Neurocomputing 191 (2016) 286–295

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Heterogeneous discriminant analysis for cross-view action recognition

Wanchen Sui, Xinxiao Wu*, Yang Feng, Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, PR China

ARTICLE INFO

Article history:
Received 10 August 2015
Received in revised form
9 December 2015
Accepted 15 January 2016
Communicated by Yongzhen Huang
Available online 18 February 2016

Keywords:
Cross-view action recognition
Transfer learning
Discriminant analysis
Heterogeneous domain adaption

ABSTRACT

We propose an approach of cross-view action recognition, in which the samples from different views are represented by features with different dimensions. Inspired by linear discriminant analysis (LDA), we introduce a discriminative common feature space to bridge the source and target views. Two different projection matrices are learned to respectively map the action data from two different views into the common space by simultaneously maximizing the similarity of intra-class samples, minimizing the similarity of inter-class samples and reducing the mismatch between data distributions of two views. In addition, the locality information is incorporated into the discriminant analysis as a constraint to make the discriminant function smooth on the data manifold. Our method is neither restricted to the corresponding action instances in the two views nor restricted to a specific type of feature. We evaluate our approach on the IXMAS multi-view action dataset and N-UCLA dataset. The experimental results demonstrate the effectiveness of our method.

1. Introduction

Human action recognition in videos plays an important role in computer vision due to its wide applications in human–computer interaction, smart surveillance, and video retrieval. In order to accurately recognize human actions, lots of approaches focus on developing effective action representation, such as 2D shape matching [1–3], optical flow patterns [4], spatio-temporal interest points [5–7], and trajectory-based descriptors [8–10]. Especially, dense trajectories-based methods have achieved impressive results on a variety of datasets [11,12]. These methods are effective for action recognition from a single viewpoint. However, the problem of viewpoint changes has posed a real challenge to human action recognition for the fact that the same action appears quite different when observed from different views. Both the data distribution and the feature space can vary drastically from one view to another. As a result, action models learned in one view tend to be incapable of the recognition in another different view [13–16].

Recently, lots of efforts have been made towards the problem of cross-view action recognition. A number of geometry-based approaches are motivated to perform by using the geometry measurement of body joints [17–20] or inferring 3D models of human subjects [21–23], usually requiring robust joint estimation which is still a challenging task. Another group of approaches tries to compute view-invariant human action representations that are stable across different viewpoints, such as temporal self-similarity matrix descriptors [24], temporal dynamics representation [25]. Several methods [26–32] have resorted to transfer learning, which constructs the mappings or connections to bridge the gap between different views. Methods [26,27,31,32] rely on either feature-to-feature correspondence or video-to-video correspondence to transfer knowledge across views. Methods [28,30] require action features of the same type in different views. However, the corresponding data and homogeneous features in both views are not always available easily. In [29], Wu et al. proposed an iterative optimization algorithm to learn a common subspace for cross-view action recognition over heterogeneous feature spaces. Their method has less restrictions except that each action sample must be represented by a sequence of image features.

In this paper, we present a new transfer learning approach for cross-view action recognition in heterogeneous feature spaces, called Heterogeneous Linear Discriminant Analysis (*HLDA*). Our method is neither restricted to the corresponding action instances in the two views nor restricted to action features of the same type. Moreover, in this work, each action sample is represented by a commonly used feature vector. All these make our method more general than the existing ones. Specifically, in order to effectively utilize these features, we are encouraged to align the features from the two views via a discriminative common space, where the action samples captured from different viewpoints can be compared directly and the data from different classes can be separated well.

Our paper focuses on the construction of the common space. We aim to learn two different projection matrices to respectively map the data from the source and target views to the common feature

* Corresponding author.
E-mail address: wuxinxiao@bit.edu.cn (X. Wu).

http://dx.doi.org/10.1016/j.neucom.2016.01.051
0925-2312/© 2016 Elsevier B.V. All rights reserved.

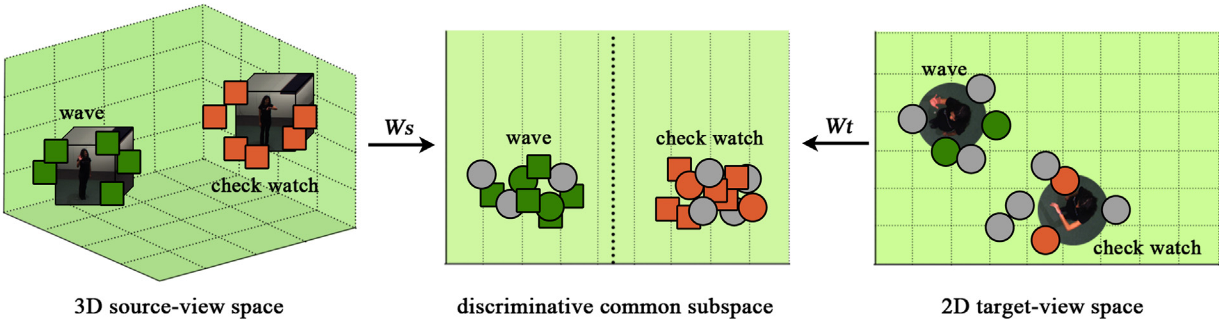


Fig. 1. An illustration of our framework. Samples from different views are represented by features with different dimensions, painted as different shapes (i.e., square and circle). Samples from different classes are denoted by different colors (i.e., green and red). The gray ones are unlabeled data. *Ws* and *Wt* are the projection matrices respectively mapping the heterogeneous data from two views to the derived common subspace. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

space. The two projections are learned by simultaneously maximizing the variance of intra-class samples, minimizing the variance of inter-class samples. Taking Fig. 1 for example, given the 3D data as source-view data and the 2D data as target-view data, we propose to explore the common subspace where the data points belonging to different classes (denoted in different colors) are well-separated from each other, while those from the same class (denoted in the same color) are closely related to each other. In order to reduce the mismatch between data distributions of different views, we also add an effective nonparametric criterion into the objective function. As Fig. 1 shows, the two distributions are similar in the projected subspace, even though they look quite different in the original 3D and 2D spaces. Moreover, a valid locality constraint is incorporated into the discriminant analysis, which preserves the local manifold structure and makes the discriminant function as smooth as possible on the data manifold. It is designed to make the neighbor data points in original space still close to each other in the new projection subspace, which keeps the similarity of the original neighbor data by using the labeled and unlabeled data. This framework can be naturally generalized to the corresponding kernel version using the kernel trick [33], called Heterogeneous Kernel Discriminant Analysis (*HKDA*), which leads to better performance.

The rest of the paper is organized as follows. Section 2 describes the recent works related to our approach. The proposed HLDA method and HKDA are introduced in Sections 3 and 4, respectively. Extensive experimental results are presented in Section 5, followed by conclusions in Section 6.

2. Related work

2.1. Cross-view action recognition

Recently, several transfer learning based methods have been proposed for cross-view action recognition. Farhadi et al. [26] employed maximum margin clustering (MMC) to generate split-based features in source view, and then transferred the split values to the corresponding frames in target view. Zhang et al. [32] imposed temporal regularization on the traditional MMC. These methods require the feature-to-feature corresponding relation at the frame-level. Liu et al. [27] presented a bipartite-graph-based approach to learn bilingual-words from two view-dependent vocabularies in an unsupervised manner, and then transferred actions from different views by a bag-of-bilingual-words model instead of bag-of-visual-words model. Zheng and Jiang [31] proposed a dictionary learning framework to exploit the video-to-video correspondence, by jointly learning a set of view-specific dictionaries for aligning view-specific features and a common dictionary for modeling view-shared features. Different from these approaches, our method possesses the view-shared action representations without any feature-to-feature correspondence or video-to-video correspondence. Li and Zickler [28] tried to connect the source and target views by a smooth virtual path, which is represented as a sequence of linear transformations of action descriptors. Similarly, Zhang et al. [30] intended to bridge two views via a continuous virtual path keeping all the visual information. These methods require the action features of different views with the same dimension. Jia et al. [34] proposed to transfer the depth information from the source RGB-D database to the target RGB database, and use the additional source information to recognize human actions in RGB videos. It can be applied to cross-view action recognition, but it emphasized the data type of source and target database. Wu et al. [29] extended discriminant-analysis of canonical correlations (DCC) [35] to accomplish cross-view action recognition over heterogeneous feature spaces. In their work, each action sample must be represented by a set of image features, and the framework cannot be combined with some impressive action features, such as spatio-temporal features [5], dense trajectory features [8], while ours relax the restriction of data type. Another distinct difference is that our method permits kernelization, which is necessary for learning the projections with non-linear effect. In addition, our method can easily get the closed-form solution, while theirs find the optimized solution by an iterative optimization algorithm.

