Multi-view Laplacian Least Squares For Human Emotion Recognition

Lin Feng[a,*], Shuai Guo[a], Zhan-Bo Feng[b], Yi-Hao Li[b], Yang Wang[a], Sheng-Lan Liu[a], Hong Qiao[c]

[a] *School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian 116024, China*
[b] *School of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China*
[c] *State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

Human emotion recognition is an emerging and important area in the field of human-computer interaction and artificial intelligence, which has been more and more related with multi-view learning methods. As an important direction of multi-view learning, most existing subspace learning methods focus on either correlation learning, scatter discriminant or locality protection. Moreover, traditional partial least squares (PLS) has advantage in performance compared with canonical correlation analysis (CCA), while suffers from linear relationship assumption and two-view limitation in subspace learning. In this paper, we propose a new nonlinear multi-view laplacian least squares (MvLLS) that extend PLS to multi-view learning by considering the distance between each two samples in the subspace, and we construct a global laplacian weighted graph to introduce category discriminant information as well as protect the locality. The proposed methods has great extendibility and robustness. Then we make the out-of-sample extension of MvLLS based on the idea of maintaining the original-space manifold structures. Finally we verify the proposed method on two human-emotion datasets, the experiment results indicate the effectiveness of MvLLS over some state-of-art methods.

[*] Corresponding author
Email address: fenglin@dlut.edu.cn (Lin Feng)

Keywords: Human Emotion Recognition, Global Weighted Graph, Iteration Method, Locality, Scatter Discriminant
2010 MSC: 00-01, 99-00

1. Introduction

Human emotion recognition is a long-standing and emerging problem in computer vision and human-computer interaction, where there have been many significant researches. Most of early human emotion recognition researches focused on single view learning methods, such as image-based face emotion recognition [1], video-based body emotion recognition [2, 3], speech emotion recognition [4], and physiological-signal emotion recognition [5, 6]. However, human emotion can be observed at various viewpoints, even by different sensors. And multi-view methods are more robust than single-view methods in most cases. With the rapid expansion of digital multimedia contents and the promotion of contest like Emotion Recognition in the Wild Challenge (EmotiW) [7, 8] in recent years, multi-view human emotion recognition is getting more and more attention [9, 10, 11]. As a result, human emotion recognition and multi-view learning are more and more closely connected. One effective approach of multi-view learning is subspace learning, which aims at getting a common subspace shared by various views. Here we introduce some existing two-view or multi-view subspace learning methods that can further run in supervised mode or unsupervised mode. And in this paper we classify them into three categories according to their key ideas:

Maximize the cross correlations. This line of methods mainly come from canonical correlation analysis (CCA) [12], which attempts to learn two linear transforms for each view to maximize their cross correlation in the subspace. Kernel CCA (KCCA) [13], multiview CCA (MCCA) [14] are the kernel and multiview extensions of CCA respectively. Partial least squares (PLS) [15] uses iterative method and least square regression to approximate a optimization problem that is similar to CCA, but with different constraints. In [16], the kernel version of PLS was introduced. And in [17], a single optimization termed method SVM-2K that combines SVM and KCCA was proposed. Multi-view SVM-2K [18] presents a multi-view modification for SVM-2K. In [19], a non-parametric sparse matrix and automatic model was proposed for multi-feature fusion, by imposing label information across all views to exploit the correlations. Recently, J.-X. Li et al. proposed a hierarchical multi-view multi-feature fusion method HMMF [20] which used a sparse covariance matrix to represent the correlations over all views. However, HMMF has a great spatiotemporal complexity.

Balance the scatter discriminant information. Inspired by linear discriminant analysis (LDA) [21], the scatter discriminant information of multi-view learning refers chiefly to within-class, between-class, intra-view or inter-view discriminants. Correlation discriminant analysis (CDA) [22] is a supervised improvement of CCA base on the definitions of within-class correlation and between-class correlation. In [23], Sharma et al. proposed a generalized multi-view analysis framework (GMA) which considers the intra-view discriminant information in the common subspace. Multi-view discriminant analysis (MvDA) [24] gets a linear transform for each view by maximizing the between-class variations and minimizing the within-class variations across all views. And MvDA-VC [24] was proposed based on MvDA by adding view-consistency. Besides, multi-view uncorrelated discriminant analysis (MULDA) [25] combines uncorrelated LDA [26] with CCA to preserve both category discriminant information and correlation information. G.-Q. Chao et al. developed a consensus and complementarity maximum entropy discrimination (MED-2C), which takes good advantage of consensus and complementarity principles. In [27], a generalized multi-view embedding method was proposed for CCA, PLS and LDA, using intrinsic and penalty graphs to characterize the intra-view and inter-view discriminant information. And laplacian multi-set canonical correlations (LaMCCs) [28] constructs the nearest neighbor graphs to take intra-view and inter-view correlations into consideration. Multi-view local discrimination and canonical correlation analysis (MLDC^{2}A) [29] aims at optimizing a combination of between-class scatter, within-class scatter and correlation.

