Pattern Recognition 43 (2010) 4190–4197

Contents lists available at ScienceDirect

Pattern Recognition

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

Incremental discriminant-analysis of canonical correlations for action recognition

Xinxiao Wu, Yunde Jia*, Wei Liang

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P.R. China

ARTICLE INFO

Article history:
Received 18 October 2009
Received in revised form
12 June 2010
Accepted 2 July 2010

Keywords:
Human action recognition
Incremental discriminant-analysis
Computer vision

ABSTRACT

Human action recognition from video sequences is a challenging problem due to the large changes of human appearance in the cases of partial occlusions, non-rigid deformations, and high irregularities. It is difficult to collect a large set of training samples to learn the discriminative model with covering all possible variations of an action. In this paper, we propose an online recognition method, namely incremental discriminant-analysis of canonical correlations (IDCC), in which the discriminative model is incrementally updated to capture the changes of human appearance, and thereby facilitates the recognition task in changing environments. As the training sets are acquired sequentially instead of being given completely in advance, our method is able to compute a new discriminant matrix by updating the existing one using the eigenspace merging algorithm. Furthermore, we integrate our method into the graph-based semi-supervised learning method, linear neighbor propagation, to deal with the limited labeled training data. Experimental results on both Weizmann and KTH action data sets show that our method performs better than state-of-the-art methods on accuracy and efficiency.

1. Introduction

Human action recognition from video sequences has recently attracted increasing interests from computer vision for a wide range of promising applications, such as video indexing, visual surveillance, human–computer interaction, sports video analysis, and intelligent systems. As motion speeds and body sizes are associated with individuals, the same action executed by different persons may exhibit large variations; while the environmental conditions such as lighting and view point may make the observations of different actions similar. In order to reduce the variations of actions within the same class and suppress the environmental contributions to the similarities of actions in different classes, our method utilizes a discriminant matrix to maximize the canonical correlations of actions within the same class and minimize the canonical correlations of actions between different classes. Each action is represented by an orthogonal linear subspace of 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 many cases several principal images even a single image is sufficient to recognize what a person is doing.

Another problem in action recognition rises from high irregularities of actions undergoing various non-stationary scenarios. Taking the walking action for example, people may walk with a dog, walk when swinging a bag, walk with a briefcase or walk when being partially occluded by other objects. It is difficult to account for all the possible variations of an action during learning the discriminative model. Therefore, we aim to find a discriminative model that can be online learned to describe the changes of human appearance and accurately classify the actions even in the irregular performances. Our method is capable of online updating the discriminative model with capturing images as the new training data, and the updated discriminative model can reflect the appearance variations. By merging eigenspace models [1], our method updates the principal components of the total canonical correlations and between-class canonical correlations separately, and then computes the discriminant matrix directly from both updated principal component sets. To improve the computation efficiency of eigen-analysis, the sufficient spanning set [1] is adopted in the solution.

Our method is an incremental learning solution for discriminant analysis of canonical correlations (DCC) [20], yielding approximate solution of the batch computation with far lower computational cost. The discriminant matrix can be incrementally updated when more data becomes available without requiring the complete training data to be given in advance. Moreover, we

* Corresponding author. Tel./fax: +86 10 6891 4849.
E-mail addresses: wuxinxiao@bit.edu.cn (X. Wu), jiayunde@bit.edu.cn (Y. Jia), liangwei@bit.edu.cn (W. Liang).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.07.012

apply our method to the graph-based semi-supervised method, linear neighbor propagation [2], to handle with the situation where the training sets are partly labeled. It is able to continuously label the new data sets and simultaneously update the discriminant transformation matrix.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the background of our method. The *incremental discriminant-analysis of canonical correlations* (IDCC) is presented in Section 4. Section 5 demonstrates experimental results. The integration of our method with the semi-supervised learning method is given in Section 6. We draw conclusion and discuss future work in Section 7. A preliminary version of this paper appeared in ICCV 2009 [3].

2. Related work

2.1. Action recognition

Many approaches for human action recognition have been proposed in the past decade. Wang and Suter [4] used kernel principal component analysis to obtain the low-dimensional representation of human silhouette, and introduced factorial conditional random field to model the motion, with the aim of recognizing human activities performed by different people with different body builds as well as different motion styles and speeds. Jia and Yeung [5] reported a manifold embedding method to discover both the local spatial and temporal discriminant structures of human silhouette. They designed a two-stage recognition scheme to improve the recognition on low sensitivity to the temporal shape variation in the same action. Rodriguez et al. [6] introduced a template-based method for action recognition which is capable of capturing intra-class variability by synthesizing a single action MACH filter for a given action class. Jhuang et al. [7] applied a biological model of motion processing to the action recognition by accounting for the dorsal stream of the visual cortex. Some other approaches could extract local spatio-temporal features to exploit rich and intrinsic representation, and introduce statistical models to classify the action in large intra-class variations [8–14].

However, most of these approaches offline learn the recognition model and lack the adaptability to classify the different irregular actions which are not included in the training data. Our method is able to online update the discriminative model to learn the changes of human appearance with superior adaptability to recognize high irregular actions.

2.2. Incremental learning

Recently, a number of incremental learning approaches have been proposed and applied to computer vision. Hall et al. [15] proposed incremental principle component analysis (IPCA) based on the update of covariance matrix through a residue estimating procedure. Later on they improved their method by merging and splitting eigenspace models that allow a chunk of new samples to be learned in a single step [1]. Pang et al. [16] proposed an incremental linear discriminant analysis (ILDA) in two forms: sequential ILDA and chunk ILDA. The discriminant eigenspace is updated for classification when bursts of data are added to an initial discriminant eigenspace in the form of random chunks. As an improvement of ILDA, Kim et al. [17] applied the concept of the sufficient spanning set approximation in updating the between-class scatter matrix, the projected data matrix, and the total scatter matrix. Lin et al. [18] handled with the online update of discriminative models for tracking objects undergoing large pose and lighting changes. In the image set-based recognition, Kim et al. [19] proposed an incremental method of learning orthogonal subspace. With the concept of the sufficient spanning set, the algorithm separately updates the principal components of the class correlation and total correlation matrices, and then computes the orthogonal components of the updated few principal components.

