

View-invariant human action recognition via robust locally adaptive multi-view learning^{*}

Jia-geng FENG[†], Jun XIAO

(Institute of Artificial Intelligence, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

[†]E-mail: fengjiageng@126.com

Received Mar. 18, 2015; Revision accepted Sept. 14, 2015; Crosschecked Oct. 12, 2015

Abstract: Human action recognition is currently one of the most active research areas in computer vision. It has been widely used in many applications, such as intelligent surveillance, perceptual interface, and content-based video retrieval. However, some extrinsic factors are barriers for the development of action recognition; e.g., human actions may be observed from arbitrary camera viewpoints in realistic scene. Thus, view-invariant analysis becomes important for action recognition algorithms, and a number of researchers have paid much attention to this issue. In this paper, we present a multi-view learning approach to recognize human actions from different views. As most existing multi-view learning algorithms often suffer from the problem of lacking data adaptiveness in the nearest neighborhood graph construction procedure, a robust locally adaptive multi-view learning algorithm based on learning multiple local L1-graphs is proposed. Moreover, an efficient iterative optimization method is proposed to solve the proposed objective function. Experiments on three public view-invariant action recognition datasets, i.e., ViHASi, IXMAS, and WVU, demonstrate data adaptiveness, effectiveness, and efficiency of our algorithm. More importantly, when the feature dimension is correctly selected (i.e., >60), the proposed algorithm stably outperforms state-of-the-art counterparts and obtains about 6% improvement in recognition accuracy on the three datasets.

Key words: View-invariant, Action recognition, Multi-view learning, L1-norm, Local learning

doi:10.1631/FITEE.1500080

Document code: A

CLC number: TP391

1 Introduction

In recent years, human action recognition is one of the most popular and important topics in computer vision. It has been widely used in many applications, such as human-computer interaction, medical treatment, sports, and video surveillance, and has tremendous economic importance. However, certain extrinsic factors, such as human body shape, cloth change, gesture, camera state, light, and background, are barriers to its development. Most of the existing human action recognition research works have to

focus on objective constraints on environmental conditions, the action type, and the state of camera to make it solvable. Additionally, they require that the performed human actions are parallel to the camera plane (Bobick and Davis, 2001; Yu *et al.*, 2005; Chen *et al.*, 2006), which sometimes is not possible in real-world applications. Thus, most of them cannot be applied to realistic environments. To overcome these problems, some researchers have started to bring independent property into their recognition methods. View invariance is one of the hot topics that is important in video surveillance. It has already been used in object recognition (Srestasathien and Yilmaz, 2008; Raytchev *et al.*, 2010), face recognition (Ashraf *et al.*, 2008; Tian *et al.*, 2008), and gait recognition (Jean *et al.*, 2008). View-invariant human action recognition is trying to overcome the influence of different view angles of the input video and make it possible to recognize human action from any given

^{*} Project supported by the National Natural Science Foundation of China (No. 61572431), the National Key Technology R&D Program (No. 2013BAH59F00), the Zhejiang Provincial Natural Science Foundation of China (No. LY13F020001), and the Zhejiang Province Public Technology Applied Research Projects, China (No. 2014C33090)

 ORCID: Jia-geng FENG, <http://orcid.org/0000-0003-4577-4520>

© Zhejiang University and Springer-Verlag Berlin Heidelberg 2015

view. In the perspective of data representation, view invariance refers to the requirement that feature representations of the same human action from different viewpoints are the same or similar to each other. Indeed, it is a data representation problem that needs to be solved for view-invariant human action recognition. In light of this, a lot of researchers have proposed various efficient methods for solving the view-invariant problem for human action recognition.

As the low-level visual features usually cannot fully describe the high-level semantics of objects in computer vision and pattern recognition, multiple visual features are extracted and adopted for feature representation (Fu and Xian, 2001; Zheng and Ye, 2006; Tang *et al.*, 2011). Different from the traditional single features, these multiple features are from different feature space and have special physical meanings, feature dimensions, and discriminative power. We call these multiple feature data as multi-view data (Xia *et al.*, 2010). Many existing multi-view learning algorithms are based on the spectral analysis theory, which usually needs to build a nearest neighborhood graph to represent the geometric structure of data (de Sa Virginia, 2005; Zhou and Burges, 2007; Xia *et al.*, 2010). Thus, the performances of these algorithms rely on the reasonableness of the graphs. There are two widely used graph construction methods: the k -nearest neighborhood graph and the ε -ball graph. These methods are not data adaptive and sensitive to noise and outliers. Moreover, the former needs the user to set the nearest neighborhood number in advance and the latter requires the kernel parameter. However, these two parameters are often sensitive to the input data, and there is no natural way to automatically tune them (Wen *et al.*, 2011). As a result, it is time-consuming to obtain the best parameters in these two methods. In addition, in some previous works, as the Euclidean distance measure was used to find the neighborhoods, the effect of noise was not taken into account which may lead to poor performance. Therefore, how to improve data adaptability and the robustness of graph construction method is an urgent problem.

To resolve the aforementioned issues, we propose a robust locally adaptive multi-view learning (RAML) algorithm, which is based on sparse representation theory and has the advantages of the L1-graph with respect to the noise and outliers. In contrast to other

existing L1-graph-related work, the L1-graph used in our work is local and robust. As the global L1-graph suffers from the problem of high computational cost, we adopt multiple local L1-graphs instead of the traditional global L1-graph during graph construction. As the approximated k -nearest neighborhood search is very efficient, our graph learning method is also very fast and efficient.

2 Related work

View-invariant human action recognition has a large number of real applications, such as intelligent video surveillance which aims to detect, track, and understand human behaviors from image sequences, video research, and computer games (Luo *et al.*, 2003; Brémond *et al.*, 2006). The existing view-invariant human action recognition methods can be broadly divided into the following four categories (Feng and Xiao, 2013): (1) spatial-temporal feature based methods, (2) probabilistic state-space based methods, (3) dimension reduction methods, and (4) motion trajectory based methods.

The spatial-temporal feature based methods accumulate the observed values from motion sequences with respect to the time line to form spatiotemporal volumes and then extract visual features from these volumes, followed by human action recognition and classification using some classifiers with these features. The first work that used the spatiotemporal volume was proposed by Syeda-Mahmood *et al.* (2001). They defined an action cylinder. Yilmaz and Shah (2005) proposed spatiotemporal volumes, which were derived from continuous contours. Weinland *et al.* (2006) made a visual hull based on multi-view information and then obtained motion history volumes by accumulating the time line. Roh *et al.* (2010) introduced the definition of volume motion template and projected the motion template. Additionally, Yan *et al.* (2008) proposed an innovative four-dimensional motion feature model.

The probabilistic state-space based methods treat each kind of human gestures as a state, calculate the probability of the conversion between states, and use the probability as the edge of the state graph. Therefore, any sequence of actions can be regarded as a chain of these states. Then human action recognition

is achieved by obtaining the most probabilistic chain that matches the input sequence. Specially, there are three kinds of methods that have been used in view-invariant human action recognition: Bayesian network (Natarajan *et al.*, 2010), hidden Markov model (Wang *et al.*, 2007; Weinland *et al.*, 2007), and conditional random fields (Natarajan and Nevatia, 2008).