Protect the local geometry structure. Protecting the local characteristics would help in many problems. In [30], multi-view spectral embedding (MSE) was developed to obtain a smooth low-dimensional embedding across multiple views. And locality-preserving CCA (LPCCA) [31] integrates the local neighborhood information together with CCA to discover the low-dimensional manifold structures for different views. To solve the problem of large pose variations and facial expression recognition, locality-constrained linear coding is utilized to construct the model in [32]. Y.-H. Yuan et al. presented a graph regularized multi-set canonical correlations (GrMCC) [33] to utilize discriminative and intrinsic geometrical information under the framework of correlation analysis. Recently, noticing that Hessian can exploit the intrinsic local geometry of data, Hessian multiset CCA [34] shows superior extrapolating capability with nonlinear multi-view features.

Of course, some of these methods may have more than two of the three ideas mentioned above, we still classify them with the most import idea. Specially, with the rapid development of deep learning in recent years, many multi-view deep learning methods were proposed. In [35], Hang Su et al. proposed a multi-view Convolutional neural network (MVCNN) that combines information from multi-views into a single shape. And multi-view deep network [36] uses a view-specific sub-network to remove the view-specific variations, followed with a common sub-network to obtain the common representation of all views. Besides, multiple deep neural networks are combined for different data view to perform the EmotiW contest in [9].

1.1. Motivations

Although many significant subspace learning methods have been proposed, there is still much room for improvement. The main motivations of this paper are elaborated as follows:

(1) Most of two-view learning methods like CCA and CDA are difficult to extend to more views, and the unsupervised methods like MCCA and PLS

would not make full use of the category information. And most methods above are linear methods, there are few nonlinear subspace learning methods proposed. Moreover, the lack of category information or features of some samples in particular views could make most of existing methods hard to work.

(2) After some researches, we found that PLS get much better experimental results compared with CCA, even they have similar optimization targets. This may boils down to the iteration and regression methods used in PLS to approximate the properest results. However, on the one hand, in the recognition and classification tasks, traditional PLS suffers from the basic assumption that latent linear relationships exist between different views, which does not stand always. On the other hand, the basic framework of PLS could neither extend to more views nor make use of category information.

1.2. Contributions

In this paper, we present a nonlinear multi-view laplacian least squares (MvLLS) method for the human emotion recognition, and a weighted local preserving embedding (WLPE) method is introduced to get the out-of-sample extensions of MvLLS. The main contributions and characteristics of our works are summarized as below:

(1) MvLLS combines laplacian eigenmaps (LE) [37] with PLS, and uses iterative method as well as generalized eigen-decomposition to solve the optimization problem. By minimizing the sum of weighted distance between samples of all views, MvLLS learns nonlinear low-dimensional embeddings for each view. MvLLS makes use of category discriminant information and maintains the local geometric information in the subspace. As far as we are concerned, the proposed method gets the best accuracy in the datasets of our experiments, compared with state-of-art methods.

(2) Inspired by LE, a global laplacian weighted graph (GLWP) is constructed across all views where the local neighbor information is maintained. The

category discriminant information is used to determine whether samples are connected or not. And samples of different views are projected into a common dimension to metric the weights of their connections. Moreover, the global weighted graph is not only insensitive to the missing of part samples or labels, but also easy to extend to more views. The robustness of GLWP could even help MvLLS to perform the semi-supervised or unsupervised tasks.

(3) Traditional PLS is upgraded to a supervised, nonlinear and multi-view modification MvLLS, which is more appropriate to human emotion recognition. Having abandoned the basic assumption that latent linear relationships exist between views, MvLLS gets a score vector of subspace in each iteration with GLWP. Specially, as GLWP upgrades after each iteration, we could predict the intrinsic dimension of the subspace as GLWP tends to be stable.

(4) As the proposed MvLLS is a nonlinear method, we present an out-of-sample method to process the new samples. To create the embeddings of new samples, we examine the local geometric structures of the new samples in the training set. Then each new sample is represented as the weighted sum of its neighbors. We keep the local geometric structure in the subspace, and the neighbors that connected with more samples are expected to have higher weights. Afterwards, the embeddings of new samples can thus be thought as weighted sum of embeddings of their neighbors.