Most of these incremental learning methods do not exploit the concept of multiple image sets in a single class. Considering that multiple image sets represent multiple actions executed by different subjects, our method maximizes the canonical correlations between multiple image sets within the same class, and is more robust to the intra-class changes.

3. Background

Table 1 demonstrates the important notations used throughout the paper.

Discriminant-analysis of canonical correlations (DCC) [20] 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 $\{X_1,X_2,\ldots X_m\}$, here 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 $P_i \in R^{N\times d}$ s.t. $X_i X_i^T = P_i \Lambda_i P_i^T$. Λ_i is the d largest eigenvalues, P_i the corresponding eigenvectors, and N the dimension of column vector. The discriminant transformation matrix $T=[t_1,\ldots,t_n]\in R^{N\times n}$ is defined by $Y_i=T^T X_i$ to make the transformed image sets more discriminative using canonical correlations. Orthonormal basis matrices of the subspaces of the transformed data are given by

$$Y_i Y_i^T = (T^T X_i)(T^T X_i)^T = (T^T P_i)\Lambda_i (T^T P_i)^T. \tag{1}$$

Canonical correlations are only defined for orthonormal basis matrices of subspace. Because $T^T P_i$ is not generally orthonormal, the matrix P_i is normalized to P_i' so that the columns of $T^T P_i'$ are orthonormal. By the SVD computation $(T^T P_i')^T (T^T P_j') = Q_{ij} \Lambda Q_{ji}^T$, the similarity of two transformed data sets is defined as the sum of canonical correlations

$$F_{ij} = \max_{Q_{ij},Q_{ji}} tr\{T^T P_j' Q_{ji} Q_{ij}^T P_i'^T T\}. \tag{2}$$

T is determined to maximize the similarities of any pairs of within-class sets and minimize the similarities of pair-wise sets of different classes by

$$T = \arg\max_T \frac{\sum_{i=1}^m \sum_{k\in W_i} F_{ik}}{\sum_{i=1}^m \sum_{l\in B_i} F_{il}}. \tag{3}$$

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

Table 1
Notations.

Notations	Descriptions
X_i	ith image set with each column describing an image
C_i	class label of X_i
P_i	orthonormal basis matrix representing the linear subspace of X_i
T	discriminant transformation matrix
S_b, S_t	transformed canonical correlations of between-class sets and total sets
V, Σ	eigenvector and eigenvalue matrices of S_b
U, Λ	eigenvector and eigenvalue matrices of S_t
m	number of image sets

set of class C_i. By the simple linear algebra

$$T^T P_j' Q_{ji} Q_{ij}^T P_i'^T T = I - T^T (P_j' Q_{ji} - P_i' Q_{ij})(P_j' Q_{ji} - P_i' Q_{ij})^T T/2, \quad (4)$$

the discriminative function in Eq. 3 is rewritten as

$$T = \underset{T}{\operatorname{argmax}} \frac{tr(T^T S_b T)}{tr(T^T S_w T)}, \quad (5)$$

where $S_b = \sum_{i=1}^{m} \sum_{l \in B_i} (P_l' Q_{li} - P_i' Q_{il})(P_l' Q_{li} - P_i' Q_{il})^T$, $B_i = \{j | C_j \neq C_i\}$, $S_w = \sum_{i=1}^{m} \sum_{k \in W_i} (P_k' Q_{ki} - P_i' Q_{ik})(P_k' Q_{ki} - P_i' Q_{ik})^T$, $W_i = \{j | C_j = C_i\}$. Finally, the optimal T is computed by eigen-decomposition of $(S_w)^{-1} S_b$. Without losing generality, we assume in the rest of the paper all the P_i are normalized.

4. Incremental discriminant-analysis of canonical correlations (IDCC)

Incremental discriminant-analysis of canonical correlations (IDCC) is able to update the discriminant transformation matrix for classification when new training data is being added. It includes three steps: updating the canonical correlations of the total sets, updating the canonical correlations of the between-class sets, and computing the discriminant transformation matrix. Two equivalent criterions to obtain the discriminant transformation matrix T are given by

$$\max_{\text{arg}T} \frac{\sum_{i=1}^{m} \sum_{k \in W_i} F_{ik}}{\sum_{i=1}^{m} \sum_{l \in B_i} F_{il}} = \max_{\text{arg}T} \frac{\sum_{i=1}^{m} \sum_{h \in T_i} F_{ih}}{\sum_{i=1}^{m} \sum_{l \in B_i} F_{il}}. \quad (6)$$

$W_i = \{j | C_j = C_i\}$, $B_i = \{j | C_j \neq C_i\}$, $T_i = W_i \cup B_i$ indexes the total sets. $\sum_{i=1}^{m} \sum_{k \in W_i} F_{ik}$ and $\sum_{i=1}^{m} \sum_{l \in B_i} F_{il}$ represent the canonical correlations of within-class sets and between-class sets, respectively. The total canonical correlations are represented by $\sum_{i=1}^{m} \sum_{h \in T_i} F_{ih} = \sum_{i=1}^{m} \sum_{k \in W_i} F_{ik} + \sum_{i=1}^{m} \sum_{l \in B_i} F_{il}$. In this paper, the algorithm uses the second criterion in Eq. (6). By the simple linear algebra (Eq. (4)), the discriminative function is

$$T = \underset{T}{\operatorname{argmax}} \frac{tr(T^T S_b T)}{tr(T^T S_t T)}, \quad (7)$$

where $S_b = \sum_{i=1}^{m} \sum_{l \in B_i} (P_l Q_{li} - P_i Q_{il})(P_l Q_{li} - P_i Q_{il})^T$, $B_i = \{j | C_j \neq C_i\}$, $S_t = \sum_{i=1}^{m} \sum_{h \in T_i} (P_h Q_{hi} - P_i Q_{ih})(P_h Q_{hi} - P_i Q_{ih})^T$, $T_i = \{j | j = 1, \ldots, m\}$.