2.2. Transfer learning

In terms of transfer learning, our view-transfer problem has much in common with the heterogeneous domain adaptation problem, and the methods [36–44] are closely related to our work for constructing an effective common feature space. Shawe-Taylor and Cristianini [36] proposed kernel-based canonical correlation analysis (KCCA) to learn the common feature subspace by maximizing the correlation between the source and target training data in an unsupervised manner. Several approaches [37,38] extend KCCA to deal with the problem of cross-view action recognition. Shi et al. [39] employed spectral embedding to unify the different feature spaces without using any label information of training data. Different from these approaches, our method does not require the simultaneous multi-view observations of the same action instance. Wang and Mahadevan [40] proposed a manifold alignment based approach to project samples into a latent space, simultaneously matching the samples with the same labels, separating the samples with different labels and preserving the topology of each domain. They assumed that the data should have a manifold structure. Kulis et al. [41] tried to learn an asymmetric, non-linear transformation for domain adaptation by a supervised learning approach. Hoffman et al. [42] extended this work to simultaneously learn the transformation of features and the classifier parameters by using the same classification loss. These two methods explored the discriminative information contained in the labeled data, while our method can learn the transformation with both insufficient labeled data and sufficient unlabeled data. Duan et al. [43] presented a heterogeneous feature augmentation (HFA) method to align different domains by two projection matrices, using the standard support vector machine (SVM) with the hinge loss. Li et al. [44] equivalently reformulated the optimization problem in [43] into a convex optimization problem, which shares a similar formulation with the well-known Multiple Kernel Learning (MKL) problem. Their methods are formulated to solve a binary problem, so a new feature transformation must be learned for each class, while our method can learn a transformation that generalized to novel target class. Some other methods [45,46] also tried to seek a common subspace to mitigate the semantic gap among different views. However, they cannot deal with the data from heterogeneous feature spaces directly.

3. Heterogeneous Linear Discriminant Analysis

In this paper, we focus on the problem of cross-view action recognition in which the data from different views are represented by features from heterogeneous spaces. Given a large number of labeled training samples from the source view $\{(x_i^s, y_i^s)|_{i=1}^{n_s}\}$ with $x_i^s \in \mathbb{R}^{d_s}$, a few labeled samples from the target view $\{(x_i^l, y_i^l)|_{i=1}^{n_l}\}$ with $x_i^l \in \mathbb{R}^{d_t}$ and some unlabeled samples from the target view $\{(x_i^u)|_{i=1}^{n_u}\}$ with $x_i^u \in \mathbb{R}^{d_t}$, where n_s, n_l are the numbers of labeled samples from source and target views, n_u is the number of unlabeled samples from target view, $y_i{}^s$ and $y_i{}^l$ represent the labels of the samples $x_i{}^s$ and $x_i{}^l$, and $y_i^s, y_i^l \in \{1, \ldots, c\}$ with c being the number of action classes. In general, the feature dimensions in source and target views are not equal, i.e., $d_s \neq d_t$.

We aim to find two projection matrices w_s and w_t for respectively mapping data from source and target views into a common subspace via simultaneously minimizing the variance of intra-class samples and maximizing the variance of inter-class samples. Fukunaga [47] pointed out that there are equivalent variants of Fisher criterion to get the projection matrix w:

$$w = \arg\max_w \frac{|w^T S_B w|}{|w^T S_W w|} = \arg\max_w \frac{|w^T S_B w|}{|w^T S_T w|}, \tag{1}$$

where S_B is the between-class scatter matrix, S_W is the within-class scatter matrix, and $S_T = S_B + S_W$ is the total scatter matrix. As for this work, we use the second criterion in Eq. (1). Meanwhile, in order to reduce the mismatch between data distributions of different views, we add a nonparametric criterion into the discriminative function. We also impose a locality constraint on the objective function to respectively preserve the local structure of the source-view and target-view data. Hence, denoting $w = \begin{bmatrix} w_s \\ w_t \end{bmatrix}$ as the total projection matrix, our objective function can be formulated as

$$w = \arg\max_w \frac{|w^T S_B w|}{|w^T S_T w| + \gamma_1 |w^T S_D w| + \gamma_2 |w^T S_L w|}, \tag{2}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are the tradeoff parameters, $w^T S_T w$ is the scatter matrix of all data in the common space, $w^T S_B w$ the scatter matrix of between-class data projected in the common space, $w^T S_D w$ the between-view distribution difference, $w^T S_L w$ the locality constraint matrix. Specifically, these four matrices can be defined as follows.

3.1. Total scatter matrix

The total scatter matrix $w^T S_T w$ in the common subspace is formulated as

$$w^T S_T w = \sum_{i=1}^{n_s} (w_s^T x_i^s - \mu)(w_s^T x_i^s - \mu)^T + \sum_{i=1}^{n_l} (w_t^T x_i^t - \mu)(w_t^T x_i^t - \mu)^T, \tag{3}$$

where μ indicates the global mean of all the projected labeled data from source and target views, defined by

$$\mu = \frac{n_s w_s^T \mu_s + n_l w_t^T \mu_l}{n_s + n_l}, \tag{4}$$

with μ_s and μ_l being the mean of the labeled source and target samples.

Substituting Eq. (4) into Eq. (3), $w^T S_T w$ can be reformulated as

$$w^T S_T w = w^T \begin{bmatrix} S_{T,s} & 0 \\ 0 & S_{T,t} \end{bmatrix} w + \frac{n_s n_l}{n_s + n_l} w^T \begin{bmatrix} \mu_s \\ -\mu_l \end{bmatrix} \begin{bmatrix} \mu_s \\ -\mu_l \end{bmatrix}^T w, \tag{5}$$

where $S_{T,s}$ and $S_{T,t}$ are defined as $S_{T,s} = \sum_{i=1}^{n_s} (x_i^s - \mu_s)(x_i^s - \mu_s)^T$, $S_{T,t} = \sum_{i=1}^{n_l} (x_i^l - \mu_l)(x_i^l - \mu_l)^T$.

Therefore, S_T can be written as

$$S_T = \begin{bmatrix} S_{T,s} & 0 \\ 0 & S_{T,t} \end{bmatrix} + \frac{n_s n_l}{n_s + n_l} \begin{bmatrix} \mu_s \\ -\mu_l \end{bmatrix} \begin{bmatrix} \mu_s \\ -\mu_l \end{bmatrix}^T. \tag{6}$$

3.2. Between-class scatter matrix

The between-class scatter matrix $w^T S_B w$ in the common subspace is defined by

$$w^T S_B w = \sum_{j=1}^{c} n_j \left(\frac{w_s^T m_{sj} + w_t^T m_{tj}}{n_j} - \mu \right) \left(\frac{w_s^T m_{sj} + w_t^T m_{tj}}{n_j} - \mu \right)^T, \tag{7}$$

where n_j is the total number of the j-th class training samples from both source and target views, m_{sj} and m_{tj} represent the sum of the j-th class samples from source and target views, respectively. Combining Eqs. (4) and (7), we can rewrite $w^T S_B w$ as

$$w^T S_B w = \sum_{j=1}^{c} n_j w^T \left(\begin{bmatrix} \frac{m_{sj}}{n_j} \\ \frac{m_{lj}}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s}{n_s + n_l} \\ \frac{n_l \mu_l}{n_s + n_l} \end{bmatrix} \right) \left(\begin{bmatrix} \frac{m_{sj}}{n_j} \\ \frac{m_{lj}}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s}{n_s + n_l} \\ \frac{n_l \mu_l}{n_s + n_l} \end{bmatrix} \right)^T w, \tag{8}$$

and S_B can be formulated as

$$S_B = \sum_{j=1}^{c} n_j \left(\begin{bmatrix} \frac{m_{sj}}{n_j} \\ \frac{m_{tj}}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s}{n_s + n_l} \\ \frac{n_l \mu_l}{n_s + n_l} \end{bmatrix} \right) \left(\begin{bmatrix} \frac{m_{sj}}{n_j} \\ \frac{m_{tj}}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s}{n_s + n_l} \\ \frac{n_l \mu_l}{n_s + n_l} \end{bmatrix} \right)^T. \tag{9}$$

3.3. Between-view distribution difference matrix

The between-view distribution difference matrix $w^T S_D w$ is defined by

$$w^T S_D w = w^T \begin{bmatrix} \mu_s \\ -\mu_t \end{bmatrix} \begin{bmatrix} \mu_s \\ -\mu_t \end{bmatrix} w, \tag{10}$$

where $\mu_t = \frac{1}{n_l + n_u} (\sum_{i=1}^{n_l} x_i^l + \sum_{i=1}^{n_u} x_i^u)$ indicates the mean of all target samples, and S_D can be written as

$$S_D = \begin{bmatrix} \mu_s \\ -\mu_t \end{bmatrix} \begin{bmatrix} \mu_s \\ -\mu_t \end{bmatrix}. \tag{11}$$