The dimension reduction based methods are the most widely used ones. They are efficient and robust in many cases. The main idea of these methods is to extract low-level features from the image sequences and then apply dimension-reduction algorithms to reduce feature dimension and recognize human action with special classifiers using the reduced features as the input. These methods can also be divided into two groups: linear methods, such as principal component analysis (PCA) (Wold *et al.*, 1987), latent Dirichlet allocation (LDA) (Balakrishnama and Ganapathiraju, 1998), and non-negative matrix factorization (Lee and Seung, 1999), and nonlinear methods, such as temporal lobe epilepsy (Lewandowski *et al.*, 2010) and ISOMap (Balasubramanian and Schwartz, 2002).

The motion trajectory based methods focus on the trajectories of human joints or interesting points and extract visual features from them. Rao *et al.* (2002) presented a motion trajectory based method by computing the degree of trajectory curvature and the magnitude of direction change. Parameswaran and Chellappa (2006) calculated the geometric invariances of five key points on the same plane. Junejo *et al.* (2008) used the self-similarity of human motion trajectory and exploited the self-similarity matrix.

In summary, the spatiotemporal feature based methods require much computational cost, and the probabilistic state-space based methods need much time in training the model. The main difficulty in the motion trajectory based methods is how to fast and correctly detect the motion trajectory. This process usually relies on a robust and efficient tracking algorithm. The dimension reduction based methods are independent of the input video sequences and can be widely used with various visual features. Therefore, the dimension reduction based methods have attracted much attention in computer vision. In this study, our proposed approach belongs to this kind of methods.

Multi-view learning was first proposed by Blum and Mitchell (1998) in their co-training algorithm. Under the assumption that different views are mutu-

ally independent, if there is a reasonably accurate classification result of one view, it can be used in other views to improve the recognition accuracy. However, if there is no mutual independence between different views, the algorithm's result will be influenced.

In the application of multi-view learning, researchers have used feature information from different views to obtain better performance than just using them from a single view. Spectral embedding learning has been widely used within many multi-view learning algorithms. Previous researchers liked to learn models from multiple views separately. Thus, the relevant information among different feature views are not taken into consideration. Based on the work of Long *et al.* (2008), Xia *et al.* (2010) proposed a multi-view spectrum embedding (MSE) algorithm. Although the MSE algorithm has the advantage of relevant information among different feature views, the user has to set the number of neighbors in advance to build the nearest neighborhood graph. Hence, this method has the disadvantage of lacking adaptability. Furthermore, some researchers overlooked the noise problem that may arise in many real-world applications. Hence, their proposed methods are not robust with the real-world noise.

The fact that compared with the L2-norm, L1-norm (Donoho, 2006) is sparse has a strong physical meaning. In addition, L1-norm has the property of good data adaptability and is robust to noise. Therefore, we will propose the RAML algorithm in this study. The RAML algorithm takes not only the local geometric structure information of data but also the useful complementary information between different views into account. Different from many existing L1-graph based methods, the RAML algorithm uses multiple local L1-graphs to speed up the graph construction phase. In the next section, we will introduce the details of our proposed RAML algorithm and apply it to deal with the view-invariant human action recognition problem.

3 Visual feature representation

In this study, we use three public view-invariant human action datasets, namely, ViHASi (<http://dipersec.king.ac.uk/VIHASI/>), IXMAS (<http://4drepository.inrialpes.fr/public/viewgroup/6>), and

WVU (<http://csee.wvu.edu/~vkkulathumani/wvu-action.html>).

ViHASi has synthesized silhouette images and motion capture data, including 31 viewpoints, 20 types of actions, and 9 actors. IXMAS is captured from 5 different viewpoints and includes 14 types of actions performed by 11 actors. It provides videos, silhouette images, and volume data. WVU consists of 12 types of actions, which are performed by 2 actors and captured from 8 viewpoints. More details about these three datasets are given in Section 5.

For these three datasets, we use the silhouette image sequences as the input and extract visual features from them. Specially, four kinds of silhouette features (Chen *et al.*, 2010), i.e., Fourier descriptor, geometric signature, grid descriptor, and distance transform descriptor, are extracted from each frame of the input sequences. Hence, we can obtain many frame based visual features, which can be used to represent actions.

As the frame number of different silhouette image sequences may not be the same, to obtain a fixed-length video based feature representation, we adopt the bag-of-words scheme (Yang *et al.*, 2007). In detail, we first cluster the frame based visual features from different viewpoints and actions into multiple bag-of-visual-words (BOVW) using the training data. Then each silhouette image sequence can be converted to a fixed-length BOVW feature via the sparse coding method. Note that we process each kind of silhouette features separately; therefore, each video can be represented by four kinds of BOVW features.

Therefore, each dataset can be represented as $X = \{X^i \in \mathbb{R}^{d_i \times N}\}_{i=1}^m$ using the above learned BOVW features, wherein X^i is the i th kind of features, d_i the feature dimension of the i th feature, N the total number of silhouette image sequences, and m the number of features.

4 Robust locally adaptive multi-view learning

In computer vision and machine learning, many researchers have found that high-dimensional visual data often lie on a low-dimensional special subspace. The relationship between different visual data approximates a linear function. Thus, for any given datum, it can be represented by a linear combination of the other data. Meanwhile, different visual features have special discriminative power in action recognition, while the complementary information between different views is helpful in feature representation.

To preserve the geometric structure information of data and exploit other useful information (Roweis and Saul, 2000; He *et al.*, 2005), we propose the RAML algorithm in this study. The flowchart of the RAML algorithm is shown in Fig. 1. The main data process of our proposed algorithm includes three steps: (1) extraction of silhouette image sequences from the input multi-view videos, (2) extraction of multiple visual features based on the silhouette image sequences, and (3) application of the RAML algorithm to learn a more compact and discriminative feature representation. Once such a new feature representation is learnt, we can use many existing

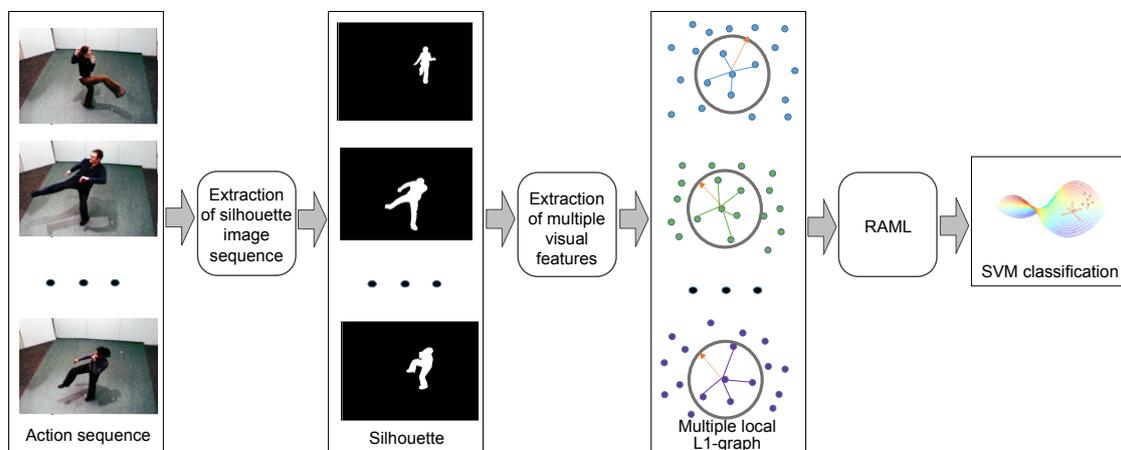


Fig. 1 Flowchart of our proposed RAML algorithm

classifiers, such as support vector machine (SVM) and decision tree, in the following action recognition phase. Next, we introduce our proposed RAML algorithm in detail.