Note that S_b and S_t are the linear algebra transformation (see Eq. (4)) of between-class canonical correlations and total canonical correlations, respectively, so our algorithm actually involves the update of principal components of S_t, the update of principal components of S_b, and the computation of T from the updated S_t and S_b.

4.1. Updating the total canonical correlations

Let the total canonical correlations of existing image sets be $\{U, \Delta, P_i, i = 1, 2, \ldots, m\}$, here $U$and Δ, respectively, denote the eigenvectors and eigenvalues of S_t, s.t. $S_t = U \Delta U^T$, and $P_i \in R^{N \times d}$ is a normalized orthonormal basis matrix of the ith existing set. Assume P_{m+1} is the normalized orthonormal basis matrix of a new data set, the update is defined by

$$\xi_1(U, \Delta, P_i, P_{m+1}) = (U', \Delta') \quad i = 1, 2, \ldots, m. \quad (8)$$

Assume

$$A = 2 \sum_{i=1}^{m} (P_{m+1} Q_{m+1,i} - P_i Q_{i,m+1})(P_{m+1} Q_{m+1,i} - P_i Q_{i,m+1})^T, \quad (9)$$

where $Q_{m+1,i}$ and $Q_{i,m+1}$ are obtained by the SVD solution $P_i^T T T^T P_{m+1} = Q_{i,m+1} \Lambda Q_{m+1,i}^T$, then the updated S_t is computed by $S_t' = S_t + A = U \Delta U^T + A$. We wish to calculate the eigenvectors U' and eigenvalues Δ' of S_t', i.e. $S_t' = U' \Delta' U'^T$. To reduce the dimension of eigenvalue problem, the concept of the sufficient spanning set [1] is used. Let the SVD of A be $A = W \Psi W^T$, here W and Ψ are the eigenvectors and eigenvalues, respectively. The sufficient spanning set of S_t' can be calculated by $\Phi_t = h([U, W])$ with h an orthonormalization function. Then U' is written as $U' = \Phi_t R_t$, and R_t is a rotation matrix. Thus, we solve a smaller eigen-problem to obtain R_t and Δ'

$$S_t' = U' \Delta' U'^T = \Phi_t R_t \Delta' R_t^T \Phi_t^T$$
$$\Rightarrow \Phi_t^T (S_t + A) \Phi_t = \Phi_t^T (U \Delta U^T + A) \Phi_t = R_t \Delta' R_t^T. \quad (10)$$

Suppose that d_t and d_A are the number of eigenvectors of U and W, respectively, the matrix $\Phi_t^T S_t' \Phi_t$ has the reduced size $d_t' = d_t + d_A$, and the eigen-analysis of S_t' requires a computational cost $O((d_t + d_A)^3)$. Let m be the number of existing training sets, the eigen-analysis of A requires a computational cost $O(m^3)$, and the total cost of our method is $O((d_t + d_A)^3 + m^3)$. While the eigen-analysis of S_t' in batch mode requires $O(m^6)$. Typically, d_t and d_A are (much) less than m^2 and m, respectively.

4.2. Updating the between-class canonical correlations

The between-class canonical correlations of existing data sets are represented by $\{V, \Sigma, P_i, C_i, i = 1, 2, \ldots, m\}$, here V and Σ are the eigenvectors and eigenvalues of S_b, s.t. $S_b = V \Sigma V^T$. P_i and C_i, respectively, represent the normalized orthonormal component matrix and class label of the ith set. Given a new image set represented by a normalized orthonormal basis matrix P_{m+1} and the corresponding class label C_{m+1}, the update is described as

$$\xi_2(V, \Sigma, P_i, C_i, P_{m+1}, C_{m+1}) = (V', \Sigma') \quad i = 1, 2, \ldots, m. \quad (11)$$

The updated S_b is computed by $S_b' = V \Sigma V^T + F$. F is calculated by

$$F = 2 \sum_{i \in E} (P_{m+1} Q_{m+1,i} - P_i Q_{i,m+1})(P_{m+1} Q_{m+1,i} - P_i Q_{i,m+1})^T, \quad (12)$$

where $E = \{j | C_j \neq C_{m+1}\}$. Let Z be the eigenvectors of F obtained by SVD solution, the sufficient spanning set of S_b' can be given by $\Phi_b = h([V, Z])$. $V' = \Phi_b R_b$, where R_b is a rotation matrix. Accordingly, the new small dimensional eigen-problem is given by

$$S_b' = V' \Sigma' V'^T$$
$$\Rightarrow \Phi_b^T S_b' \Phi_b = \Phi_b^T (V \Sigma V^T + F) \Phi_b = R_b \Sigma' R_b^T. \quad (13)$$

Let n_k be the number of sets belonging to class k, then the eigen-analysis of F requires a computational cost $O((m - n_{Cm+1})^3)$. Suppose d_b and d_F are the number of eigenvectors of V and F, respectively, the matrix $\Phi_b^T S_b' \Phi_b$ has the reduced size $d_b' = d_b + d_F$. The eigen-analysis of S_b' takes at most $O((d_b + d_F)^3)$ computation, whereas the eigen-analysis of the new between-class canonical correlations in batch mode takes $O((m^2 - \sum_k n_k^2)^3)$ computation, where $m = \sum_k n_k$. Typically, d_b and d_F are (much) less than $(m^2 - \sum_k n_k^2)$ and $(m - n_{Cm+1})$, respectively.