3.4. Locality constraint matrix

As revealed in the previous studies [48–50], locality information is an important clue in manifold learning. In this work, we formulated the locality constraint matrix $w^T S_L w$ as

$$w^T S_L w = w^t \begin{bmatrix} X_s L_s X_s^T & 0 \\ 0 & X_t L_t X_t^T \end{bmatrix} w, \tag{12}$$

where $X_s = [x_1^s, x_2^s, \ldots, x_{n_s}^s] \in \mathbb{R}^{d_s \times n_s}$, $X_t = [x_1^l, x_2^l, \ldots, x_{n_l}^l, x_1^u, x_2^u, \ldots, x_{n_u}^u] \in \mathbb{R}^{d_t \times (n_l + n_u)}$ are the data matrices of the source and target view, L_s and L_t are the normalized graph Laplacian matrices of source view and target view, defined by

$$L_s = I_s - D_s^{1/2} A_s D_s^{1/2}$$

$$L_t = I_t - D_t^{1/2} A_t D_t^{1/2}, \tag{13}$$

where I_s and I_t are the identity matrices of size $n_s \times n_s$ and $(n_l + n_u) \times (n_l + n_u)$ respectively, the diagonal matrices D_s and D_t satisfy $D_{s(ii)} = \sum_j A_{s(ij)}$ and $D_{t(ii)} = \sum_j A_{t(ij)}$, the adjacency matrices $A_s \in \mathbb{R}^{n_s \times n_s}$ and $A_t \in \mathbb{R}^{(n_l + n_u) \times (n_l + n_u)}$ are respectively defined by

$$A_{s(ij)} = \begin{cases} \exp\left\{ -\dfrac{\| x_i^s - x_j^s \|^2}{2\sigma^2} \right\} & \text{if } x_i^s \in N_k(x_j^s) \text{ or } x_j^s \in N_k(x_i^s) \\ 0 & \text{otherwise} \end{cases}$$

$$A_{t(ij)} = \begin{cases} \exp\left\{ -\dfrac{\| x_i^t - x_j^t \|^2}{2\sigma^2} \right\} & \text{if } x_i^t \in N_k(x_j^t) \text{ or } x_j^t \in N_k(x_i^t) \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

with $x_i^t \in \mathbb{R}^{d_t}$ indicates the i-th target data (including the labeled or unlabeled target data), $N_k(x_i^s)$ is the k nearest neighbor set of $x_i{}^s$ in the source view, $N_k(x_i^t)$ is the k nearest neighbor set of $x_i{}^t$ in the target view. We set σ as the mean distance of all near neighbor pairs, i.e. $\sigma = \frac{1}{n \times k} \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} \| x_i - x_j \|$.

It is easy to verify that by minimizing $w^T S_L w$, the neighbor data in the input space are still close in the projected space, which keeps the smoothness of labeled and unlabeled data. According to Eq. (12), S_L can be formulated as

$$S_L = \begin{bmatrix} X_s L_s X_s^T & 0 \\ 0 & X_t L_t X_t^T \end{bmatrix}. \tag{15}$$

3.5. Solution

By solving the generalized eigen-decomposition problem

$$S_B w = \lambda (S_T + \gamma_1 S_D + \gamma_2 S_L) w \tag{16}$$

with its leading eigenvalues, we can obtain the optimal projection matrix w and split it into w_s and w_t as $w_s = w(1 : d_s, :)$, $w_t = w(d_s + 1 : d_s + d_t, :)$. Then we use the two projection matrices to map the data from heterogeneous spaces into the common space and apply k-Nearest Neighbor (k-NN) classifier to the projected labeled training data from both source and target views.

4. Heterogeneous Kernel Discriminant Analysis

Considering linear discriminant is not complex enough for most real-world data, we extend the aforesaid heterogeneous discriminant analysis from linear space to kernel space, in order to increase the expressiveness of the discriminant. Like in [51], in order to yield the projection matrices in the kernel space, we first map the data non-linearly into a high-dimensional kernel space and compute heterogeneous linear discriminant analysis there.

Let $\phi : x \rightarrow \mathcal{F}$ be the non-linear function used to map the data to a high-dimensional Hilbert space, where the data is more linearly separable. $\phi(x_i^s)$, $\phi(x_i^l)$ and $\phi(x_i^u)$ denote the transformed representation of $x_i{}^s$, $x_i{}^l$ and $x_i{}^u$ in the kernel space. Considering the kernel trick [33], we try to replace the explicit mapping with the inner product $k(x_i, x_j) = (\phi(x_i) \cdot \phi(x_j))$. According to the theory of reproducing kernels [52], the projections learned from the training samples lie in the span of all training samples in $\mathcal{F}$. Therefore, the

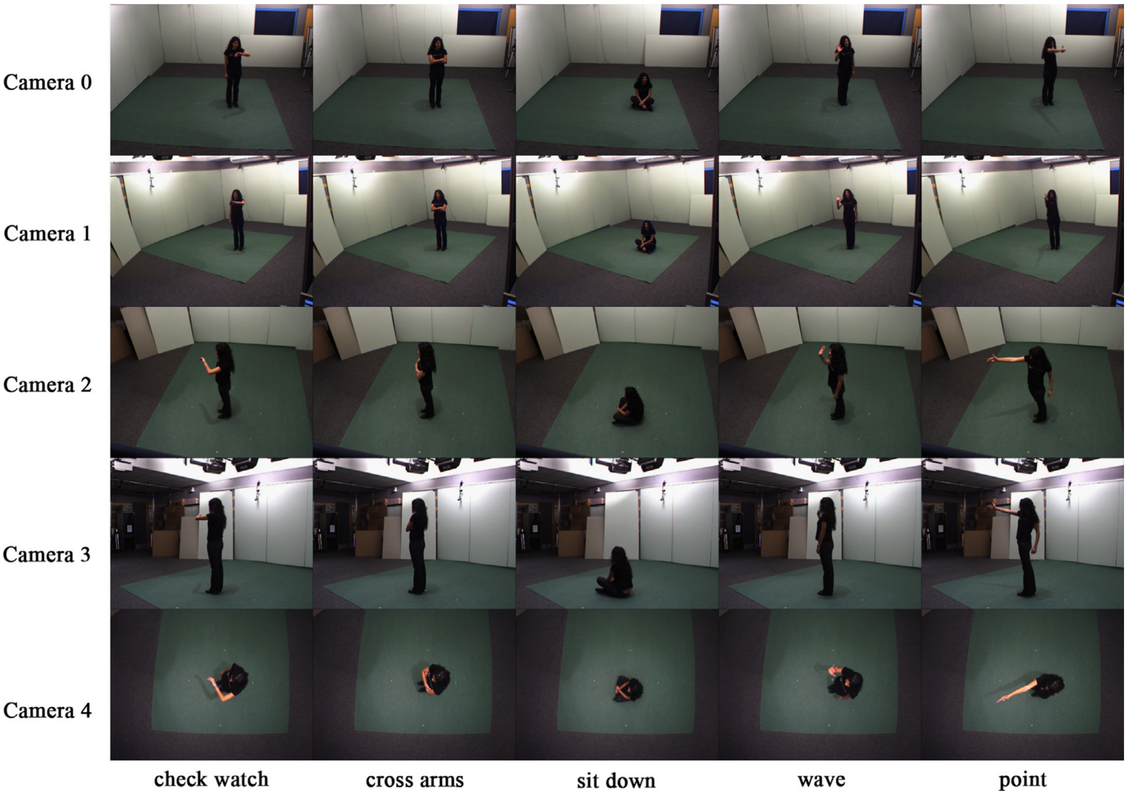


Fig. 2. Exemplar frames from IXMAS multi-view action dataset. Each column shows one action captured from different viewpoints.