1.3. Organization

The remainder of this paper is organized as below: In Section 2 we define the common notations and review some related works. Then Section 3 introduces the formulations and optimizations of MvLLS in detail, as well as some discussions. The experimental results are presented in Section 4 with qualitative and quantitative evaluations, followed with a conclusion.

2. Related works

In this section, we define the common notations that are used throughout this paper firstly. Then we introduce some researches that are related to our work, including PCA, CCA, LE, PLS and MvDA.

2.1. Notations

Suppose that we are give samples of various different views and their labels. For each view, we expect to learn a low-dimensional embedding in the common subspace. Important notations in this paper are listed in Table 1:

Table 1: Definitions of important notations

Notation	Description
$X_i \in \mathbb{R}^{d_i \times n_i}$	All n_i samples of i^{th} view
$Label_i \in \mathbb{R}^{n_i}$	Labels that correspond to samples in X^i
$x_{ia}, x_{jb} \in \mathbb{R}^d$	The a^{th} and b^{th} of view X_i and X_j
$label_{ia}, label_{jb}$	The labels of x_{ia} and x_{jb}
v	The number of views
c	The number of class across all views
dim	The number of dimensions of the subspace
$w_i \in \mathbb{R}^{d_i}$	A basic vector of the linear transform of X_i
$Y_i \in \mathbb{R}^{dim \times n_i}$	The embedding of X_i in the subspace
$y_{ia}, y_{jb} \in \mathbb{R}^{dim}$	The embeddings of x_{ia} and x_{jb}
$y^i \in \mathbb{R}^{n_i}$ $y^j \in \mathbb{R}^{n_j}$	A basic vector of Y_i and Y_j
W_{ab}^{ij}	The weight of connection between x_{ia} and x_{jb}
I	The identity matrix
$tr(X)$	Trace of matrix X

When we refer to single-view learning methods, the view $X \in \mathbb{R}^{d \times n}$ is used. And when we refer to methods that require the equality of number of samples, we use n instead of n_i. In some other specific situations, we would remove upper and lower corner markers as needed.

2.2. PCA and CCA

PCA [38] is a classical dimensional reduction and single view learning method. With normalization as the first step, PCA finds a group of standard orthogonal basis to maximum the variance of view X after projection, which can be transformed as the trace of covariance matrix of projection of X:

$$\begin{aligned} &\max_{w} w^T X X^T w \\ &s.t.\ w^T X X^T w = 1 \end{aligned} \tag{1}$$

CCA [12] is a two-view learning method that attempts to find two linear transforms w_1 and w_2 for the normalized feature matrices X_1 and X_2, such that their embeddings in the common subspace are most correlated:

$$\begin{aligned} &\max_{w_1, w_2} w_1^T X_1 X_2^T w_2 \\ &s.t.\ w_1^T X_1 X_1^T w_1 = 1, w_2^T X_2 X_2^T w_2 = 1 \end{aligned} \tag{2}$$

Eq. 1 and Eq. 2 can be solved by using Lagrange multiplier method and resorting the eigenvalue decomposition. CCA can be ragarded as the the two-view version of PCA. And one of the main drawbacks of CCA is that the number of samples of the two views should be equal.

2.3. Laplacian eigenmaps

Laplacian Eigenmaps (LE) [37] is an effective single-view nonlinear manifold learning method. Given the n samples of X, LE has great locality preserving properties by constructing a weighted graph W:

$$W_{ab} = \begin{cases} exp(-\frac{\|x_a - x_b\|_2^2}{t}), & if\ \|x_a - x_b\|_2^2 < \epsilon \\ 0, & else \end{cases} \tag{3}$$

In W, the neighboring samples are connected with weighted edges to record the locality information. Let y_a and y_b denote the low-dimensional representation of a^{th} and b^{th} sample, LE maps the weighted graph W to a low-dimensional space with connected samples stay as close together as possible:

$$
\begin{aligned}
\min_{Y} \frac{1}{2} \sum_{a,b} \|y_a - y_b\|_2^2 W_{ab} = \min_{Y} tr(Y^T L Y) \\
s.t.\ Y^T D Y = I \\
D_{kk} = \sum_{j=1}^{n} W_{jk},\ L = D - W
\end{aligned} \tag{4}
$$

In Eq. 4, D is a diagonal matrix and L is the laplacian matrix. This equation can be solved by computing eigenvalues and eigenvectors of a generalized eigenvector problem.