4.3. Updating the discriminant transformation matrix

The discriminant transformation matrix is computed using the updated total canonical correlations and between-class canonical correlations

$$\xi_3(U', \Delta', V', \Sigma') = T'. \quad (14)$$

In order to further reduce the computation complexity, we introduce new sufficient spanning set to change eigen-analysis into a smaller dimensional eigenvalue problem with the cost of $O(d_b'^3)$ rather than $O(d_t'^3)$. Let $G = U' \Delta'^{-1/2}$, then $G^T S_t' G = I$. As the denominator of the second criterion in Eq. (6) is the identity

matrix, the problem becomes to find the discriminant components that maximize $G^T S_b' G$, s.t. $G^T S_b' G = H \Lambda H^T$. The final discriminant components are obtained by $T' = GH$. The sufficient spanning set of the projection data can be constructed by $\Omega = h([G^T V'])$, and the eigenvalue problem is

$$G^T S_b' G = \Omega R \Lambda R^T \Omega^T \Rightarrow \Omega^T G^T V' \Sigma' V'^T G \Omega = R \Lambda R^T. \quad (15)$$

The updated discriminant matrix is given by

$$T' = GH = G\Omega R. \quad (16)$$

Let d_b' be the number of eigenvectors V', the computational time of eigen-problem in Eq. (15) is $O(d_b'^3)$. The dimension d'_t of U' is usually larger than d'_b, so the computation efficiency of T' improves from $O(d_t'^3)$ to $O(d_b'^3)$. The IDCC algorithm is summarized in Algorithm 1.

Algorithm 1. Incremental discriminant-analysis canonical correlations (IDCC)

Input: The total and between-class canonical correlations eigen-models $\{U, \Delta, V, \Sigma, P_i, C_i, i=1, 2, \ldots, m\}$ of the existing data sets, and the normalized orthonormal basis matrix P_{m+1} of the new data set with its label C_{m+1}

Output: Updated discriminant matrix T'

1. Update the total canonical correlations.
   Compute A by Eq. 9.
   Do SVD: $A = W \Psi W^T$.
   Set Φ_t by $\Phi_t = h([U, W])$.
   Compute eigenvectors R_t of $\Phi_t^T (U \Delta U^T + A) \Phi_t$.
   $U' = \Phi_t R_t$.
2. Update the between-class canonical correlations.
   Compute F by Eq. 12.
   Do SVD: $F = Z \Pi Z^T$.
   Set Φ_b by $\Phi_b = h([V, Z])$.
   Compute eigenvectors R_b of $\Phi_b^T (V \Sigma V^T + F) \Phi_b$.
   $V' = \Phi_b R_b$.
3. Update the discriminant matrix.
   Compute $G = U' \Delta'^{-1/2}$ and $\Omega = h([G^T V'])$.
   Eigendecompose $\Omega^T G^T V' \Sigma' V'^T G \Omega$ for the eigenvectors R.
   $T' = G \Omega R$.

5. Combination with semi-supervised learning

This section deals with the update of discriminant matrix in IDCC when the class labels of new additional training sets are not given. We integrate IDCC with the semi-supervised learning method to simultaneously label the new training sets and update the discriminant matrix. The semi-supervised learning addresses the problem of machine learning from both labeled and unlabeled training data, and the graph-based methods have been widely for semi-supervised learning. In this paper, we adopt a graph-based semi-supervised learning method, called the linear neighborhood propagation (LNP), which assumes that the label of each data set can be linearly reconstructed from its neighbors' labels. The labels from LNP can be sufficiently smooth with respect to the intrinsic structure collectively revealed by both labeled and unlabeled sets. In the combination of IDCC and LNP, firstly project an input unlabeled set into the discriminative subspace via the discriminant transformation matrix and predict its label by LNP on the discriminative subspace, and then incrementally update the discriminant transformation matrix with the newly labeled set. The combination of IDCC and LNP is listed in Algorithm 2.

Algorithm 2. Combination of IDCC and LNP

Input: m labeled image sets represented by the normalized orthonormal subspace $\Gamma_M = \{P_1, P_2, \ldots, P_m\}$ with the corresponding labels $Y_M = \{y_1, y_2, \ldots, y_m\}$; n unlabeled image sets $\Gamma_N = \{P_{m+1}, P_{m+2}, \ldots, P_{m+n}\}$; discriminant transformation matrix T; number of nearest neighbors k.

Output: updated discriminant matrix T'; labels y_j of unlabeled sets P_j, $j = m+1, m+2, \ldots, m+n$.