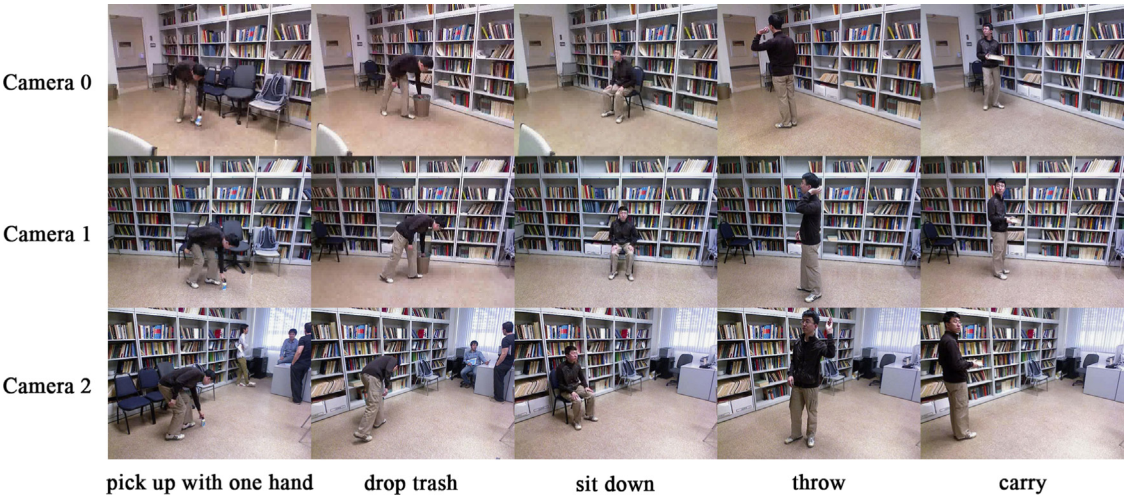


Fig. 3. Exemplar frames from Northwestern-UCLA dataset. Each column shows one action captured from different viewpoints.

projection matrices w_s and w_t can be linearly represented as

$$w_s = \sum_{i=1}^{n_s} \alpha_i \phi(x_i^s)$$

$$w_t = \sum_{i=1}^{n_l} \beta_i \phi(x_i^l) + \sum_{i=1}^{n_u} \beta_{n_l+i} \phi(x_i^u). \tag{17}$$

For the convenience of discussion, we denote $A = [\alpha_1, \ldots, \alpha_{n_s}]^T$, $B = [\beta_1, \ldots, \beta_{n_l+n_u}]^T$, and $v = \left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$. By using the definition of v, our objective function can be written as

$$v = \arg\max_v \frac{|v^T S_B^\phi v|}{|v^T S_T^\phi v| + \gamma_1 |v^T S_D^\phi v| + \gamma_2 |v^T S_L^\phi v|}, \tag{18}$$

where the total scatter matrix $w^T S_T^\phi w$, the between-class scatter matrix $w^T S_B^\phi w$, the between-view distribution difference matrix $w^T S_D^\phi w$ and the locality constraint matrix $w^T S_L^\phi w$ can be reformulated as follows.

4.1. Total scatter matrix

Substituting Eq. (17) into Eq. (5), we have

$$v^T S_T^\phi v = v^T \begin{bmatrix} S_{T,s}^\phi & 0 \\ 0 & S_{T,t}^\phi \end{bmatrix} v + \frac{n_s n_l}{n_s + n_l} v^T \begin{bmatrix} \mu_s^\phi \\ -\mu_l^\phi \end{bmatrix} \begin{bmatrix} \mu_s^\phi \\ -\mu_l^\phi \end{bmatrix}^T v, \tag{19}$$

where $S_{T,s}^\phi$ and $S_{T,t}^\phi$ are defined by

$$S_{T,s}^\phi = \sum_{i=1}^{n_s} (\zeta_i^s - \mu_s^\phi)(\zeta_i^s - \mu_s^\phi)^T$$

$$S_{T,t}^\phi = \sum_{i=1}^{n_l} (\zeta_i^l - \mu_l^\phi)(\zeta_i^l - \mu_l^\phi)^T, \tag{20}$$

in which $\zeta_i^s = [k(x_i^s, x_1^s), \ldots, k(x_i^s, x_{n_s}^s)]^T$ and $\zeta_i^l = [k(x_i^l, x_1^l), \ldots, k(x_i^l, x_{n_l}^l), k(x_i^l, x_1^u), \ldots, k(x_i^l, x_{n_u}^u)]^T$, $k(x_1, x_2)$ is the inner product of $\phi(x_1)$ and $\phi(x_2)$; $\mu_s = \frac{1}{n_s}\sum_{i=1}^{n_s} \zeta_i^s$ and $\mu_l = \frac{1}{n_l}\sum_{i=1}^{n_l} \zeta_i^l$ are the mean vectors of all ζ^s and ζ^l, respectively. Therefore, ${S_T}^\phi$ can be formulated as

$$S_T^\phi = \begin{bmatrix} S_{T,s}^\phi & 0 \\ 0 & S_{T,t}^\phi \end{bmatrix} + \frac{n_s n_l}{n_s + n_l} \begin{bmatrix} \mu_s^\phi \\ -\mu_l^\phi \end{bmatrix} \begin{bmatrix} \mu_s^\phi \\ -\mu_l^\phi \end{bmatrix}^T, \tag{21}$$

.

Table 1
Cross-view action recognition accuracies (%) of HKDA with different settings on the IXMAS dataset. The numbers are the mean recognition accuracies of each target view, e.g. Cam0 is the average accuracies when camera 0 is used as the target view and the other camera views are respectively used as the source views. Each time, only one camera view is used for the source view.

Methods	Cam0	Cam1	Cam2	Cam3	Cam4	Ave.
$\gamma_1 = \gamma_2 = 0$	67.36	67.77	70.83	69.85	51.79	65.52
$\gamma_1 = 0$	68.40	69.33	72.57	72.51	53.18	67.20
$\gamma_2 = 0$	70.83	70.54	72.92	71.70	56.42	68.48
Ours	72.57	72.11	74.88	73.84	58.04	70.29

4.2. Between-class scatter matrix

Combining Eqs. (8) and (17), we find

$$v^T S_B^\phi v = \sum_{j=1}^{c} n_j v^T \left(\begin{bmatrix} \frac{m_{sj}^\phi}{n_j} \\ \frac{m_{tj}^\phi}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s^\phi}{n_s + n_l} \\ \frac{n_l \mu_l^\phi}{n_s + n_l} \end{bmatrix} \right) \left(\begin{bmatrix} \frac{m_{sj}^\phi}{n_j} \\ \frac{m_{tj}^\phi}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s^\phi}{n_s + n_l} \\ \frac{n_l \mu_l^\phi}{n_s + n_l} \end{bmatrix} \right)^T v, \tag{22}$$

with m_{sj}^ϕ and m_{tj}^ϕ being the sum vectors of ζ^s and ζ^l from the j-th class, and ${S_B}^\phi$ can be formulated as

$$S_B^\phi = \sum_{j=1}^{c} n_j \left(\begin{bmatrix} \frac{m_{sj}^\phi}{n_j} \\ \frac{m_{tj}^\phi}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s^\phi}{n_s + n_l} \\ \frac{n_l \mu_l^\phi}{n_s + n_l} \end{bmatrix} \right) \left(\begin{bmatrix} \frac{m_{sj}^\phi}{n_j} \\ \frac{m_{tj}^\phi}{n_j} \end{bmatrix} - \begin{bmatrix} \frac{n_s \mu_s^\phi}{n_s + n_l} \\ \frac{n_l \mu_l^\phi}{n_s + n_l} \end{bmatrix} \right)^T. \tag{23}$$

4.3. Between-view distribution difference matrix

Considering the denominator, we can obtain a similar transformation as in Eq. (10), and ${S_D}^\phi$ can be formulated as

$$S_D^\phi = \begin{bmatrix} \mu_s^\phi \\ -\mu_t^\phi \end{bmatrix} \begin{bmatrix} \mu_s^\phi \\ -\mu_t^\phi \end{bmatrix}^T \tag{24}$$

where $\mu_t^\phi = \frac{1}{n_l + n_u}(\sum_{i=1}^{n_l} \zeta_i^l + \sum_{i=1}^{n_u} \zeta_i^u)$ is the mean vectors of all ζ^l and ζ^u, $\zeta_i^u = [k(x_i^u, x_1^l), \ldots, k(x_i^u, x_{n_l}^l), k(x_i^u, x_1^u), \ldots, k(x_i^u, x_{n_u}^u)]^T$.

4.4. Locality constraint matrix

In the same way, with the definition of Eqs. (12) and (17), we have

$$S_L^{\phi} = \begin{bmatrix} K_s L_s K_s^T & 0 \\ 0 & K_t L_t K_t^T \end{bmatrix}, \tag{25}$$

where $K_s = [\zeta_1^s, \zeta_2^s, \ldots, \zeta_{n_s}^s]$, $K_t = [\zeta_1^l, \zeta_2^l, \ldots, \zeta_{n_l}^l, \zeta_1^u, \zeta_2^u, \ldots, \zeta_{n_u}^u]$.