To avoid confusion, we clarify that the word ‘view’ in view-invariance means different viewpoints of camera, and in a multi-view, single-view learning means the homogeneous features from objects having the same single viewpoint.

Our objective function is represented as follows (meanings for variables will be given later):

$$\begin{aligned} & \arg \min \sum_{i=1}^m \alpha_i^r \text{tr}(\mathbf{Y}\mathbf{L}_n^i\mathbf{Y}^T) \\ & \text{s.t. } \mathbf{Y}\mathbf{Y}^T = \mathbf{I}, \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0. \end{aligned}$$

We have obtained four heterogeneous features above (Fourier descriptor, geometric signature, grid descriptor, and distance transform descriptor). Furthermore, we show how to learn the homogeneous features using the sample data and these heterogeneous features.

4.1 Single-view robust local L1-graph

Assume we have N samples of human action data and the i th kind of visual features is clustered into d_i BOVWs. Therefore, the dataset is represented as $\mathbf{X}^i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_N^i] \in \mathbb{R}^{d_i \times N}$ in terms of the i th kind of visual features, and any datum \mathbf{x}_p^i ($1 \leq p \leq N$) can be represented by a linear combination of the other data in the dataset (Mao, 2013):

$$\min_{\mathbf{w}} \|\mathbf{w}_p\|_0 \quad \text{s.t. } \mathbf{x}_p^i = \mathbf{X}^i \mathbf{w}_p, \quad (1)$$

where $\|\cdot\|_0$ is the L0-norm and $\mathbf{w}_p \in \mathbb{R}^{N \times 1}$.

However, to solve the optimization problem, we have to consider all the possible combinations. The L0-norm in Eq. (1) is a non-convex and NP-hard problem. Donoho *et al.* (2006) proved that if \mathbf{X}^i satisfies the following condition:

$$\text{Spark}(\mathbf{X}^i) \geq 2 \|\mathbf{w}_p\|_0,$$

where $\text{Spark}(\mathbf{X}^i)$ represents the smallest number of the linear dependent column vectors, then the optimization problem has a unique solution.

Afterwards, Candès and Romberg (2005) proved that under the restricted isometry property (RIP) condition, the results of the L0-norm and L1-norm are very similar (Donoho, 2006). To make Eq. (1) solvable, we replace the L0-norm constraint with the L1-norm constraint. We tend to construct the L1-graph to model the geometric structure information under each single view. As we know, the L1-norm constraint is of high computational complexity if the problem is a large-scale problem. To overcome this problem, we modify the global L1-graph by using a robust local L1-graph. Specially, we learn the robust local L1-graph as follows:

1. As each datum relates only to a few locally similar data in the dataset, we find M nearest neighbors for each datum. To speed up the searching procedure, we adopt the kd-tree algorithm to preprocess the data and search neighbors. The computational complexity of constructing the kd-tree structure is nearly $O(N \log N)$, which just requires $O(\log N)$ for searching the nearest neighbors.

2. Using the nearest neighbors to construct a sub-dataset and solve a local L1-norm constrained problem, i.e., Eq. (2), we can obtain the local reconstruction coefficient for each datum:

$$\min_{\mathbf{w}} \|\mathbf{w}_p\|_1 \quad \text{s.t. } \mathbf{x}_p^i = \mathbf{X}_M^i \mathbf{w}_p, \quad (2)$$

where $\mathbf{X}_M^i = [\mathbf{x}_{i_1}^i, \mathbf{x}_{i_2}^i, \dots, \mathbf{x}_{i_M}^i] \in \mathbb{R}^{d_i \times M}$ ($i_1, i_2, \dots, i_M \leq N, M \leq N$) represents the M neighbors of \mathbf{x}_p^i under the i th feature representation.

In our experiment, we set M to 900. The performance variation of our proposed RAML algorithm with respect to M is discussed in Section 5.

Therefore, we reduce the high computational cost of the global L1-graph via learning local L1-graphs. Meanwhile, the advantages of L1-graph are included in our proposed RAML algorithm.

In objective function (2), as we use the L1-norm constraint, \mathbf{w}_p is a sparse vector, which is the reconstruction coefficient of datum \mathbf{x}_p^i . Once all the reconstruction coefficients of data are learnt, we can construct a local L1-graph according to them. The local L1-graph is defined as $G = \{\mathbf{X}^i, \mathbf{W}^i\}$, where the data sample set \mathbf{X}^i represents graph vertices and \mathbf{W}^i is the graph weight matrix. We set \mathbf{W}^i as follows:

$$W_{jk}^i = \begin{cases} w_{jk}^i, & j > k, \\ w_{j(k-1)}^i, & j < k, \\ 0, & j = k. \end{cases} \quad (3)$$

4.2 Graph based local optimization

After obtaining the graph weight matrix (W^i) for each view, we wish to learn a more compact and representative low-dimensional feature representation of the original high-dimensional multi-view features. During feature dimension reduction, we hope to exploit the geometric structure information of the original data. Recall that in the i th view, the feature matrix is $X^i = [x_1^i, x_2^i, \dots, x_N^i] \in \mathbb{R}^{d_i \times N}$ and the corresponding graph weight matrix is W^i .

For each data point x_j^i and its k nearest neighborhoods, we define a local region of datum x_j^i as $X_j^i = [x_j^i, x_{j_1}^i, \dots, x_{j_k}^i] \in \mathbb{R}^{d_i \times (k+1)}$. In a local patch, the relationship between the data points can be regarded as a linear function (Roweis and Saul, 2000; Zhou and Burges, 2007; Shen and Si, 2010). Thus, we adopt a local linear function to predict the low-dimensional feature of each datum in X_j^i , i.e., $f_j^i : X_j^i \rightarrow Y_j^i$, where $Y_j^i = [y_j^i, y_{j_1}^i, \dots, y_{j_k}^i] \in \mathbb{R}^{d \times (k+1)}$. Here, d is the dimension of the learned new feature representation. To hold the local geometric structure information embedded in the original feature space, we try to minimize the following local optimization objective function:

$$\arg \min_{Y_j^i} \sum_{l=1}^k \|y_j^i - y_{j_l}^i\|^2 (w_{j_l}^i), \quad (4)$$

where $w_{j_l}^i$ is a weight vector selected from the j th row of W^i . Eq. (4) can be rewritten as follows:

$$\arg \min_{Y_j^i} \text{tr} \begin{bmatrix} (y_j^i - y_{j_1}^i)^T \\ (y_j^i - y_{j_2}^i)^T \\ \dots \\ (y_j^i - y_{j_k}^i)^T \end{bmatrix}$$

$$\begin{aligned} & \cdot [y_j^i - y_{j_1}^i, y_j^i - y_{j_2}^i, \dots, y_j^i - y_{j_k}^i] \text{diag}(w_j^i) \\ & = \arg \min_{Y_j^i} \text{tr} \left(Y_j^i \begin{bmatrix} -e_k^T \\ I_k \end{bmatrix} \text{diag}(w_j^i) \cdot [-e_k, I_k] (Y_j^i)^T \right) \\ & = \arg \min_{Y_j^i} \text{tr}(Y_j^i L_j^i (Y_j^i)^T), \end{aligned} \quad (5)$$

where $e_k = [1, 1, \dots, 1]^T$, I_k a $k \times k$ identity matrix, $\text{tr}(\cdot)$ the trace operator, and $L_j^i \in \mathbb{R}^{(k+1) \times (k+1)}$ satisfying

$$L_j^i = \begin{bmatrix} -e_k^T \\ I_k \end{bmatrix} \text{diag}(w_j^i) [-e_k, I_k] = \begin{bmatrix} \sum_{l=1}^k (w_{j_l}^i)_l, & -(w_j^i)^T \\ -w_j^i, & \text{diag}(w_j^i) \end{bmatrix}. \quad (6)$$

Thus, the local optimization objective function is equivalent to the following equation:

$$\arg \min_{Y_j^i} \text{tr}(Y_j^i L_j^i (Y_j^i)^T). \quad (7)$$

To jointly take the multi-view data into account, a non-negative view-weight vector is used to measure the importance of different kinds of visual features. So, we summarize all the m kinds of features and obtain the following objective function:

$$\arg \min_{Y = \{Y_j^i\}_{i=1}^m, \alpha} \sum_{i=1}^m \alpha_i \text{tr}(Y_j^i L_j^i (Y_j^i)^T). \quad (8)$$

4.3 Global coordinates alignment

For each local area X_j^i , there exists a low-dimensional feature representation $Y_j^i = [y_j^i, y_{j_1}^i, \dots, y_{j_k}^i]$. For all the data, we denote the corresponding low-dimensional feature representation as $Y = [y_1, y_2, \dots, y_n]$. If we use a selection matrix $S_j^i \in \mathbb{R}^{n \times (k+1)}$, we obtain $Y_j^i = Y S_j^i$. The objective function described in Eq. (8) is equal to

$$\arg \min_{Y, \alpha} \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y} \mathbf{S}_j^i \mathbf{L}_j^i (\mathbf{S}_j^i)^T \mathbf{Y}^T). \quad (9)$$

Summarizing all the data points, we can obtain the following equation:

$$\begin{aligned} & \arg \min_{Y, \alpha} \sum_{j=1}^n \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y} \mathbf{S}_j^i \mathbf{L}_j^i (\mathbf{S}_j^i)^T \mathbf{Y}^T) \\ & = \arg \min_{Y, \alpha} \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y} \mathbf{L}^i \mathbf{Y}^T), \end{aligned} \quad (10)$$

where $\mathbf{L}^i = \sum_{j=1}^n \mathbf{S}_j^i \mathbf{L}_j^i (\mathbf{S}_j^i)^T$. As $\mathbf{L}^i = \mathbf{D}^i - \mathbf{W}^i$ (\mathbf{D}^i is a transformation matrix which is diagonal), we have

$$\mathbf{L}_n^i = (\mathbf{D}^i)^{-1/2} \mathbf{L}^i (\mathbf{D}^i)^{-1/2} = \mathbf{I} - (\mathbf{D}^i)^{-1/2} \mathbf{W}^i (\mathbf{D}^i)^{-1/2}.$$

Therefore, the abovementioned objective function (8) can be reformed as follows:

$$\begin{aligned} & \arg \min_{Y, \alpha} \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T) \\ & \text{s.t. } \mathbf{Y} \mathbf{Y}^T = \mathbf{I}, \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0, \end{aligned} \quad (11)$$

where $\alpha \in \mathbb{R}^{m \times 1}$ is a non-negative view-weight vector, which represents the weight for different kinds of visual features. α is set to be non-negative because we want to ensure that there is only positive relationship between different features, avoiding the negative additive effect. As \mathbf{Y} is orthogonal, the solution is unique. To prevent the trivial problem of Eq. (7), we use the r power of α_i instead of α_i (Cheng et al., 2010). It brings in two benefits: one is the elimination of the trivial solution of objective function (11) and the other is the improvement in the original simple linear weight, furthermore strengthening the impact of discriminative features on obtaining low-dimensional embedding. Then we have

$$\begin{aligned} & \arg \min_{Y, \alpha} \sum_{i=1}^m \alpha_i^r \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T) \\ & \text{s.t. } \mathbf{Y} \mathbf{Y}^T = \mathbf{I}, \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0. \end{aligned} \quad (12)$$

4.4 Optimization method

To solve Eq. (12), we propose an iterative optimization scheme. First, we use a Lagrange multiplier and convert Eq. (12) to the following form:

$$L(\alpha, \lambda) = \sum_{i=1}^m \alpha_i^r \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T) - \lambda \left(\sum_{i=1}^m \alpha_i - 1 \right). \quad (13)$$

Then we calculate the derivation of L with respect to α_i and λ , and then set them to zeros, respectively. We have the following equations:

$$\frac{\partial L(\alpha, \lambda)}{\partial \alpha_i} = r \alpha_i^{r-1} \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T) - \lambda = 0, \quad (14)$$

$$\frac{\partial L(\alpha, \lambda)}{\partial \lambda} = \sum_{i=1}^m \alpha_i - 1 = 0. \quad (15)$$

Based on Eqs. (13)–(15), we can obtain

$$\alpha_i = \frac{[1 / \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T)]^{1/(r-1)}}{\sum_{i=1}^m [1 / \text{tr}(\mathbf{Y} \mathbf{L}_n^i \mathbf{Y}^T)]^{1/(r-1)}}. \quad (16)$$

Now, objective function (12) degenerates to

$$\min_Y \text{tr}(\mathbf{Y} \mathbf{G} \mathbf{Y}^T) \text{ s.t. } \mathbf{Y} \mathbf{Y}^T = \mathbf{I}, \quad (17)$$

where $\mathbf{G} = \sum_{i=1}^m \alpha_i^r \mathbf{G}^i$, $\mathbf{G}^i = \mathbf{L}_n^i$. Eq. (17) has an optimal solution, which can be obtained via eigenvalue decomposition. The optimal value of \mathbf{Y} is the eigenvector of \mathbf{G} with respect to d smallest eigenvalues.

Now, we obtain the optimization method for solving our proposed RAML algorithm (Algorithm 1).

Algorithm 1 Robust locally adaptive multi-feature learning

Input: the input dataset $X = \{\mathbf{X}^i \in \mathbb{R}^{d_i \times N}\}_{i=1}^m$, visual feature dimension d , and power value r .