For $j = m+1$ to $m+n$ **do**

1. $\forall P_l \in \Gamma_M \cup \{P_j\}$, find the k-nearest neighborhood $N(P_l) = \{P_s | F_{ls} > \varepsilon, s = 1, 2, \ldots, k\}$ using the canonical correlations on discriminative subspace, i.e. $F_{ls} = Tr\{T^T P_s Q_{sl} Q_{ls}^T P_l^T T\}$, where $(T^T P_l)^T (T^T P_s) = Q_{ls} \Lambda Q_{sl}^T$.
2. Estimate the weights w_{ls} that best reconstruct P_l from its neighbors by minimizing $\| T^T P_l - \sum_{s=1}^{k} w_{ls} T^T P_s \|^2$ with the constraint $\sum_{s=1}^{k} w_{ls} = 1$.
3. Predict the label y_j of P_j by solving the quadratic optimization $y_j = \mathrm{argmin}_{f_j} \sum_{l=1}^{j} \| f_l - \sum_{P_s \in N(P_l)} w_{ls} f_s \|^2$ with the constraint $f_i = y_i$, $i = 1, 2, \ldots, j-1$.
4. Update the matrix T by Algorithm 1 with the newly labeled set P_j and its label y_j. The labeled sets are updated by $\Gamma_M = \Gamma_M \cup \{P_j\}$ and $Y_M = Y_M \cup \{y_j\}$.

End

6. Experiments

The experiments are conducted to evaluate our method on two publicly available data sets: KTH human dataset and Weizmann human dataset. For all experiments, the non-optimized Matlab codes run on a Dell PC with Intel Pentium D 3.4 GHz CPU and 1 G RAM.

6.1. Weizmann action recognition

In Weizmann action dataset [21], there are about 90 low-resolution (180×144, 25 fps) video sequences showing nine different subjects, each performing 10 actions including bending (bend), jumping jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), skipping (skip), galloping-sideways (side), walking (walk), waving-one-hand (wave1), and waving-two-hands (wave2). The centered silhouettes extracted in Ref. [21] are normalized to the same 64×48 dimension and converted into 3072 dimensional vectors in a raster-scan manner. Fig. 1 shows some normalized silhouettes of the ten actions mentioned above. The classification accuracy is evaluated under ninefold cross validation. Each time we take the silhouette frames of eight subjects for training and use those of the remaining one subject for testing. The training dataset is further partitioned into an initial set which is used for learning the initial discriminative model and the remaining sets which are added successively for re-training. In the semi-supervised learning method, the initial sets are labeled and the additional sets are unlabeled.

Our method (IDCC) has been compared with DCC [20], ILDA [16], and IPCA [15] in efficiency and accuracy. Particularly, we are interested in evaluating the discriminability and execution time of IDCC with the increasing data sets. In IDCC and DCC, the best dimension of the linear subspace of each image set is around 19 to represent 99 percent information, and the nearest neighbor (NN) classification is utilized based on the similarity between subspaces. PCA is performed to learn the linear subspace of each set in IDCC and DCC. For ILDA and IPCA, the dimensions of eigenspace

are set to 8 and 28, respectively, and the k-nearest neighbor is used for classification. Fig. 2 demonstrates the recognition accuracy of IDCC and the related methods with the increasing training data. The initial sets contain about 25% of the total training data, randomly constructed by two subjects with 10 actions. The additional sets are about 75% of the total training data, constructed by one subject with 10 actions at each incremental learning stage. IDCC achieves approximate accuracy as DCC, provided that enough components of the total and between-class canonical correlations are stored. k-NN-ILDA and k-NN-IPCA ($k=10$) perform worse since they are based on single image matching without exploiting the multiple image sets. The comparison of computational costs between DCC and IDCC is illustrated in Fig. 3. Whereas the execution time of DCC increases significantly with the training samples arriving successively, the time of the IDCC remains low. Table 2 shows the mean and standard deviation of recognition accuracy between different methods. Both IDCC and DCC provide significant improvements on recognition accuracy over IPCA and ILDA. Table 3 is the recognition rates of IDCC as well as some art-of-state methods, and all these methods adopt the evaluation scheme of leaving one out cross validation. As shown in Table 3, our method is significantly superior to the previous methods. We evaluate the recognition results of semi-supervised IDCC in Fig. 4. Since the prediction error exists in labeling the unlabeled training sets, the accuracy of semi-supervised IDCC is lower than that of the supervised IDCC.

6.2. KTH action recognition

The KTH human action dataset [10] contains six types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping. These actions are performed several times by twenty-five subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors with lighting variation (s4). Somebody tracking methods (e.g. [24]) can be applied to locate the areas and the geometric centers of human bodies in each frame, and the centered body region is normalized to the size of 50×50 pixels. Since the scenario s2 is only the scale variation of s1, the normalized human images of s2 are very similar to that of s1, and we conduct the experiment on s1, s3, and s4. Fig. 5 illustrates some samples from the KTH dataset. Leave-one-out cross-validation is performed, i.e. for each run the image sets of 24 subjects are used for training and the image sets of the remaining subject are for testing. Fig. 6 is the recognition rates of incremental solution between IDCC, DCC, IPCA, and ILDA. The initial sets contain about 25% of the total training data, randomly constructed by six subjects with 25 actions. The additional sets

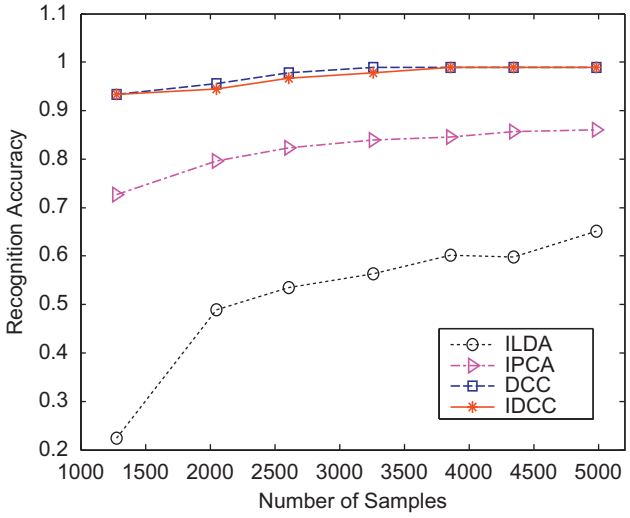

Fig. 2. Recognition accuracy of incremental solution.