2.4. Partial least squares

Partial least squares regression (PLS) [15] has similar optimization target with CCA, it can be regarded as a linear unsupervised two-view learning method. Traditional PLS maximizes the correlations of the two views after projection. PLS supposes that the two views can be driven by a few latent variables, which are not directly observed or measured. PLS uses iterative method to predict X_2 with X_1 by finding the component of X_1 and using it as the regressor of both views. After each iteration step, X_1 and X_2 are covered by the residual matrices E and F:

$$
\begin{aligned}
\max_{w_1,w_2} y_1 y_2^T &= \max_{w_1,w_2} w_1^T X_1 X_2^T w_2 \\
s.t.\ w_1^T w_1 &= 1, w_2^T w_2 = 1 \\
X_1 &= P^T Y_1 + E \\
X_2 &= Q^T Y_2 + F \\
Y_2 &= D Y_1 + H
\end{aligned} \tag{5}
$$

PLS supposes X_1 and X_2 have the same number of samples like CCA, that is $n_1 = n_2 = n$. In Eq. 5, $P \in \mathbb{R}^{dim \times d_1}$, $Q \in \mathbb{R}^{dim \times d_2}$ are the loading matrices. $E \in \mathbb{R}^{d_1 \times n}$, $F \in \mathbb{R}^{d_2 \times n}$ and $H \in \mathbb{R}^{dim \times n}$ are the residual matrices. And $D \in \mathbb{R}^{dim \times dim}$ shows the latent scores of two views. But in the cross-view classification problems, the latent linear relationships between different views do not always exist.

2.5. Multi-view discriminant analysis

Inspired by LDA [21], the main idea of MvDA [24] is to find v linear transforms to project samples of all views to a common subspace, where the between-class variation S_B^y is maximized and the within-class variation S_W^y is minimized.

$$
\begin{aligned}
&\max_{w_1,\cdots,w_v} \frac{tr(S_B^y)}{tr(S_W^y)} \\
S_W^y &= \sum_{i=1}^{v}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}}(y_{ijk}-\mu_j)(y_{ijk}-\mu_j)^T \\
S_B^y &= \sum_{i=1}^{c} n_i(\mu_i-\mu)(\mu_i-\mu)^T
\end{aligned} \tag{6}
$$

where n_{ij} is the number of samples that are labeled as j in X_i, y_{ijk} is the label of k^{th} sample in X_i that is labeled as j, μ_j is the mean of low-dimensional embeddings of all samples in the v views that are labeled as j, and μ is the mean of low-dimensional embeddings of all samples in the v views. Supposing that each view can be a flipping of any other one, MvDA-VC is proposed by adding a view-consistency term to the denominator of Eq. 6.

MvDA and MvDA-VC take both between-class and within-class discriminant information into consideration. They can deal with the cases where there are different number of samples or classes for the v views. However, the integrity and local structure of samples in a single view is broke up, as S_W^y and S_B^y are calculated across all views.

3. Multi-view laplacian least squares

In this section we present the detailed ideas and optimizations of multi-view laplacian least squares (MvLLS), followed with some discussions.

3.1. Overview

Basing on the framework of LE and PLS, MvLLS is proposed to get the nonlinear subspace embedding of each view directly. By constructing a global laplacian weighted graph (GLWP) $\mathcal{W}$ over all views, MvLLS aims to find a subspace for all views where the connected samples stay as close together as possible. The global graph can not only protect the local geometry structure, but also introduce the category discriminant information. As features of the v views lie on different dimensions, a preprocessed dimension reduction algorithm PCA [38] is applied to make samples of different views measurable. The basic idea is shown in Fig. 1. And inspired by PLS, we use iterative and regression methods to solve the optimization problem. We remove the estimated variations and update $\mathcal{W}$ in each iteration, the variations are predicted by low-dimensional embeddings of views. After that, an out-of-extension method weighted local preserving embedding (WLPE) is introduced to get the embeddings of new samples, which attempts to maintain the high-dimensional local neighborhood information in the subspace.

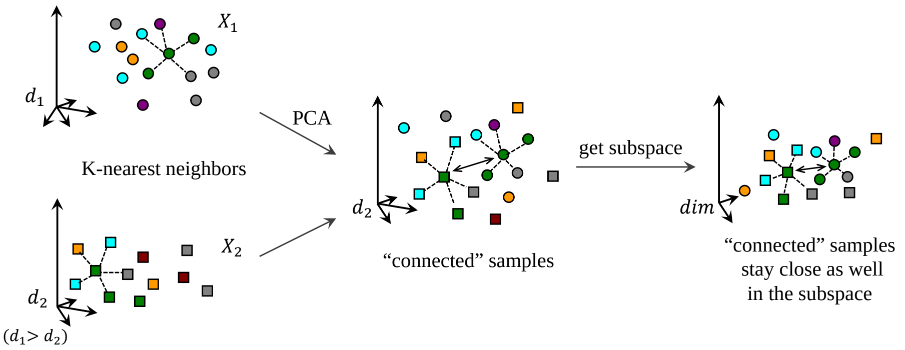

Figure 1: The basic idea of MvLLS. Suppose that we are given samples of X_1 and X_2, PCA is used to get a common space for X_1 and X_2. Whether samples are connected or not depends on labels of themselves and their neighbors, which is explained in detail in Eq. 8. And the connected samples would stay close as well in the subspace got by MvLLS. Squares and circulars denote samples from X_1 and X_2 respectively, labels of the samples are represented by different colours.