Output: $\mathbf{Y} \in \mathbb{R}^{d \times n}$, where $d \geq d$ ($1 \leq i \leq m$).

Optimization procedure:

1. Search M neighbors for each datum in X .
2. Construct local L1-graph by solving Eq. (2).
3. Calculate \mathbf{L} according to Eq. (6).
4. Calculate \mathbf{G}^i ($1 \leq i \leq m$) for each feature (initially $\alpha = [1/m, 1/m, \dots, 1/m]$).

5. Do the following iteration until reaches convergence:
- (1) $Y=U^T$, where $U=[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d]$ ($\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d$ are the corresponding eigenvectors of d smallest eigenvalues of matrix G).
 - (2) Calculate α_i (Eq. (16)).

5 Experiments

In this study, we apply the proposed RAML algorithm to the view-invariant human action recognition problem. We compare the RAML algorithm with other four widely used methods: principal component analysis (PCA), sparse representation for classification (SRC) (Wright *et al.*, 2009), latent Dirichlet allocation (LDA), and multi-view spectral embedding (MSE) (Xia *et al.*, 2010). We use three public view-invariant human action datasets in this experiment. The details about these datasets are listed in Table 1.

For each dataset, four kinds of silhouette image visual features are extracted following Chen *et al.*

(2010). These features are Fourier descriptor, geometric signature, grid descriptor, and distance transform descriptor. Some examples from these three datasets are shown in Fig. 2.

First of all, we randomly select 10% of data samples as the training dataset and the remainder the testing dataset. To make all the competitions fair, we tune the model-related parameters for each algorithm and use the optimal parameter settings, respectively. Once the new feature representation is learned, we adopt the SVM classifier to recognize the human action based on these new features. All the experiments repeat 30 times, and we report the average accuracy of each algorithm.

In the following experiment, we show performance variation of our proposed RAML algorithm with respect to M , which is the number of nearest neighbors for each datum. From Fig. 3, we find that when M is larger than 900, the performances are almost the same; hence, we set $M=900$ in our experiment, considering the efficiency of the algorithm.

As shown in Fig. 4, when the feature dimension is 120, our proposed algorithm can improve 7% recognition accuracy on WVU and 6% on ViHASi and IXMAS. It is interesting to see that multi-view learning algorithms such as MSE and RAML

Table 1 Information of experimental datasets

Dataset	Number of viewpoints	Number of actors	Number of actions
ViHASi	31	9	20
IXMAS	5	11	14
WVU	8	2	12

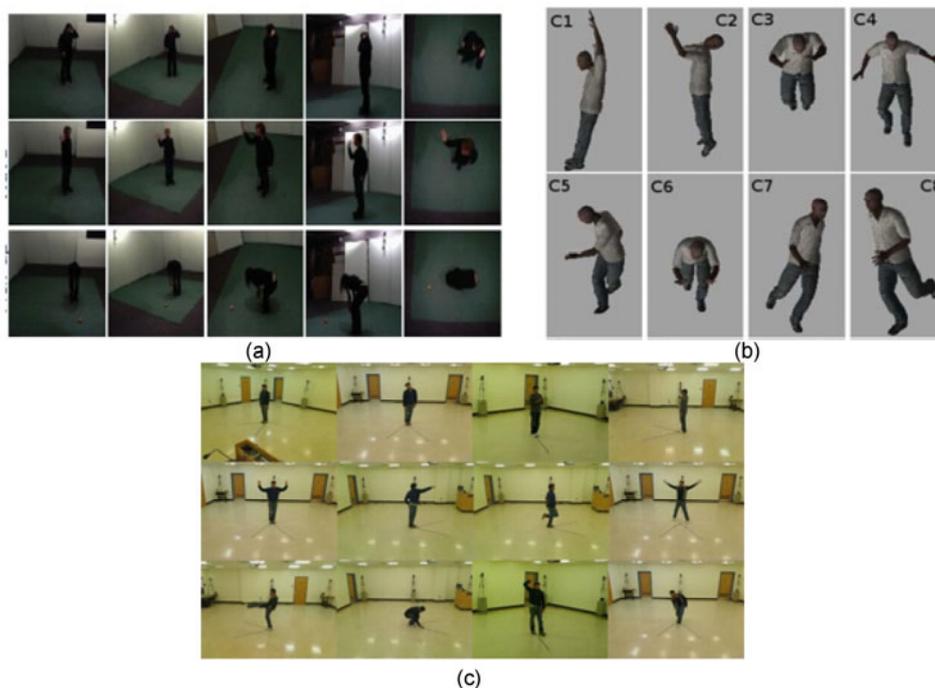


Fig. 2 Some sample images from IXMAS (a), ViHASi (b), and WVU (c)

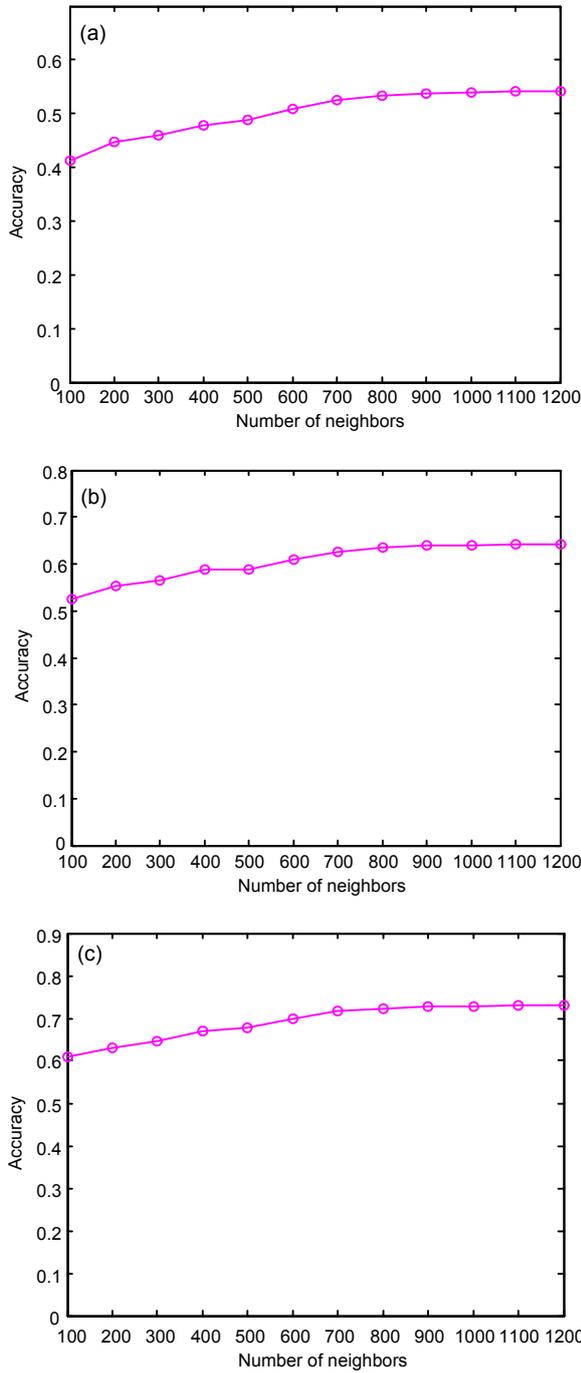


Fig. 3 Performance variation of the RAML algorithm with respect to M on WVU (a), ViHASi (b), and IXMAS (c)

outperform the other methods. It means that using multi-view features can improve the human action recognition accuracy.