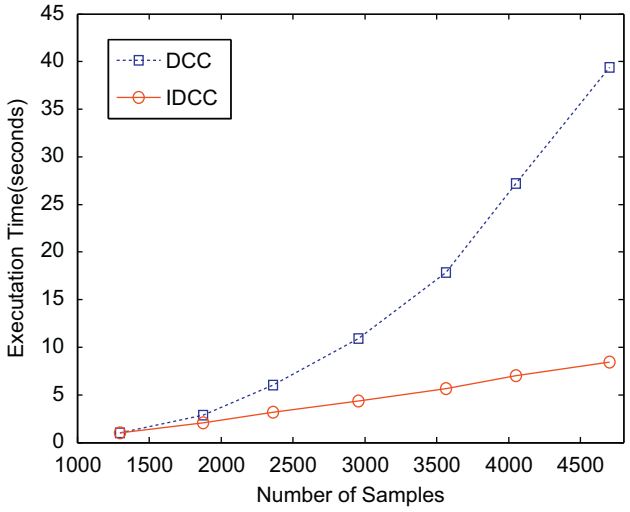

Fig. 3. Computation efficiency of IDCC and DCC.

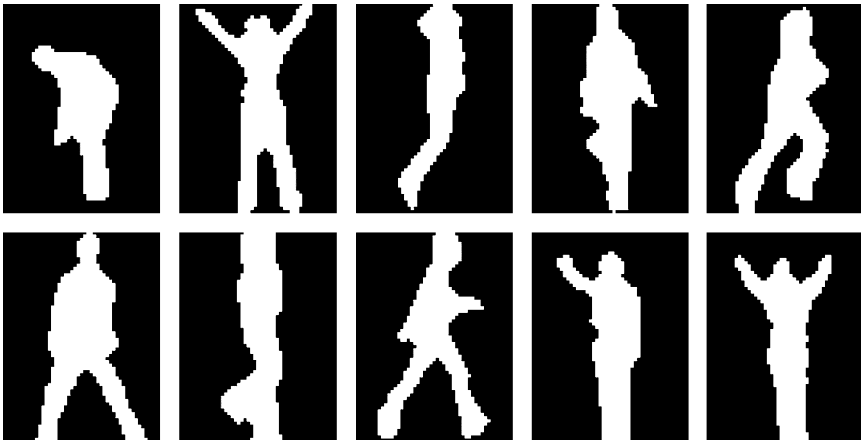

Fig. 1. Examples of the normalized silhouettes from Weizmann dataset.

Table 2
Mean and standard deviation of recognition accuracy for different methods. The number in parentheses represents the dimensionality of linear subspace.

Methods	Weizmann accuracy (%)
k-NN-IPCA($k=10$)	$86.09 \pm 0.03(28)$
k-NN-ILDA($k=10$)	$65.15 \pm 0.08(8)$
DCC	$98.89 \pm 0.02(19)$
IDCC	$98.89 \pm 0.02(19)$

Table 3
Recognition accuracy of some related recognition approaches. All these approaches use the evaluation scheme of leaving one out cross validation.

Methods	Weizmann Accuracy (%)
Our method	98.9
Wang and Suter [4]	97.8
Bregonzio et al. [22]	96.6
Zhang et al. [13]	92.9
Ali et al. [23]	92.6
Jia and Yeung [5]	90.9
Scovanner et al. [14]	84.2
Niebles and Li [8]	72.8

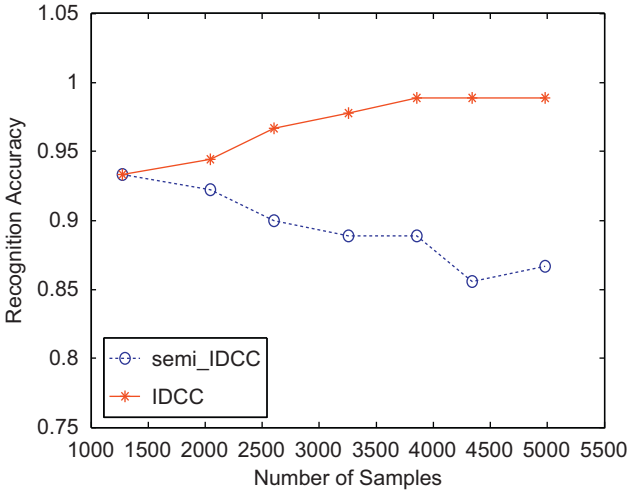

Fig. 4. Recognition accuracy of semi-IDCC and IDCC.

Fig. 5. Some sample frames from the KTH dataset.

are about 75% of the total training data, constructed by one subject with 25 actions at each incremental learning stage. Fig. 7 compares the computational costs between DCC and IDCC. Table 4 shows the mean and standard deviation of recognition accuracy for different methods on s1, s3, and s4 data sets of KTH, and

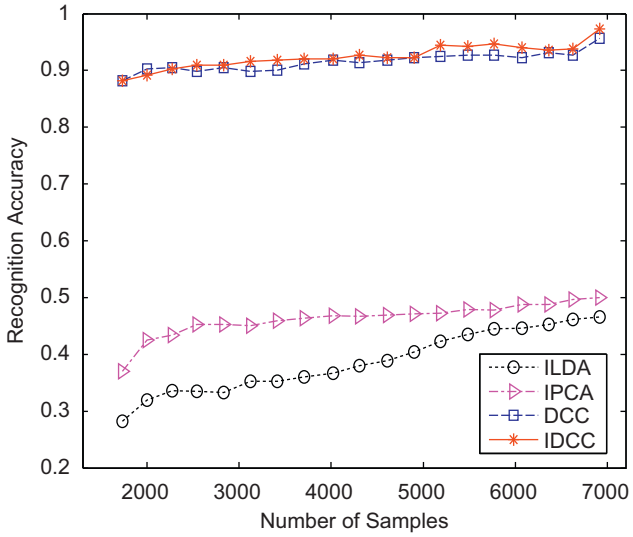

Fig. 6. Recognition rates of incremental solutions.