4.5. Solution

Similarly, by solving the generalized eigen-decomposition problem

$$S_B^{\phi} v = \lambda (S_T^{\phi} + \gamma_1 S_D^{\phi} + \gamma_2 S_L^{\phi}) v \tag{26}$$

A and B can be obtained from the optimal projection matrix $v = \left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$, then the optimal projection matrices w_s and w_t can be finally gotten by different modalities of Eq. (17).

5. Experiments

5.1. Datasets

IXMAS multi-view action dataset [22] consists of eleven daily-life actions: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, and pick up. Each action is performed three times by twelve subjects and taken from five different views including four side views and one top view. Fig. 2 shows some exemplar actions of the IXMAS dataset.

Northwestern-UCLA Multiview Action3D (N-UCLA) dataset [53] contains RGB, depth and human skeleton data captured simultaneously by three Kinect cameras. This dataset includes ten action categories: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, and carry. Each action is performed by ten subjects from one to six times. Several samples of the MVU dataset are shown in Fig. 3.

Table 2
Cross-view action recognition accuracies (%) of HKDA with different settings on the N-UCLA dataset. The numbers are the mean recognition accuracies of each target view, e.g. Cam0 is the average accuracies when camera 0 is used as the target view and the other camera views are respectively used as the source views. Each time, only one camera view is used for the source view.

Methods	Cam0	Cam1	Cam2	Ave.
$\gamma_1 = \gamma_2 = 0$	47.21	40.50	50.88	46.20
$\gamma_1 = 0$	47.68	40.63	54.24	47.52
$\gamma_2 = 0$	47.83	43.82	53.44	48.37
Ours	49.30	43.40	56.79	49.83

5.2. Setup

We extract the three dense trajectory features (i.e., HOG, HOF, MBH) proposed by Wang et al. [8]. For each descriptor, we use the

Table 3
Cross-view action recognition accuracies (%) of different heterogeneous transfer learning approaches on the IXMAS dataset. The numbers are the average recognition accuracies over all rounds of experiments.

source → target	SVMt	[40]	[41]	[42]	[44]	HLDA	HKDA
c0 → c1	59.95	60.19	61.11	65.51	68.29	63.43	71.99
c0 → c2	66.20	67.13	67.12	67.13	70.83	68.75	76.16
c0 → c3	66.90	65.74	65.97	69.44	68.98	67.13	74.31
c0 → c4	50.69	50.69	51.39	53.24	56.71	50.00	61.11
c1 → c0	62.73	61.11	61.57	67.13	66.44	64.35	72.92
c1 → c2	66.20	65.51	65.28	68.29	70.60	70.83	73.38
c1 → c3	66.90	65.28	65.51	69.44	68.98	65.05	74.31
c1 → c4	50.69	50.23	50.69	51.62	55.56	45.83	58.10
c2 → c0	62.73	63.19	62.04	66.20	65.74	63.19	74.07
c2 → c1	66.20	61.34	60.42	65.51	68.75	63.19	72.92
c2 → c3	66.90	65.97	66.20	69.44	68.75	64.35	71.76
c2 → c4	50.69	51.16	50.93	52.31	56.25	49.54	56.94
c3 → c0	62.73	62.27	61.81	66.20	65.97	62.27	72.22
c3 → c1	66.20	60.42	61.57	65.05	68.52	66.20	71.53
c3 → c2	66.90	67.13	67.36	67.82	70.83	68.06	73.84
c3 → c4	50.69	51.16	51.39	51.39	56.71	47.92	56.02
c4 → c0	62.73	63.66	63.19	66.67	65.97	64.35	71.06
c4 → c1	66.20	61.81	62.04	66.67	67.82	65.05	71.99
c4 → c2	66.90	69.68	70.14	68.52	70.83	70.14	76.16
c4 → c3	50.69	67.13	66.90	69.21	68.75	68.06	75.00
Average	61.30	61.57	61.63	64.34	66.06	62.38	70.29

Table 4
Cross-view action recognition accuracies (%) of different heterogeneous transfer learning approaches on the N-UCLA dataset. The numbers are the average recognition accuracies over all rounds of experiments.

source → target	SVMt	[40]	[41]	[42]	[44]	HLDA	HKDA
c0 → c1	39.82	40.75	41.02	39.55	42.75	38.60	43.12
c0 → c2	50.96	49.75	49.57	51.76	53.65	47.59	56.53
c1 → c0	44.04	48.41	47.88	45.98	47.52	41.23	49.60
c1 → c2	50.96	51.32	50.21	52.14	53.45	46.93	57.06
c2 → c0	44.04	47.84	47.09	46.36	47.91	41.24	49.01
c2 → c1	39.82	40.68	40.61	38.83	42.71	38.34	43.68
Average	44.94	46.46	46.06	45.77	48.00	42.32	49.83

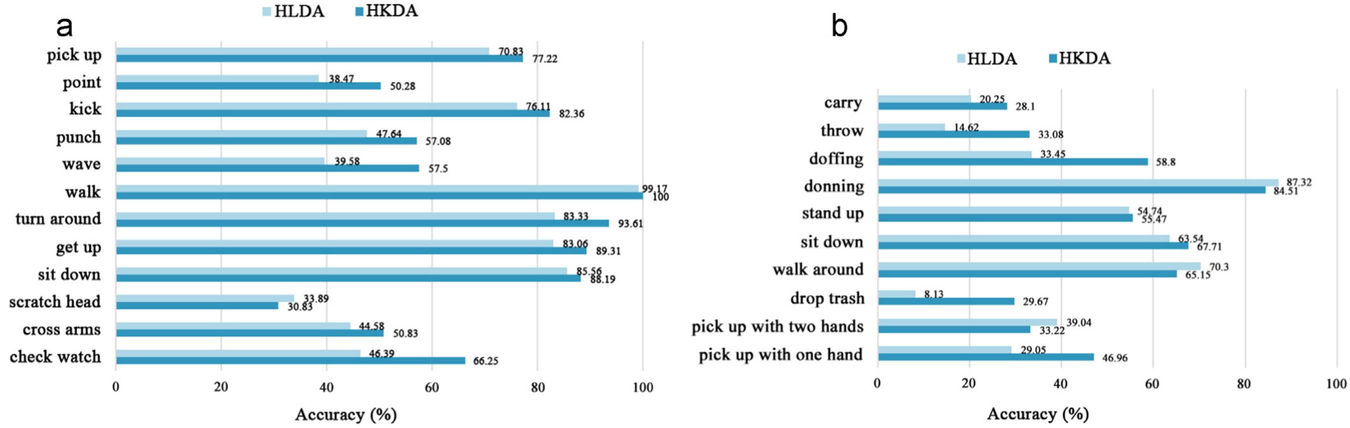


Fig. 4. Cross-view action recognition accuracies of HLDA and HKDA for each action. (a) Performances on the IXMAS dataset. (b) Performances on the N-UCLA dataset.

bag-of-words approach and uniformly set the number of visual words 400. The 800-dimensional HOG/HOF feature (the concatenation of 400-dimensional HOG and 400-dimensional HOF) is adopted for source view, and 400-dimensional MBH feature for target view.

When evaluating on these two datasets, we take one view as source view and another different view as target view. To enable appropriate verification, we look into all possible pairwise view combinations (20 in total for 5 views in the IXMAS dataset and 6 in total for 3 views in the N-UCLA dataset). The leave-one-subject-out cross validation strategy (i.e., 12-fold cross validation for the IXMAS dataset and 10-fold cross validation for the N-UCLA dataset) is employed. Specifically, for each time, we use videos of one subject from the target view for testing, the remaining videos (i.e. videos of the rest subjects from the target view and all subjects from the source view) for training, in which only one target subject and all source subjects labeled.