The convergence curves of our proposed optimization method on the three datasets are shown in

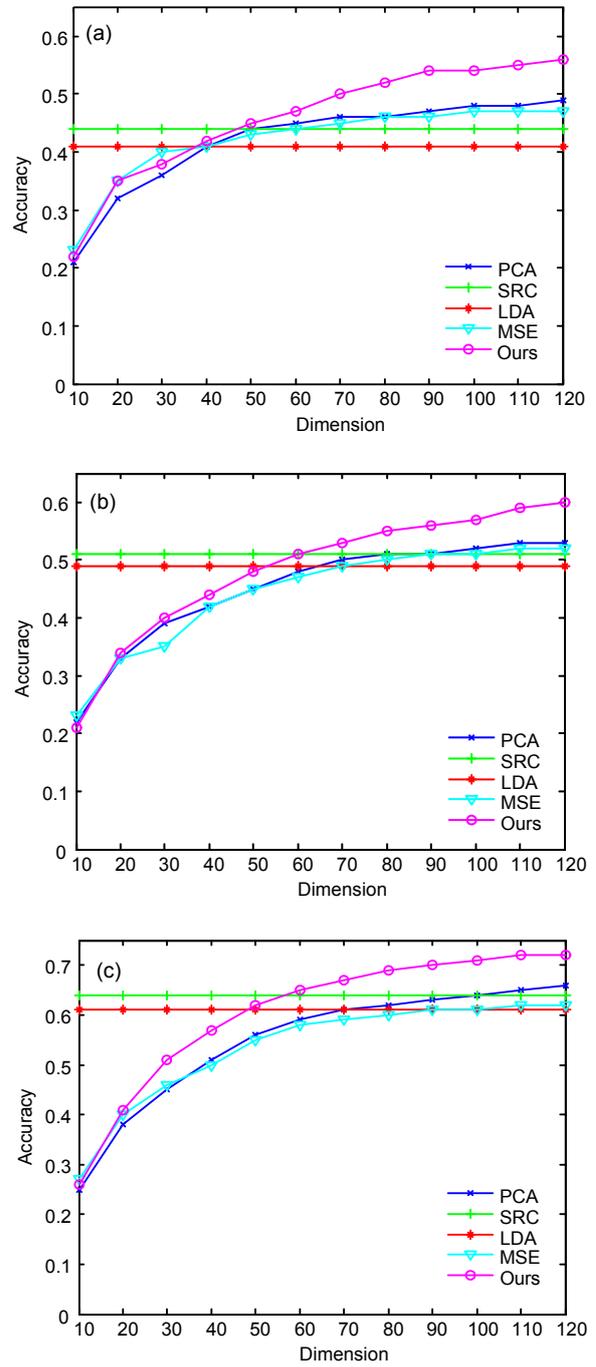


Fig. 4 Performance comparison results of different algorithms on WVU (a), ViHASi (b), and IXMAS (c)

Fig. 5. We can see that our method converges within 200 iterations. So, it indicates that the proposed optimization method is very fast.

Similarly, we conduct another set of experiments, wherein half of the performers are randomly selected

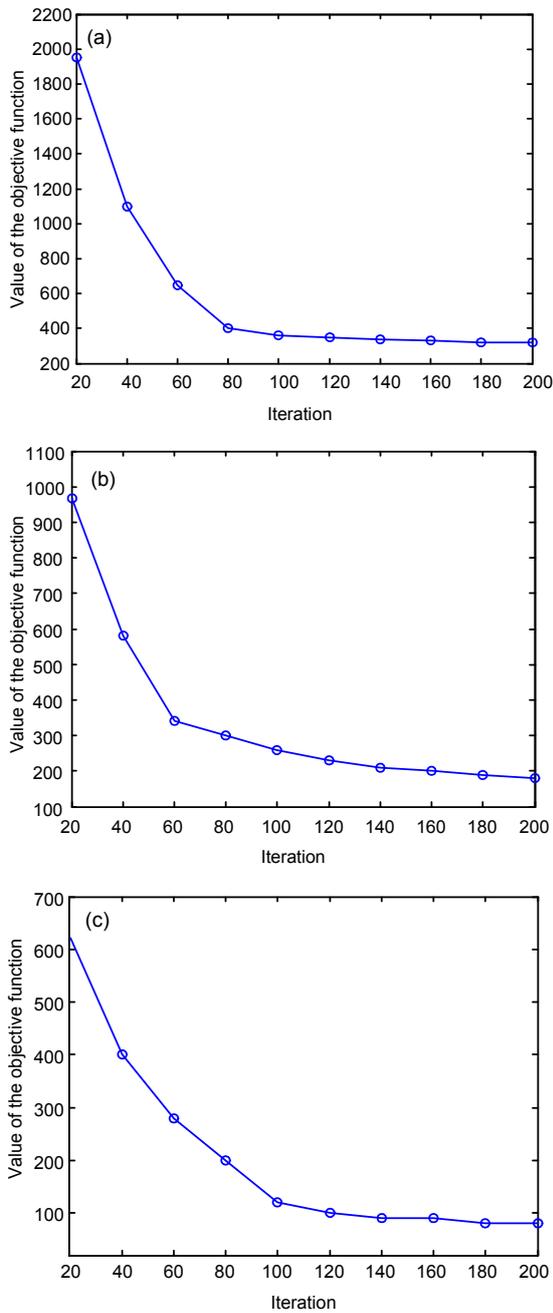


Fig. 5 Convergence curves on WVU (a), ViHASi (b), and IXMAS (c)

as the training dataset, while the remainder are used as the testing dataset. In fact, it is a cross-subject validation. Using the same parameters and repeating all the experiments 30 times, we obtain the results which are shown in Fig. 6. From Fig. 6, we find that if we correctly set the feature dimension (>60), our