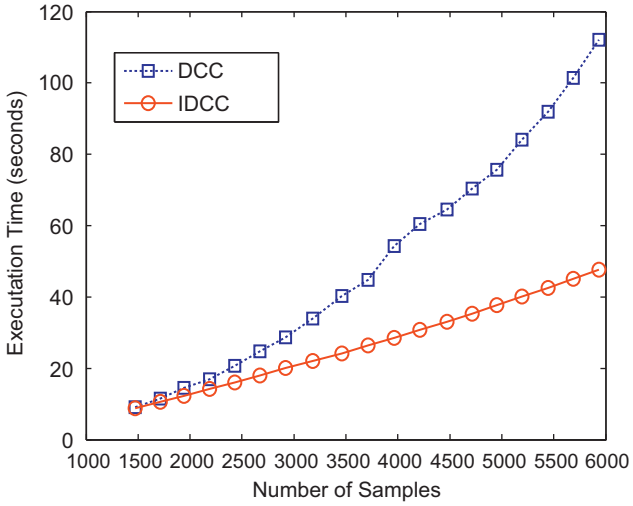

Fig. 7. Computation efficiency of IDCC and DCC on the KTH dataset.

Table 4
Comparisons of recognition rates. s1, s3, and s4 correspond to different conditions of the KTH database and avg. to the mean performance across the sets.

	KTH s1	KTH s3	KTH s4	Avg.
k-NN-IPCA($k=10$)	51.46 ± 0.16	49.82 ± 0.14	50.69 ± 0.15	50.66 ± 0.15
k-NN-ILDA($k=10$)	48.07 ± 0.13	39.77 ± 0.13	53.62 ± 0.08	47.15 ± 0.11
DCC	94.67 ± 0.05	89.33 ± 0.10	97.67 ± 0.02	93.89 ± 0.06
IDCC	96.00 ± 0.06	90.67 ± 0.10	98.67 ± 0.02	95.11 ± 0.06

Table 5 presents the recognition rates between other related recognition methods. Moreover, we compare the accuracy between the semi-supervised IDCC with partially labeled training sets and the supervised IDCC with all labeled sets as shown in Fig. 8.

6.3. Robustness test

To evaluate the adaptability and robustness of IDCC to the irregular actions in changing scenarios, we conduct the

experiment on 10 video sequences of people walking in various difficult scenarios [21], including walking with a dog, walking when swinging a bag, walking in a skirt, walking with partially occluded legs, walking occluded by pole, sleepwalking, limping, walking with knees up, walking when carrying a briefcase, and normal walking. Some selected example frames and their associated segmented silhouettes are illustrated in Fig. 9.

In this experiment, the testing action is recognized on a frame-by-frame basis, and the recognition accuracy is measured in terms of the percentage of the correctly recognized frames among the whole sequence. At each frame, we collect its local temporal neighbors as test image set and acquire the class label by computing the canonical correlations between the test set and those training sets. With the time process, several recognized frames are accumulated to construct the new training set for online re-training the discriminative model. For the trade-off between computational efficiency and effectiveness, we update the discriminative model at each interval of several frames rather than each frame. Table 6 shows the comparison results of LSTDE [5], DCC, and IDCC, from which we can see that IDCC indeed perform better on the irregular action recognition from ever-changing silhouettes via online updating the discriminative model.

Table 5
Recognition accuracy on KTH dataset comparison between related recognition approaches. All these approaches use the evaluation scheme of leaving one out cross validation.

Methods	KTH accuracy (%)
Our method	95.1
Bregonzio et al. [22]	93.2
Zhang et al. [13]	91.3
Savarese et al. [9]	86.8
Wang et al. [25]	85.0
Niebles et al. [11]	81.5
Dollar et al. [12]	81.7

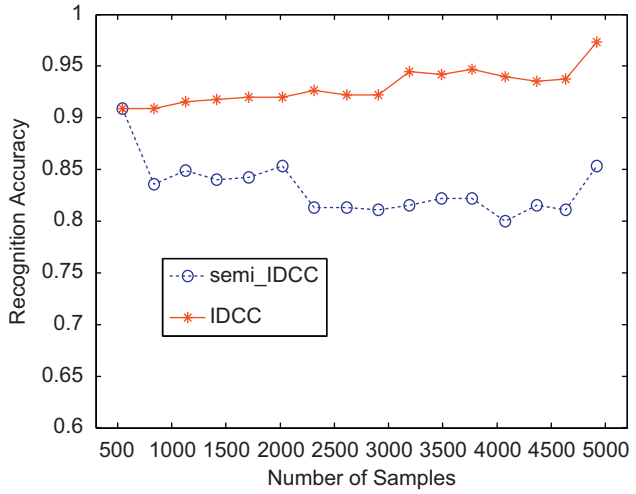

Fig. 8. Recognition accuracy of semi-IDCC and IDCC.

Table 6
Comparison of robustness test results between LSTDE, DCC, and IDCC.