The proposed method takes both advantages of scatter discriminant balance and locality protection. With a robust weighted graph, MvLLS can deal with the lack of partial samples or category information. One of the great advantages over other methods is that, MvLLS can also work on unsupervised or semi-supervised models, even the cases that the number of samples are different for the given views. Moreover, as the weighted graph varies by iteration, the change rate can be used to predict the intrinsic dimension of the subspace got by MvLLS.

3.2. Global laplacian weighted graph

In this subsection we construct the global laplacian weighted graph that weights the connections of samples across all views, and get the embeddings of all samples on the first dimension of the subspace, which are $y^1, \cdots, y^v$. Across all views, MvLLS tries to find a subspace in which the connected samples of each two views stay as close together as possible.

We talk about connections between two views firstly, with X_i and X_j as an example:

$$\begin{aligned}
\min_{y^i, y^j} \xi(i,j) &= \frac{1}{2} \sum_{\substack{1 \le a \le n_i \\ 1 \le b \le n_j}} \|y_{ia} - y_{jb}\|_2^2 W_{ab}^{ij} \\
&= \frac{1}{2} \sum_{a,b} (y_a^2 + y_b^2 - 2 y_a y_b) W_{ab}^{ij} \\
&= tr(Y_i^T L^{ij} Y_j) \\
&= y^i L^{ij} y^{jT} \\
D_{kk} = &\sum_{j=1}^{n_i} W_{jk}, \;\; L^{ij} = D - W^{ij}
\end{aligned} \tag{7}$$

In Eq. 7, $W^{ij} \in \mathbb{R}^{n_i \times n_j}$ is a weighted matrix which measures the weights of connections between samples of X_i and X_j, D is a diagonal matrix, and L^{ij} is the laplacian matrix corresponds to W^{ij}. As the dimensions of the two views are different, we use PCA to project X_i and X_j to a common dimension which is the minimum of d_i and d_j. This step is considered taken by default in the following. And in W^{ij}, the weighted edge W^{ij}_{ab} that connects X_{ia} and X_{jb} is defined as below:

$$W^{ij}_{ab} = \begin{cases} exp(-\frac{\|X_{ia}-X_{jb}\|_1}{t}), & if\ X_{ia}\ and\ X_{jb}\ are\ \text{``}connected\text{''} \\ 0, & else \end{cases} \quad (8)$$

We use $Leighby_{ia}$ and $Leighby_{jb}$ to denote the labels of the K-nearest neighbors of X_{ia} and X_{jb}, which is measured by 1-norm distance. If $label_{ia} \in Leighby_{jb}$ and $label_{jb} \in Leighby_{ia}$, we would think X_{ia} is "connected" with X_{jb}, such that the category discriminant information is introduced. Then we would put an edge between X_{ia} and X_{jb} based on 1-norm distance as Eq. 8 shows.

Then for all of the v views, MvLLS protects all connections between each two views:

$$\min_{y^1,\cdots,y^v} \sum_{1\le i,j\le v} \xi(i,j) = \sum_{1\le i,j\le v} y^i L^{ij} y^{jT} \quad (9)$$

$$= \begin{bmatrix} y^1, y^2, \cdots, y^v \end{bmatrix} \begin{bmatrix} L^{11} & L^{12} & \cdots & L^{1v} \\ L^{21} & L^{22} & \cdots & L^{2v} \\ \vdots & \vdots & \vdots & \vdots \\ L^{v1} & L^{v2} & \cdots & L^{vv} \end{bmatrix} \begin{bmatrix} y^{1T} \\ y^{2T} \\ \vdots \\ y^{vT} \end{bmatrix}$$

$$= \mathcal{Y}\mathcal{L}\mathcal{Y}^T \quad (10)$$

$$s.t.\ \mathcal{Y}\mathcal{D}\mathcal{Y}^T = 1 \quad (11)$$

where $\mathcal{Y} = [y^1, y^2, \cdots, y^v]$, $\mathcal{L}$ is composed of $[L^{11}, \cdots, L^{vv}]$ as shown above, and $\mathcal{W}$ is composed of $[W^{11}, \cdots, W^{vv}]$:

$$\mathcal{W} = \begin{bmatrix} W^{11} & W^{12} & \cdots & W^{1v} \\ W^{21} & W^{22} & \cdots & W^{2v} \\ \vdots & \vdots & \vdots & \vdots \\ W^{v1} & W^{v2} & \cdots & W^{vv} \end{bmatrix} \quad (12)$$