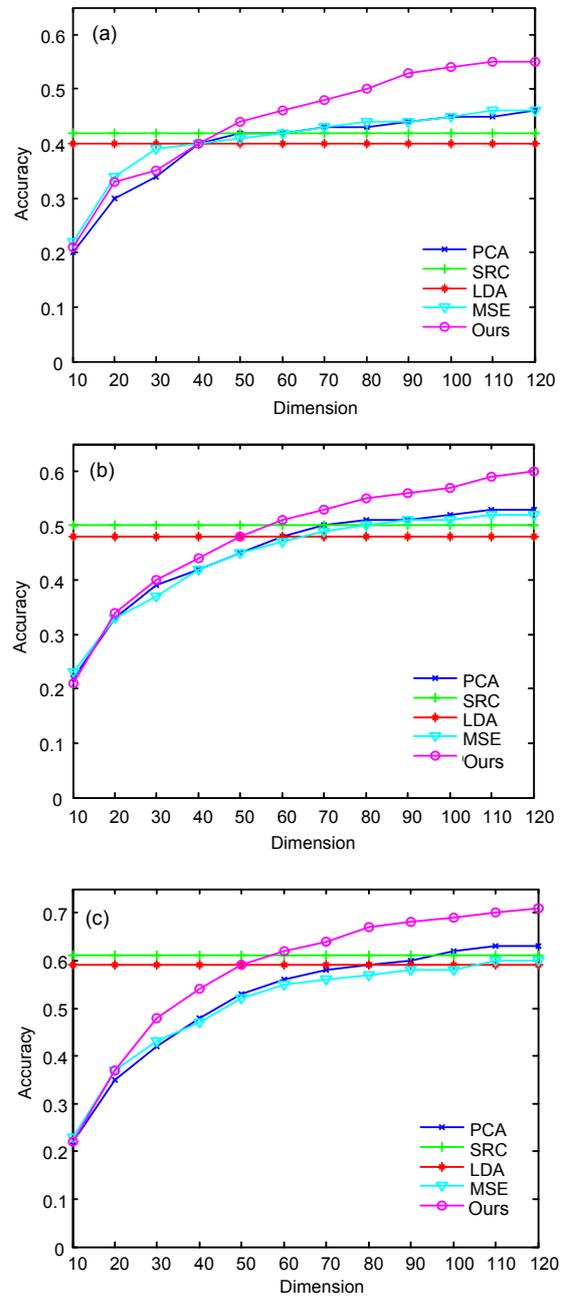


Fig. 6 Performance comparison results of different algorithms on WVU (a), ViHASi (b), and IXMAS (c) in another set of experiments

proposed method will consistently outperform the other methods.

To validate the data adaptive property of our proposed local L1-graph, we count the number of the nearest neighborhood. The results are shown in Fig. 7. As we can see, in Fig. 7a, most of the vertices have

more than 5 but less than 10 nearest neighbors, while a few vertices have less than 5 or more than 20 nearest neighbors. In Figs. 7b and 7c, most of the vertices have more than 5 but less than 10 neighbors or less than 5 neighbors. The underlying reason may be that the data samples in ViHASi are similar to each other if they belong to the same human action, because they are synthesized data, while those in the WVU and IXMAS have more differences.

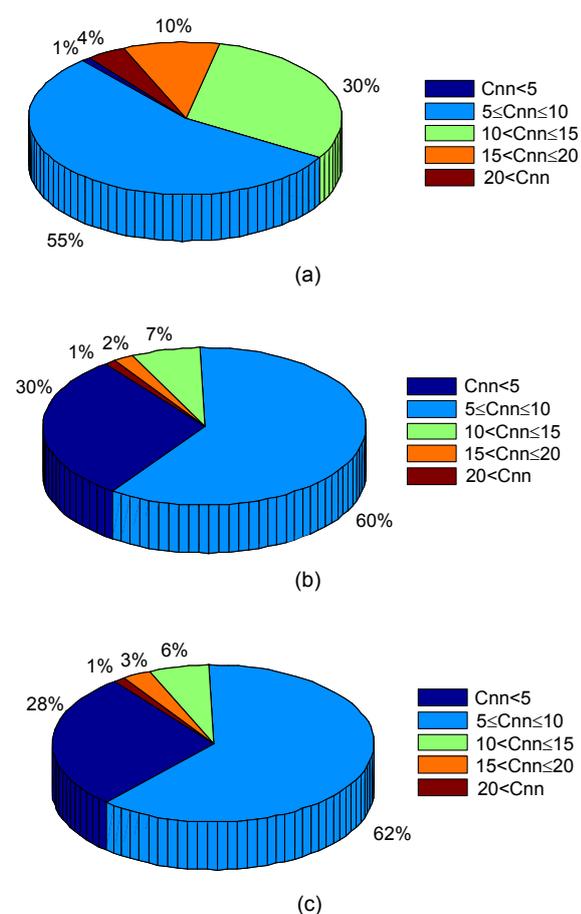


Fig. 7 Statistical results of the number of neighbors of L1-graph on ViHASi (a), WVU (b), and IXMAS (c) (References to color refer to the online version of this figure)

6 Conclusions

In this paper, we proposed a robust locally adaptive multi-view learning (RAML) algorithm based on multiple local L1-graphs and suggested an optimization method for solving it. We have applied the proposed RAML algorithm on three public da-

taset to solve the view-invariant human action recognition problem. Compared with several other methods, the RAML algorithm consistently outperforms the other methods if the selected feature dimension is higher than 60. Meanwhile, we notice that the RAML algorithm involves solving eigenvalue decomposition problem, which is time-consuming. So, there are two key issues that should be carefully considered in the further work. One is how to reduce the computation cost, and the other is how to apply the RAML algorithm to deal with large-scale real-world problems.

References

- Ashraf, A.B., Lucey, S., Chen, T., 2008. Learning patch correspondences for improved viewpoint invariant face recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2008.4587754]
- Balakrishnama, S., Ganapathiraju, A., 1998. Linear Discriminant Analysis—a Brief Tutorial. Institute for Signal and Information Processing, Mississippi State University, USA.
- Balasubramanian, M., Schwartz, E.L., 2002. The isomap algorithm and topological stability. *Science*, **295**(5552):7. [doi:10.1126/science.295.5552.9r]
- Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. Proc. 11th Annual Conf. on Computational Learning Theory, p.92-100. [doi:10.1145/279943.279962]
- Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. *IEEE Trans. Patt. Anal. Mach. Intell.*, **23**(3):257-267. [doi:10.1109/34.910878]
- Brémond, F., Thonnat, M., Zúñiga, M., 2006. Video-understanding framework for automatic behavior recognition. *Behav. Res. Methods*, **38**(3):416-426. [doi:10.3758/BF03192795]
- Candès, E., Romberg, J., 2005. *l₁-Magic: Recovery of Sparse Signals via Convex Programming*.
- Chen, C., Zhuang, Y.T., Xiao, J., 2010. Silhouette representation and matching for 3D pose discrimination—a comparative study. *Image Vis. Comput.*, **28**(4):654-667. [doi:10.1016/j.imavis.2009.10.008]
- Chen, H.S., Chen, H.T., Chen, Y., et al., 2006. Human action recognition using star skeleton. Proc. 4th ACM Int. Workshop on Video Surveillance and Sensor Networks, p.171-178. [doi:10.1145/1178782.1178808]
- Cheng, B., Yang, J., Yan, S., et al., 2010. Learning with *l¹*-graph for image analysis. *IEEE Trans. Image Process.*, **19**(4):858-866. [doi:10.1109/TIP.2009.2038764]
- de Sa Virginia, R., 2005. Spectral clustering with two views.