To verify the effectiveness of reducing the between-view distribution mismatch, we compare our method with a special version, which excludes the minimization of data distribution difference between source and target views, i.e., $\gamma_1 = 0$. To evaluate the contribution of the locality constraint, we perform our method with $\gamma_2 = 0$ for comparison. Likewise, to investigate the effect of

Table 5
Cross-view action recognition accuracies (%) of different transfer learning approaches on the IXMAS dataset. The numbers are the mean recognition accuracies of all possible pairwise view combinations.

source → target	[29]	[13]	[28]	[31]	[30]	[25]	[11]	[12]	HKDA
c0 → c1	47.2	83.7	63.6	64.9	71.5	80.3	94.8	92.7	76.2
c0 → c2	41.0	59.2	60.6	64.1	68.9	63.6	69.1	84.2	78.0
c0 → c3	61.8	57.4	61.2	67.1	67.3	68.5	83.9	83.9	78.5
c0 → c4	32.6	33.6	52.6	65.5	64.2	56.1	39.1	44.2	69.2
c1 → c0	44.4	84.3	61.0	63.6	70.5	80.0	90.6	95.5	79.6
c1 → c2	44.4	61.6	62.1	60.2	69.8	62.1	79.7	77.6	75.2
c1 → c3	57.6	62.8	65.1	66.7	74.2	59.7	79.1	86.1	77.3
c1 → c4	35.4	26.9	54.2	66.8	62.3	47.9	30.6	40.9	66.7
c2 → c0	45.8	62.5	63.2	65.4	67.8	63.6	72.1	82.4	79.4
c2 → c1	48.6	65.2	62.4	63.2	71.8	62.1	86.1	79.4	76.6
c2 → c3	54.2	72.0	71.7	67.1	79.2	79.7	77.3	85.8	77.3
c2 → c4	37.5	60.1	58.2	65.9	66.5	75.5	62.7	71.5	70.1
c3 → c0	43.8	57.1	64.2	65.4	68.7	67.0	82.4	82.4	79.6
c3 → c1	41.7	61.5	71.0	61.9	80.0	65.8	79.7	80.9	76.4
c3 → c2	43.1	71.0	64.3	65.4	70.4	83.6	70.9	82.7	78.7
c3 → c4	31.3	31.2	56.6	61.6	63.8	46.4	37.9	44.2	69.9
c4 → c0	41.0	39.6	50.0	65.8	55.4	54.5	48.8	57.1	79.4
c4 → c1	45.1	32.8	59.7	62.7	67.3	49.4	40.9	48.5	76.2
c4 → c2	41.0	68.1	60.7	64.5	72.6	72.1	70.3	78.8	77.3
c4 → c3	53.5	37.4	61.1	61.9	68.0	50.0	49.4	51.2	79.2
Average	44.6	56.4	61.2	64.5	69.2	64.4	67.4	72.5	76.0

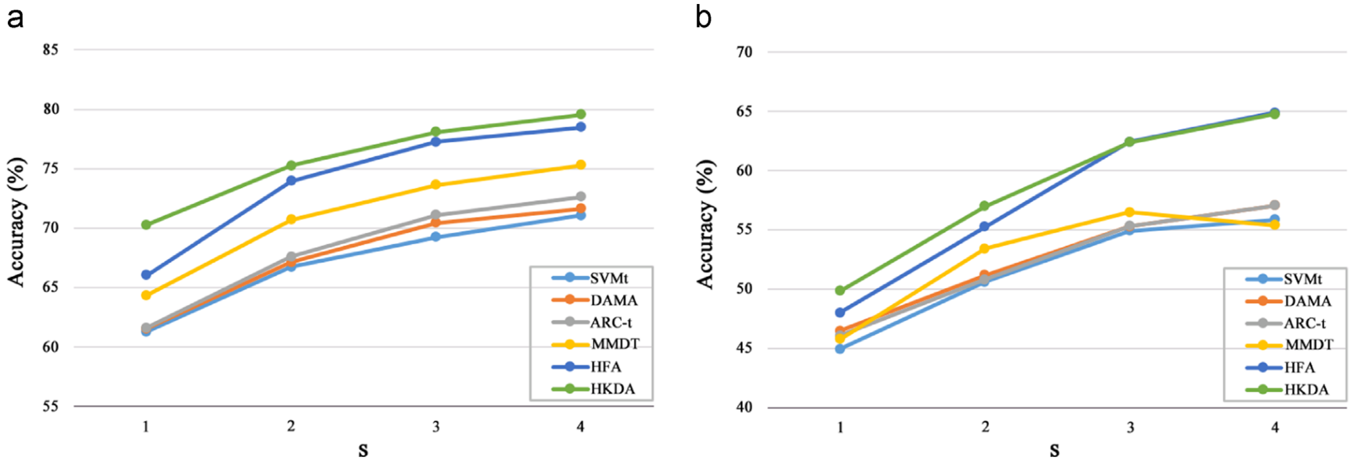


Fig. 5. Cross-view action recognition accuracies of all methods with respect to different numbers of labeled target training data (vary from one subject to four subjects, i.e., $s = 1, 2, 3, 4$). (a) Performances on the IXMAS dataset. (b) Performances on the N-UCLA dataset.

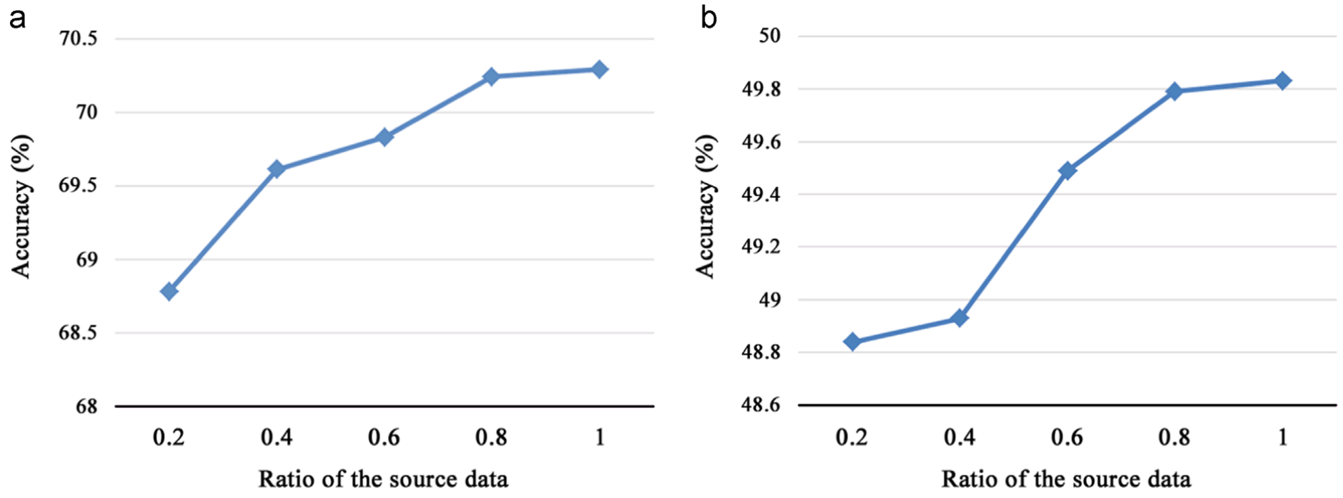


Fig. 6. Cross-view action recognition accuracies of HKDA with respect to different numbers of source data. (a) Performances on the IXMAS dataset. (b) Performances on the N-UCLA dataset.

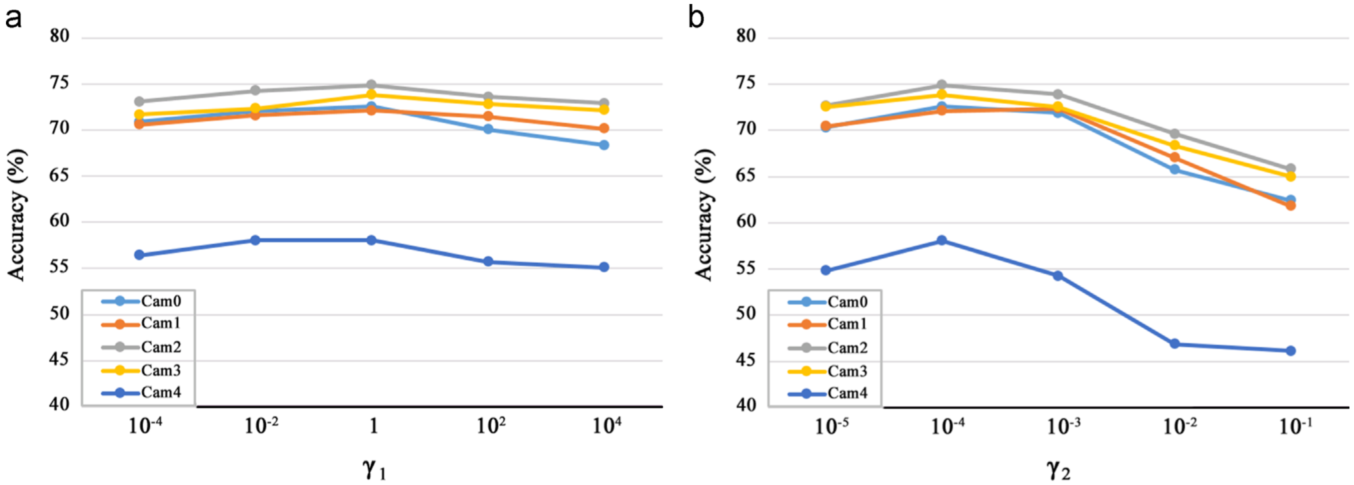


Fig. 7. Cross-view action recognition accuracies of HKDA using different parameters on the IXMAS dataset. Every line corresponds to one target view and the results are the mean recognition accuracies of each target view. (a) Performances w.r.t. γ_1. (b) Performances w.r.t. γ_2.