Test sequence	Recognition accuracy (%)		
	LSTDE	DCC	IDCC
Walk with a dog	90.74	100	100
Swinging a bag	74.58	100	100
Walk in a skirt	92.16	80.49	85.37
Occluded feet	71.19	77.55	93.88
Occluded by pole	-	84.62	92.31
Moonwalk	56.06	78.57	94.64
Limp walk	83.96	87.50	90.63
Walk with knees up	66.02	64.52	72.10
Carry a briefcase	92.86	100	100
Normal walk	95.16	100	100

Fig. 9. Example frames and segmented silhouettes for robustness test.

7. Conclusions

We have presented a novel incremental discriminant-analysis canonical correlation (IDCC) method and its application to the online human action recognition in various changing scenarios. By efficiently updating the discriminative model, IDCC can capture the appearance variations of human and accurately recognize the action even in high irregular performance. Experiments on both regular and irregular actions have shown the superior discriminability in classification, significant adaptability to changing environments, and high computational efficiency of learning. IDCC can also be incorporated into semi-supervised learning framework and applied to the situation where the training data are partially labeled. In the future, we intend to extend the method to nonlinear learning and add the temporal information for more accurate recognition

Acknowledgments

This work was partially supported by the Natural Science Foundation of China (90920009, 60905006) the 973 Program of China (2006CB303103) and the Chinese High-Tech Program (2009AA01Z323).

References

[1] P. Hall, D. Marshall, R. Martin, Merging and splitting eigenspace models, IEEE Transactions on Pattern Analysis and Machine Learning 22 (9) (2000) 1042–1049.
[2] F. Wang, C.S. Zhang, Label propagation through linear neighborhoods, IEEE Transactions on Knowledge and Data Engineering. 20 (1) (2008) 55–67.
[3] X.X. Wu, W. Liang, Y.D. Jia, Incremental discriminative-analysis of canonical correlations for action recognition, IEEE 12th International Conference on Computer Vision (2009) 2035–2041.
[4] L. Wang, D. Suter, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model, IEEE Conference on Computer Vision and Pattern Recognition (2007).
[5] K. Jia, D.Y. Yeung, Human action recognition using local spatio-temporal discriminant embedding, IEEE Conference on Computer Vision and Pattern Recognition (2008).
[6] M.D. Rodriguez, J. Ahmed, M. Shah, Action MACH: a spatio-temporal maximum average correlation height filter for action recognition, IEEE Conference on Computer Vision and Pattern Recognition (2008).
[7] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, IEEE 11th International Conference on Computer Vision (2007).
[8] J.C. Niebles, F.F. Li, A hierarchical model of shape and appearance for human action classification, IEEE Conference on Computer Vision and Pattern Recognition (2007).
[9] S. Savarese, A. DelPozo, J.C. Niebles, F.F. Li, Spatial-temporal correlatons for unsupervised action classification, IEEE Workshop on Motion and Video Computing, 2008.
[10] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, 17th International Conference on Pattern Recognition 3 (2004) 32–36.
[11] J.C. Niebles, H.C. Wang, F.F. Li, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 70 (3) (2008) 299–318.
[12] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[13] Z.M. Zhang, Y.Q. Hu, S. Chan, L.T. Chia, Motion context: a new representation for human action recognition, in: Proceedings of the Tenth European Conference on Computer Vision, 2008, pp. 817–829.
[14] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia, 2007, pp. 357–360.
[15] P. M. Hall, D. Marshall, R. Martin, Incremental eigenanalysis for classification, in: Proceedings of the British Machine Vision Conference, 1998.
[16] S.N. Pang, S. Ozawa, N. Kasabov, Incremental linear discriminant analysis for classification of data streams, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 35 (5) (2005) 905–914.
[17] T.K. Kim, S.F. Wong, B. Stenger, J. Kittler, R. Cipolla, Incremental linear discriminant analysis using sufficient spanning set approximations, IEEE Conference on Computer Vision and Pattern Recognition (2007).
[18] R.S. Lin, D. Ross, J. Lim, M.H. Yang, Adaptive discriminative generative model and its applications, Advances in Neural Information Processing Systems (2004) 801–808.
[19] T.K. Kim, J. Kittler, R. Cipolla, Incremental learning of locally orthogonal subspaces for set-based object recognition, British Machine Vision Conference (2006) 559–568.
[20] T.K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1005–1018.
[21] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE 10th International Conference on Computer Vision 2 (2005) 1395–1402.
[22] M. Bregonzio, S.G. Gong, T. Xiang, Recognizing action as clouds of space-time interest points, IEEE Conference on Computer Vision and Pattern Recognition (2009) 1948–1955.
[23] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, IEEE 11th International Conference on Computer Vision (2007).
[24] A. Bissacco, M.H. Yang, S. Soatto, Fast human pose estimation using appearance and motion via multi-dimensional boosting regression, IEEE Conference on Computer Vision and Pattern Recognition (2007).
[25] Y. Wang, P. Sabzmeydani, G. Mori, Semi-latent dirichlet allocation: a hierarchical model for human action recognition, in: IEEE International Conference on Computer Vision Workshop on Human Motion Understanding, Modeling, Capture and Animation, 2007.

Xinxiao Wu received the B.A. degree in computer science from the Nanjing University of Information Science and Technology. She is currently a doctorial candidate in the School of Computer Science at the Beijing Institute of Technology. Her thesis is focused on the problem of human action recognition and human pose estimation. Her research interests include machine learning, computer vision, and human action perception.

Yunde Jia received the Ph.D. degree in mechatronics from the Beijing Institute of Technology in 2000. He is currently a professor of computer science and also a director of the Lab of Media Computing and Intelligent Systems, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision, media computing, human computer interaction and intelligent systems.

Wei Liang received the B.A. degree in computer science from the Shandong Institute of Light Industry in 2000 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2005.She is currently a lecturer of in the School of Computer Science at the Beijing Institute of Technology. Her research interests include computer vision, human tracking and action perception.