- Proc. 22nd Annual Int. Conf. on Machine Learning, p.20-27.
- Donoho, D.L., 2006. For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Commun. Pure Appl. Math.*, **59**(6):797-829. [doi:10.1002/cpa.20132]
- Donoho, D.L., Elad, M., Temlyakov, V.N., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, **52**(1):6-18. [doi:10.1109/TIT.2005.860430]
- Feng, J.G., Xiao, J., 2013. View-invariant action recognition: a survey. *J. Image Graph.*, **18**(2):157-168 (in Chinese). [doi:10.11834/jig.20130205]
- Fu, Y., Xian, Y.M., 2001. Image classification based on multi-feature and improved SVM ensemble. *Comput. Eng.*, **37**(21):196-198. [doi:10.3969/j.issn.1000-3428.2011.21.067]
- He, X.F., Cai, D., Yan, S., et al., 2005. Neighborhood preserving embedding. Proc. 10th IEEE Int. Conf. on Computer Vision, p.1208-1213. [doi:10.1109/ICCV.2005.167]
- Jean, F., Bergevin, R., Albu, A.B., 2008. Trajectories normalization for viewpoint invariant gait recognition. Proc. 19th Int. Conf. on Pattern Recognition, p.1-4. [doi:10.1109/ICPR.2008.4761312]
- Junejo, I.N., Dexter, E., Laptev, I., et al., 2008. Cross-view action recognition from temporal self-similarities. Proc. 10th European Conf. on Computer Vision, p.293-306. [doi:10.1007/978-3-540-88688-4_22]
- Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755):788-791. [doi:10.1038/44565]
- Lewandowski, M., Martinez-del-Rincon, J., Makris, D., et al., 2010. Temporal extension of Laplacian eigenmaps for unsupervised dimensionality reduction of time series. Proc. 20th Int. Conf. on Pattern Recognition, p.161-164. [doi:10.1109/ICPR.2010.48]
- Long, B., Yu, P.S., Zhang, Z.F., 2008. A general model for multiple view unsupervised learning. *SIAM*, p.822-833.
- Luo, Y., Wu, T., Hwang, J., 2003. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. *Comput. Vis. Image Understand.*, **92**(2-3):196-216. [doi:10.1016/j.cviu.2003.08.001]
- Mao, J.L., 2013. Adaptive multi-view learning and its application to image classification. *J. Comput. Appl.*, **33**(7):1955-1959 (in Chinese). [doi:10.11772/j.issn.1001-9081.2013.07.1955]
- Natarajan, P., Nevatia, R., 2008. View and scale invariant action recognition using multiview shape-flow models. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2008.4587716]
- Natarajan, P., Singh, V.K., Nevatia, R., 2010. Learning 3D action models from a few 2D videos for view invariant action recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2006-2013. [doi:10.1109/CVPR.2010.5539876]
- Parameswaran, V., Chellappa, R., 2006. View invariance for human action recognition. *Int. J. Comput. Vis.*, **66**(1):83-101. [doi:10.1007/s11263-005-3671-4]
- Rao, C., Yilmaz, A., Shah, M., 2002. View-invariant representation and recognition of actions. *Int. J. Comput. Vis.*, **50**(2):203-226. [doi:10.1023/A:1020350100748]
- Raychev, B., Kikutsugi, Y., Tamaki, T., et al., 2010. Class-specific low-dimensional representation of local features for viewpoint invariant object recognition. Proc. 10th Asian Conf. on Computer Vision, p.250-261. [doi:10.1007/978-3-642-19318-7_20]
- Roh, M., Shin, H., Lee, S., 2010. View-independent human action recognition with volume motion template on single stereo camera. *Patt. Recogn. Lett.*, **31**(7):639-647. [doi:10.1016/j.patrec.2009.11.017]
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500):2323-2326. [doi:10.1126/science.290.5500.2323]
- Shen, B., Si, L., 2010. Nonnegative matrix factorization clustering on multiple manifolds. Proc. 24th AAAI Conf. on Artificial Intelligence, p.575-580.
- Srestasathien, P., Yilmaz, A., 2008. View invariant object recognition. Proc. 19th Int. Conf. on Pattern Recognition, p.1-4. [doi:10.1109/ICPR.2008.4761238]
- Syeda-Mahmood, T., Vasilescu, A., Sethi, S., 2001. Recognizing action events from multiple viewpoints. Proc. IEEE Workshop on Detection and Recognition of Events in Video, p.64-72. [doi:10.1109/EVENT.2001.938868]
- Tang, Y.F., Huang, Z.M., Huang, R.J., et al., 2011. Texture image classification based on multi-feature extraction and SVM classifier. *Comput. Appl. Softw.*, **28**(6):22-46 (in Chinese). [doi:10.3969/j.issn.1000-386X.2011.06.006]
- Tian, C., Fan, G., Gao, X., 2008. Multi-view face recognition by nonlinear tensor decomposition. Proc. 19th Int. Conf. on Pattern Recognition, p.1-4. [doi:10.1109/ICPR.2008.4761195]
- Wang, Y., Huang, K., Tan, T., 2007. Multi-view gymnastic activity recognition with fused HMM. Proc. 8th Asian Conf. on Computer Vision, p.667-677. [doi:10.1007/978-3-540-76386-4_63]
- Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Understand.*, **104**(2-3):249-257. [doi:10.1016/j.cviu.2006.07.013]
- Weinland, D., Boyer, E., Ronfard, R., 2007. Action recognition from arbitrary views using 3D exemplars. Proc. IEEE 11th Int. Conf. on Computer Vision, p.1-7. [doi:10.1109/ICCV.2007.4408849]
- Wen, J.H., Tian, Z., Lin, W., et al., 2011. Feature extraction based on supervised locally linear embedding for classi-

- fication of hyperspectral images. *J. Comput. Appl.*, **31**(3):715-717. [doi:10.3724/SP.J.1087.2011.00715]
- Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. *Chemometr. Intell. Lab. Syst.*, **2**(1-3):37-52. [doi:10.1016/0169-7439(87)80084-9]
- Wright, J., Yang, A.Y., Ganesh, A., et al., 2009. Robust face recognition via sparse representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(2):210-227. [doi:10.1109/TPAMI.2008.79]
- Xia, T., Tao, D.C., Mei, T., et al., 2010. Multiview spectral embedding. *IEEE Trans. Syst. Man Cybern.*, **40**(6): 1438-1446. [doi:10.1109/TSMCB.2009.2039566]
- Yan, P., Khan, S.M., Shah, M., 2008. Learning 4D action feature models for arbitrary view action recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-7. [doi:10.1109/CVPR.2008.4587737]
- Yang, J., Jiang, Y.G., Hauptmann, A.G., et al., 2007. Evaluating bag-of-visual-words representations in scene classification. Proc. Int. Workshop on Multimedia Information Retrieval, p.197-206. [doi:10.1145/1290082.1290111]
- Yilmaz, A., Shah, M., 2005. Actions as objects: a novel action representation. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.984-989. [doi:10.1109/CVPR.2005.58]
- Yu, H., Sun, G., Song, W., et al., 2005. Human motion recognition based on neural network. Proc. Int. Conf. on Communications, Circuits and Systems, p.979-982. [doi:10.1109/ICCCAS.2005.1495271]
- Zheng, S.E., Ye, S.Z., 2006. Semi-supervision and active relevance feedback algorithm for content-based image retrieval. *Comput. Eng. Appl.*, **S1**:81-87 (in Chinese).
- Zhou, D., Burges, C.J.C., 2007. Spectral clustering and transductive learning with multiple views. Proc. 24th Int. Conf. on Machine Learning, p.1159-1166. [doi:10.1145/1273496.1273642]