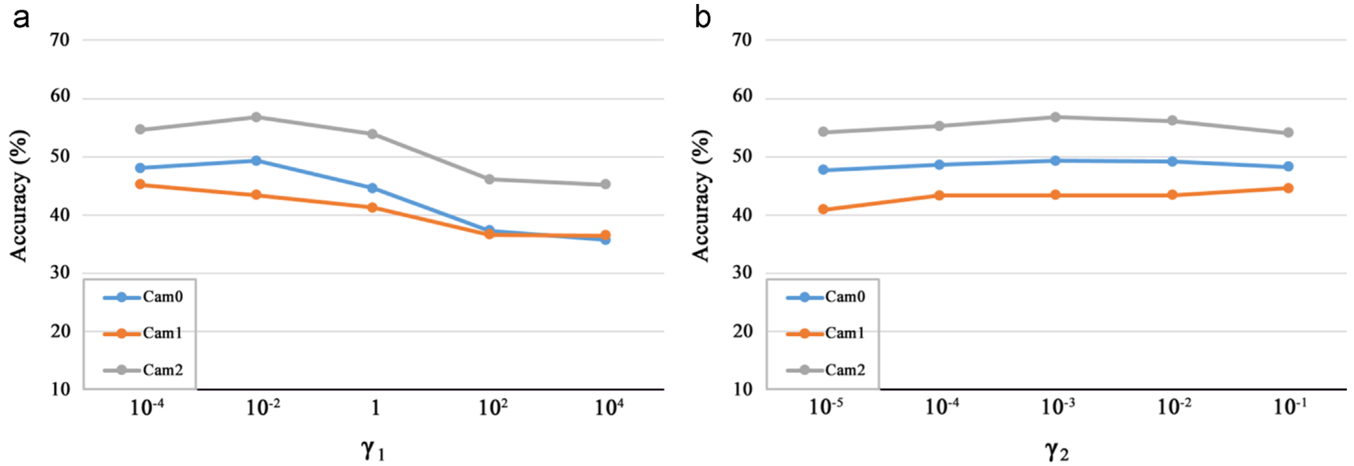


Fig. 8. Cross-view action recognition accuracies of HKDA using different parameters on the N-UCLA dataset. Every line corresponds to one target view and the results are the mean recognition accuracies of each target view. (a) Performances w.r.t. γ_1. (b) Performances w.r.t. γ_2.

unlabeled target-view data, we also report the results when $\gamma_1 = \gamma_2 = 0$ in the objective function.

In addition, as the source and target data are represented by feature vectors with different dimensions and there is no data-to-data correspondence, we compare our method with the state-of-the-art methods of transfer learning over heterogeneous feature spaces. In particular, the compared algorithms are listed as follows:

- SVMt: It utilizes the labeled target training data to train a standard support vector machine (SVM) classifier for each action class.
- DAMA [40]: It assumes the manifold structure on the dataset and learns a common feature subspace for all heterogeneous domains by simultaneously maximizing the intra-domain similarity and minimizing the inter-domain similarity.
- ARC-t [41]: A category general feature transform method that learns an asymmetric kernel transformation to transfer feature knowledge between the source and target domains by using the labeled training data from both domains.
- MMDT [42]: It uses the labeled training data from the source and target domains to learn an asymmetric category independent transform, which combines the strengths of max-margin learning with the flexibility of the feature transform.
- HFA [43]: A max-margin transform approach that augments the heterogeneous features from the source and target domains by using two feature mapping functions, respectively.

For DAMA [40], after finding the projection matrices, SVM and k-NN (k=5) are applied to train their final classifiers using the projected training data of pairwise views. For ARC-t [41], we construct the kernel matrix based on the learned asymmetric transformation metric, and then also apply SVM to train the final classifier. For MMDT [42], we find the transformation matrix using the max-margin constraints and simultaneously learn the final SVM classifier. For HFA [44], we obtain the two projection matrices using the standard SVM with the hinge loss.

For our method, the parameters are empirically set as $\gamma_1 = 1$, $\gamma_2 = 10^{-4}$ for all the datasets which generally achieves better results. And the parameter k for calculating the adjacency matrix in the Section 3.4 is fixed to 3. For other methods, we report their best results on the test data by varying their parameters in a wide range on each dataset. Specifically, we validate the parameters μ in DAMA (see Theorem 1 in [40]), λ in ARC-t (see Eq. (1) in [41]), λ in HFA (see Eq. (13) in [44]), C_S and C_T in MMDT (see Eq. (4) in [42]) from $\{10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}\}$. For all these methods, the trade-off parameter C in SVM is fixed as the default value (i.e., $C=1$) and the RBF kernel is selected for fair comparison.

5.3. Results

Tables 1 and 2 demonstrate the recognition accuracies of our method with different settings on the IXMAS dataset and N-UCLA dataset. From the result, we can observe that: (1) our method can successfully deal with the cross-view action recognition over heterogeneous feature space by minimizing the data distribution difference between source and target views; (2) the local structure of data is helpful for our method to achieve better results; (3) exploiting the information contained in the unlabeled data can contribute to improve the action recognition performance.

Fig. 4 illustrates the cross-view action recognition accuracies of HLDA and HKDA on the IXMAS dataset and N-UCLA dataset. It can be seen that the heterogeneous discriminant analysis in kernel space (HKDA) has a obvious improvement over that in linear space (HLDA) for most action classes.

Tables 3 and 4 give the cross-view action recognition accuracies of all methods on the IXMAS dataset and N-UCLA dataset, respectively. From the results, we have the following observations. The heterogeneous discriminant analysis in kernel space (HKDA) has large improvements over that in linear space (HLDA) for all possible source–target view combinations. Our HKDA performs better than other methods on the mean recognition accuracies, which clearly demonstrate the effectiveness of our proposed method. Compared with MMDT [42] and ARC-t [41], our better performance may arise from the utilization of unlabeled target training data. These data contribute to cope with the data distribution mismatch and keep the local manifold structure. The explanation for the better performance of HKDA than DAMA [40] may be the lack of the strong manifold structure on these two datasets.

We also design another experiment by using different numbers of labeled target training data (vary from one subject to four subjects). As shown in Fig. 5, all methods perform better with a larger labeled target sample number, and our HKDA achieves a considerable improvement over all other methods for most cases. In order to further verify our method and the effect of source view samples, we also plot the recognition results of our method by using different numbers of source data in Fig. 6. From the result, source data contribute to the overall performance improvement and the discriminative information is effectively transferred between views.

In addition, we compare our HKDA against some other methods [11–13,25,28–31] of cross-view action recognition, which cannot deal with the data from heterogeneous spaces. For this experiment, 800-dimensional HOG/HOF feature is adopted for both source and target views. We use 1NN classifier and a 6-fold cross validation procedure (identical cross-validation procedure as in [25,28,30,31]). Our method improves the average recognition accuracy by 3.5% compared to the next best approach (see Table 5).

We conduct experiments on the two datasets to evaluate the performance variations of our HKDA by using different parameters (i.e., γ_1, γ_2). In order to evaluate the performance variations, at each time we vary one parameter and set another parameter as the default values (i.e., $\gamma_1 = 1$, $\gamma_2 = 10^{-4}$). The results by varying different parameters on the IXMAS dataset and N-UCLA dataset are plotted in Figs. 7 and 8, respectively. From the result, we can observe that our HKDA is quite stable to the parameters in certain ranges. Specifically, by changing γ_1 in the range of $[10^{-4}, 10^2]$, the performances of HKDA vary within 2% in terms of mean recognition accuracy, which are still better than other methods reported in Tables 3 and 4. We also evaluate our HKDA by varying γ_2 in the range of $[10^{-5}, 10^{-1}]$. Our HKDA is also generally stable and better than other methods for most cases with the setting $\gamma_2 \in [10^{-5}, 10^{-3}]$.

6. Conclusions

We have presented a method for cross-view action recognition over heterogeneous feature spaces, and respectively accomplished the heterogeneous discriminant analysis in linear and kernel space. We propose to seek a discriminant common subspace, where the samples from different views are comparable. In this way, two different matrices are learned to map the data from source and target views into the common subspace, by simultaneously maximizing the similarity of intra-class samples, minimizing the similarity of inter-class samples, and reducing the mismatch between data distributions of different views and keeping the data locality. Our proposed method is capable of exploring the effective information contained in the labeled samples and unlabeled target samples, without any instance correspondence between two views. The promising results of our approach have been achieved on the IXMAS dataset and N-UCLA dataset for cross-view action recognition.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No. 61203274 and No. 61472038.