$\mathcal{W}$ is what we called the global laplacian weighted graph (GLWP). Despite L^{ij} is not a square matrix, as n_i may differ from n_j. But when we come to all v views in Eq. 10, $\mathcal{L}$, $\mathcal{W}$ and $\mathcal{D}$ are exactly square matrices. Lagrange multiplier method is then employed to obtain the optimum solution of this problem, and the constraint that $\mathcal{Y}^T\mathcal{D}\mathcal{Y}^T = 1$ simplifies the problem greatly. Finally this problem boils down to a problem of getting the smallest nonzero generalized eigenvalue and feature vector:

$$\mathcal{L}\mathcal{Y}^T = \lambda\mathcal{D}\mathcal{Y}^T \tag{13}$$

3.3. Laplacian partial least squares

Inspired by PLS, after getting the embeddings of samples $\mathcal{Y}$ which is the embeddings of all views on the first dimension of the subspace, we use $\mathcal{Y}$ as the regressor of $\mathcal{W}$ to predict it. In another word, MvLLS remove all estimated variation from $\mathcal{W}$ [39].

$$\begin{align} \mathcal{W} = \ & P^T\mathcal{Y} + E \tag{14} \\ & \tag{15} \end{align}$$

where $P \in \mathbb{R}^{1\times(n_1+\cdots+n_v)}$ is a loading vector that describes how strong $\mathcal{W}$ is related to $\mathcal{Y}$. E is the residual matrix.

$$\begin{align} U &= \mathcal{Y}\mathcal{W} \tag{16} \\ P &= \frac{U\mathcal{W}}{UU^T} \tag{17} \\ E &= \mathcal{W} - P^T\mathcal{Y} \tag{18} \end{align}$$

As shown in Eq. 16, we use $\mathcal{Y}$ as weight vector to produce the score of $\mathcal{W}$ that is U, and therefore finding the loading vector P. Residual matrix E is got by removing the found variation of $\mathcal{W}$. Then $(E + E^T)/2$ is assigned to $\mathcal{W}$ to keep the symmetrical characteristic of $\mathcal{W}$, and the negative values of $\mathcal{W}$ is set to zero. With the new GLWP $\mathcal{W}$, we can get a new embedding $\mathcal{Y}$ as introduced in Eq. 9. This procedure is repeated for dim times to get the embeddings on each dimension of the target subspace.

3.4. Weighted local preserving embedding

In the previous subsections, we get the embeddings of each view in the subspace with MvLLS. As MvLLS is a nonlinear method, we introduce a weighted local preserving embedding (WLPE) to get its out-of-sample extension. WLPE looks for an embedding in the subspace that maintains the local neighbors of the high-dimensional space, and the neighbors with higher degrees in GLWM are expected to have greater weights. And WLPE keeps the global nonlinear of MvLLS from the local linear fits. Firstly, we get the 2-norm distance K-nearest neighbors for every sample of each view in the original space, including its own. Then we characterize the local neighborhood structure by linear confidence, in other words, each sample is represented by a linear weighted sum of its neighbors. And the reconstruction error is minimized:

$$\min_{M} \epsilon(M) \sum_{1 \le a \le n} \|x_a - \sum_{1 \le b \le K} M_{ab} x_b\|_F^2 \tag{19}$$

$$s.t. \sum_{b} M_{ab} = 1 \tag{20}$$

where x_b is one of the K neighbors of x_a, M_{ab} shows the contribution of x_b to the construction of x_a. Lagrangian multiplier method is then employed to get the optimum solution:

$$\begin{gathered} 2M_a = \frac{1}{2}\lambda(G_a G_a^T)^{-1} e \\ G_a = \begin{bmatrix} M_{a1}, M_{a2}, \cdots, M_{ak} \end{bmatrix} \begin{bmatrix} x_a - x_1 \\ x_a - x_2 \\ \vdots \\ x_a - x_k \end{bmatrix} \end{gathered} \tag{21}$$

In Eq. 21, e is a zero column vector except the a^{th} element of which is 1. Then the neighbors that weighted greater in $\mathcal{W}$ are expected to be more confident:

$$M'_{ab} = M_{ab} + \sum_{1 \leq i \leq n_1 + \cdots + n_v} \mathcal{W}_{bi} \tag{22}$$

Then M' is normalized. Afterwords, M' is used to predict the low-dimensional embeddings of new samples. The embedding of sample a would be:

$$y_a = \sum_{1 \leq b \leq K} y_b M'_{ab} \tag{23}$$

Finally, the whole precess of MvLLS is summarized in Algorithm 1.

