / 31.1
C4	35.6 / 41.7	41.7 / 46.2	29.6 / 37.9	51.5 / 50.8	
Ave.	41.5 / 46.4	43.6 / 49.4	40.8 / 48.9	51.3 / 50.4	27.9 / 32.8

(c) fraction=2/11

	C0	C1	C2	C3	C4
C0		44.7 / 54.6	49.2 / 52.3	52.3 / 52.3	32.6 / 40.9
C1	52.3 / 56.8		50.8 / 47.0	52.3 / 51.5	37.9 / 38.6
C2	47.0 / 50.8	50.8 / 59.9		51.5 / 54.6	37.1 / 37.1
C3	43.2 / 53.8	46.2 / 51.5	48.5 / 52.3		33.3 / 40.2
C4	43.9 / 49.2	37.9 / 45.5	36.4 / 43.9	50.8 / 50.8	
Ave.	46.6 / 52.7	44.9 / 52.9	46.2 / 48.9	51.7 / 52.3	35.2 / 39.2

(d) fraction=3/11

	C0	C1	C2	C3	C4
C0		48.5 / 59.1	48.5 / 52.3	56.1 / 57.6	34.9 / 44.7
C1	60.6 / 56.8		56.8 / 55.3	63.6 / 62.9	41.7 / 42.4
C2	51.5 / 56.8	53.8 / 60.6		53.0 / 60.6	42.4 / 40.9
C3	52.5 / 56.8	53.8 / 56.1	47.7 / 53.0		40.2 / 45.5
C4	48.5 / 55.3	53.0 / 50.8	37.9 / 41.7	56.1 / 53.8	
Ave.	53.3 / 56.4	52.3 / 56.7	47.7 / 50.6	57.4 / 58.7	39.8 / 43.4

Table 2. Multiple view-transfer recognition accuracies (%) on different methods.

Methods	C0	C1	C2	C3	C4	Ave.
TDCC-ECC	59.1	62.1	54.6	60.6	47.7	56.8
TDCC-WCC	64.4	64.4	57.6	62.9	47.7	59.4

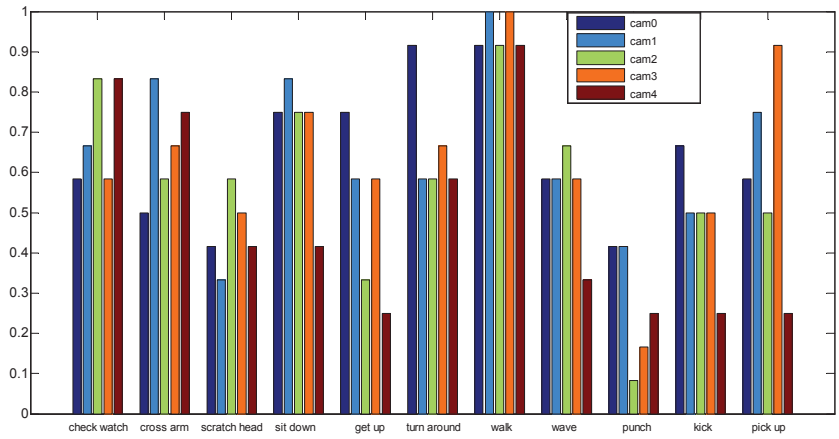

Fig. 2. Recognition results of multiple view-transfer on each action class. The horizontal and vertical axes indicate the action classes and recognition accuracies, respectively.

6 Conclusions

We have proposed a novel Transfer Discriminant-Analysis of Canonical Correlations (Transfer DCC) method for view-transfer action recognition. The data distribution inconsistency between source view and target view is minimized to improve the performance of the discriminative model for the target view when learned from the source view. Our method dose not require any corresponding action instances in the two views and operates under a weakly supervised or even non-supervised scheme. Moreover, a Weighted Canonical Correlations approach is presented to flexibly fuse multiple action predictions from multiple source views. Experiments have shown the effectiveness of our method. Future work includes extracting dynamic motion features for recognition since in this work each action is represented by an image set neglecting the temporal information of motion.

7 Acknowledgments

This work was supported by the Natural Science Foundation of China(NSFC) under Grant No. 90920009 and NSFC-Guangdong Joint Fund under Grant No. U1035004.

References

1. X.Wu, Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: Proc. CVPR. (2011) 489–496

2. J.Liu, Shah, M.: Learning human actions via information maximization. In: Proc. CVPR. (2008) 1–8
3. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: Proc. ICCV. (2007) 1–7
4. Yan, P., Khan, S., Sha, M.: Learning 4d action feature models for arbitrary view action recognition. In: Proc. CVPR. (2008) 1–7
5. Yilmaz, A., Shah, M.: Recognizing human actions in videos acquired by uncalibrated moving cameras. In: Proc. ICCV. (2005) 150–157
6. Shen, Y., Foroosh, H.: View-invariant action recognition using fundamental ratios. In: Proc. CVPR. (2008) 1–6
7. Junejo, I., Dexter, E., Laptev, I., Perez, P.: View-independent action recognition from temporal self-similarities. IEEE T-PAMI **33**(1) (2011) 172–185
8. Lewandowski, M., Makris, D., Nebel, J.: View and style-independent action manifolds for human activity recognition. In: Proc. ECCV. (2010) 547–560
9. Farhadi, A., Tabrizi, M.: Learning to recognize activities from the wrong view point. In: Proc. ECCV. (2008) 154–166
10. Liu, J., Shahz, M., Kuipersy, B., Savarese, S.: Cross-view action recognition via view knowledge transfer. In: Proc. CVPR. (2011) 3209–3216
11. Li, R., Zickler, T.: Discriminative virtual views for cross-view action recognition. In: Proc. CVPR. (2012)
12. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE T-PAMI **29**(6) (2007) 1005 – 1018
13. Pan, S., Yang, Q.: A survey on transfer learning. IEEE T-KDE **22**(10) (2010) 1345–1359
14. Duan, L., Xu, D., Tsang, I., Luo, J.: Visual event recognition in videos by learning from web data. In: Proc. CVPR. (2010) 1959–1966
15. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: Proc. CVPR. (2011) 1785–1792
16. Gopalan, R., Li, R., Chellappa, R.: Domain adaption for object recognition: an unsupervised approach. In: Proc. ICCV. (2011) 999–1006
17. Lampert, C., Kromer, O.: Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In: Proc. ECCV. (2010) 566–579
18. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: Proc. ICCV. (2011) 1863–1870
19. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning **73**(3) (2008) 243–272