References

[1] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and Viterbi path searching, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Minneapolis, MN, 2007, pp. 1–8.

[2] S. Xiang, F. Nie, Y. Song, C. Zhang, Contour graph based human tracking and action sequence recognition, Pattern Recognit. 41 (12) (2008) 3653–3664.

[3] Z. Lin, Z. Jiang, L.S. Davis, Recognizing actions by shape-motion prototype trees, in: 2009 IEEE 12th International Conference on Computer Vision (ICCV), IEEE, Kyoto, 2009, pp. 444–451.

[4] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: 2003 IEEE 9th International Conference on Computer Vision (ICCV), IEEE, Nice, France, 2003, pp. 726–733.

[5] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2005 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, IEEE, Beijing, 2005, pp. 65–72.

[6] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, IEEE, San Diego, CA, 2005, pp. 984–989.

[7] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Computer Vision—ECCV 2008, Springer, Marseille, France, 2008, pp. 650–663.

[8] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI,, 2011, pp. 3169–3176.

[9] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (1) (2013) 60–79.

[10] H. Wang, C. Schmid, Action recognition with improved trajectories, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, Sydney, NSW, 2013, pp. 3551–3558.

[11] A. Gupta, J. Martinez, J. J. Little, R. J. Woodham, 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, OH, 2014, pp. 2601–2608.

[12] H. Rahmani, A. Mian, Learning a non-linear knowledge transfer model for cross-view action recognition, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2458–2466.

[13] B. Li, O. Camps, M. Sznaier, Cross-view activity recognition using hankelets, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, 2012, pp. 1362–1369.

[14] J. Zheng, Z. Jiang, P.J. Phillips, R. Chellappa, Cross-view action recognition via a transferable dictionary pair, in: BMVC, vol. 1, 2012, p. 7.

[15] D. Wu, L. Shao, Multi-max-margin support vector machine for multi-source human action recognition, Neurocomputing 127 (2014) 98–103.

[16] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, Z.-X. Yang, Single/multi-view human action recognition via regularized multi-task learning, Neurocomputing 151 (2015) 544–553.

[17] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Int. J. Comput. Vis. 50 (2) (2002) 203–226.

[18] V. Parameswaran, R. Chellappa, View invariance for human action recognition, Int. J. Comput. Vis. 66 (1) (2006) 83–101.

[19] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Providence, RI, 2012, pp. 20–27.

[20] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 914–927.

[21] R. Li, T.-P. Tian, S. Sclaroff, Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series, in: 2007 IEEE 11th International Conference on Computer Vision (ICCV), IEEE, Rio de Janeiro, 2007, pp. 1–8.

[22] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: 2007 IEEE 11th International Conference on Computer Vision (ICCV), IEEE, Rio de Janeiro, 2007, pp. 1–7.

[23] P. Yan, S.M. Khan, M. Shah, Learning 4d action feature models for arbitrary view action recognition, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE, Anchorage, AK, 2008, pp. 1–7.

[24] I.N. Junejo, E. Dexter, I. Laptev, P. Perez, View-independent action recognition from temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 172–185.

[25] A. Ciptadi, M.S. Goodwin, J.M. Rehg, Movement pattern histogram for action recognition and retrieval, in: Computer Vision—ECCV 2014, Springer, Zurich, Switzerland, 2014, pp. 695–710.

[26] A. Farhadi, M.K. Tabrizi, Learning to recognize activities from the wrong view point, in: Computer Vision—ECCV 2008, Springer, Marseille, France, 2008, pp. 154–166.

[27] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, in: 2011 IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, CO, 2011, pp. 3209–3216.

[28] R. Li, T. Zickler, Discriminative virtual views for cross-view action recognition, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, 2012, pp. 2855–2862.

[29] X. Wu, H. Wang, C. Liu, Y. Jia, Cross-view action recognition over heterogeneous feature spaces, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, Sydney, NSW, 2013, pp. 609–616.

[30] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, C. Shi, Cross-view action recognition via a continuous virtual path, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, OR, 2013, pp. 2690–2697.

[31] J. Zheng, Z. Jiang, Learning view-invariant sparse representations for cross-view action recognition, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, Sydney, VIC, 2013, pp. 3176–3183.

[32] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, Cross-view action recognition using contextual maximum margin clustering, IEEE Trans. Circuits Syst. Video Technol. 24 (10) (2014) 1663–1668.

[33] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: Proceedings of Computational Learning Theory, Springer, Amsterdam, Netherlands, 2001, pp. 416–426.

[34] C. Jia, Y. Kong, Z. Ding, Y.R. Fu, Latent tensor transfer learning for rgb-d action recognition, in: Proceedings of the ACM International Conference on Multimedia, ACM, New York, NY, 2014, pp. 87–96.

[35] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.

[36] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, England, 2004.

[37] L. Wang, H. Sahbi, Nonlinear cross-view sample enrichment for action recognition, in: Computer Vision-ECCV 2014 Workshops, Springer, Zurich, Switzerland, 2014, pp. 47–62.

[38] Y.-R. Yeh, C.-H. Huang, Y.-C. Wang, Heterogeneous domain adaptation and classification by exploiting the correlation subspace, IEEE Trans. Image Process. 23 (5) (2014) 2009–2018.

[39] X. Shi, Q. Liu, W. Fan, P.S. Yu, R. Zhu, Transfer learning on heterogenous feature spaces via spectral transformation, in: 2010 IEEE 10th International Conference on Data Mining (ICDM), IEEE, Sydney, NSW, 2010, pp. 1049–1054.

[40] C. Wang, S. Mahadevan, Heterogeneous domain adaptation using manifold alignment, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), vol. 22, 2011, p. 1541.

[41] B. Kulis, K. Saenko, T. Darrell, What you saw is not what you get: domain adaptation using asymmetric kernel transforms, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, CO, 2011, pp. 1785–1792.

[42] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, K. Saenko, Efficient learning of domain-invariant image representations, in: International Conference on Learning Representations (ICLR), 2013.

[43] L. Duan, D. Xu, I. Tsang, Learning with augmented features for heterogeneous domain adaptation, in: Proceedings of the International Conference on Machine Learning (ICML), Omnipress, Edinburgh, Scotland, 2012, pp. 711–718.

[44] W. Li, L. Duan, D. Xu, I.W. Tsang, Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1134–1148.

[45] Z. Ding, Y. Fu, Low-rank common subspace for multi-view learning, in: 2014 IEEE International Conference on Data Mining (ICDM), IEEE, Shenzhen, 2014, pp. 110–119.

[46] M. Shao, D. Kit, Y. Fu, Generalized transfer subspace learning through low-rank constraint, Int. J. Comput. Vis. 109 (1–2) (2014) 74–93.

[47] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Salt Lake City, UT, 1990.

[48] X. Zhao, X. Li, C. Pang, S. Wang, Human action recognition based on semi-supervised discriminant analysis with global constraint, Neurocomputing 105 (2013) 45–50.

[49] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.

[50] F.R. Chung, Spectral Graph Theory, CBMS Regional Conference Series in Mathematics, no. 92, American Mathematical Society, Providence, RI, vol. 5(2), 1997.

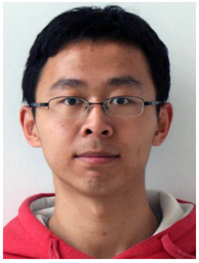
[51] B. Scholkopft, K.-R. Mullert, Fisher discriminant analysis with kernels, in: Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA, 1999, pp. 23–25.

[52] S. Saitoh, Theory of Reproducing Kernels and its Applications, vol. 189, Longman Scientific & Technical, Harlow, U.K., 1988.

[53] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning, and recognition, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, OH, 2014, pp. 2649–2656.

Wanchen Sui received the B.S. degree from Shandong Agricultural University in 2013. She is currently pursuing the M.S. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. Her research interests include computer vision, action recognition.

Xinxiao Wu received the B.A. degree in computer science from the Nanjing University of Information Science and Technology in 2005 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. She is currently an associate professor in the School of Computer Science at the Beijing Institute of Technology. Her research interests include machine learning, computer vision, and human action perception.



Yang Feng received the B.S. degree from Jilin University in 2010 and the M.S. degree from Beijing Institute of Technology in 2015. He is currently pursuing the Ph.D. degree at the University of Rochester. His research interests include computer vision, machine learning, and video retrieval.

Yunde Jia is Professor of Computer Science at BIT, and serves as the Director of the Beijing Laboratory of Intelligent Information Technology. He 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 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 at the Australian National University in 2011. His current research interests include computer vision, media computing, and intelligent systems.