Algorithm 1 Whole process of MvLLS

```
Input: Feature samples of all views $X_1, \cdots, X_v$ and their labels $Label_1, \cdots, Label_v$; Target dimensions of subspace $dim$, nearest neighbors parameter $K$.
Output: Embeddings of all views in the subspace, $Y_1, \cdots, Y_v$.
Get $W^{11}, \cdots, W^{vv}$, piece them into $\mathcal{W}$ and get $\mathcal{L}$.
Repeat
    Get the embedding of samples on a dimension of the subspace $\mathcal{Y}$ with Eq. 13;
    Project $\mathcal{W}$ on $\mathcal{Y}$, remove the predicted variation from $\mathcal{W}$ to get a new graph, using Eq. 16;
    Make the new graph a symmetric and nonzero matrix;
    Replace $\mathcal{W}$ with the new graph, and construct new $\mathcal{L}$, $\mathcal{D}$;
For $dim$ times
Get the $K$ neighbors reconstruction contribution matrix $M'$ for each view, with Eq. 21 and Eq. 22;
Make out-of-sample extension for new samples with Eq. 23.
```

3.5. Discussions

In this subsection, we further explain the strong points, disadvantages, and some details of MvLLS, then discuss the difference between MvLLS and some related works in detail. Firstly, we present some explanations and discussions in detail.

(1) The main disadvantage of MvLLS is that the time complexity tends to be high, as we should calculate a generalized eigenvalue for each interaction.

(2) The reason why we used PCA as a preprocessing step rather than other methods is that, PCA may be the most classic, effective and widely used dimensional reduction method.

(3) How does MvLLS introduce category information and protect the local structure at the same time? On the one hand, category information determines whether samples are connected or not, such that category discriminant is introduced. And samples of the same class are always connected with each other, as their K-nearest neighbors include themselves. Samples of different classes are still connected, if they are within the K-nearest neighbors of each other. On the other hand, MvLLS expects to find a subspace where the connected samples stay close as well. Therefore, all of the within-class or adjacent samples tend to stay close in the subspace, while between-class samples that are nonadjacent would have no effect to the subspace.

(4) The unsupervised and semi-supervised versions of MvLLS. As the construction of $\mathcal{W}$ is very flexible, MvLLS has its unsupervised and semi-supervised versions. We reconstruct W^{ij} with x_{ia} and x_{jb} here, then the new $\mathcal{W}$ can be pieced up. For unsupervised learning, if neither $label_a^i$ nor $label_b^j$ are given, x_{ia} and x_{jb} would be "connected" when $\|x_{ia} - x_{jb}\|_2^2 < \epsilon$, where ϵ is an adjustable parameter. For semi-supervised learning, if $label_a^i$ is given while $label_b^j$ is not, they would be "connected" when both of the two samples are within the K-nearest neighbors of each other, or $label_{ia} \in Leighby_{jb}$. There could be many other judgement methods certainly, basing on different

starting points and optimization targets.

(5) Estimation of intrinsic dimension of the subspace. As the global laplacian weighted graph $\mathcal{W}$ upgrades slowly with each iteration, the intrinsic dimension could be estimated when $\mathcal{W}$ tends to be stable. Suppose that the weighted graph after a single iteration is $\mathcal{W}'$, we would think the subspace is stable when $\|\mathcal{W}' - \mathcal{W}\|_F^2 < \epsilon$, where ϵ is an adjustable parameter.

Then we give detailed comparisons between our approach and some related or newly-proposed works, including MCCA, PLS, MULDA, MvDA, GrMCC and MLDC^{2}A.

(1) Difference from MCCA [14] and PLS [15]. Both of MCCA and MvLLS are solved by iterative method. MCCA focus on maximizing the sum of correlations between each two views, while MvLLS considers category discriminant and local structure information. From the view point of Eq. 9, MvLLS could be regarded as the weighted sum of correlations between projections of the view in the subspace. PLS is a two-view learning method that supposes the two views can be predicted by some latent variables. But instead of finding the latent relationships, MvLLS builds a global graph GLWP to connect all views. Then all views are predicted at the same time, which means that MvLLS does not rely on any latent relationship assumption. And MvLLS can be considered as a multi-view extension of PLS.

(2) Difference from MULDA [25] and MvDA [24]. MULDA is a two-view learning method that maximizes the combination of between-class variances of each view and correlation of the two views. And MvDA finds a linear transformer for the multiple views to maximize the between-class variance as well as minimize the within-class variance, both of which are calculated across all views. However, MULDA ignores the intra-view information, and seems to be difficult to extend to more views. Such that MvDA breaks up the integrity of each view. And both of MULDA and MvDA do not take local structures into accounts. MvLLS is a nonlinear method that exploits the category discriminant information and protect the locality with a global

graph GLWP, the integrity of each view is maintained as well. Between-class and non-neighbor samples would have no influence to the target subspace got by MvLLS, such that within-class variation tends to be small. However, MvLLS does not address the between-class variance in detail.

(3) Difference from GrMCC [33] and MLDC^{2}A [29]. GrMCC constructs the local between-class scatters for each view respectively, and considers the interaction relationships between each two different views. While MvLLS constructs a global graph across all views. MLDC^{2}A maximizes the between-class margin and minimizes the within-class distance, for both intra-view and inter-view samples. As a multi-view learning method based on graph construction, the main difference from GrMCC and MLDC^{2}A is that MvLLS is a nonlinear method. MLDC^{2}A builds two neighbor graphs for between-class scatter and within-class scatter respectively. The two graphs may contradict with each other sometimes, comparing with the unified and well-designed graph in MvLLS. In addition, MvLLS is optimized with iteration method.

4. Experiments

In this section, we compare the performance of MvLLS with some state-of-art methods on two human emotion datasets: RGB-D human video-emotion dataset and KDEF human face-emotion dataset. We evaluate the proposed method on human emotion recognition across sensors, poses, and features. Many experiments are designed to show the effeteness of MvLLS over the state-of-art methods, some discussions and comparisons are presented as well.

4.1. Dataset descriptions and feature representations

Firstly, RGB-D human video-emotion dataset [40] includes 4224 clips of RGB video and 4224 clips of Depth video which belong to 7 emotion categories: angry, disgusted, fearful, happy, neutral, sad, and surprised. To get these clips, professional actors were employed to perform human emotion scripts that are designed under psychological principles. Besides, these clips were collected at a changeless scene to avoid the influence of environments, and 3 Kinect-2.0 cameras were used to record the performances from angles of right, middle and left of the stage at the same time. In Fig. 2, some discontinuous frames show a performance of sad emotion, collected by the three Kinect-2.0 cameras at the same time.

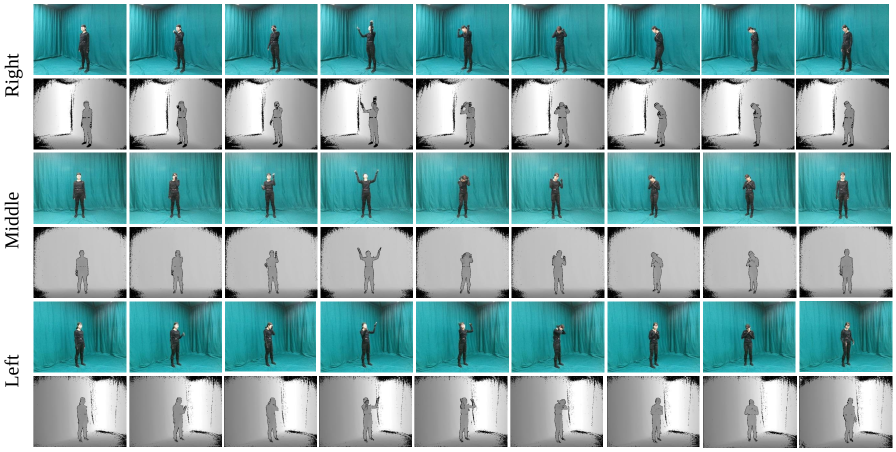

Figure 2: An example of sad emotion from RGB-D human video-emotion dataset.

For the RGB-D human video-emotion dataset, we extracted the 3-Dimensional Convolutional Neural Networks (3D-CNN) feature. 3D-CNN is a popular and effective method for video learning and feature extraction, which constructs features from dimensions of spatial and temporal. As a modified version of BVLC_caffe to support 3D-CNN, C3D-1.0 [41, 42, 43] trained on UCF-101 [44] is used to extract the features. And features of RGB view and Depth view both have a dimension of 4096.

Secondly, the Karolinska Directed Emotional Faces (KDEF) dataset [45, 46] consists of 4900 pictures of 7 human facial emotions: afraid, angry, disgusted, happy, neutral, sad, surprised. A total number of 70 participants (35 male and 35 female) were employed to perform these emotions from 5 different angles: full left, half left, straight, half right and full right. In this paper We select angles of half left, half right and straight from the 5 angles, which include 2940 images totally. Fig. 3 shows examples of surprised, angry and happy emotions from half left, straight and half right angles in KDEF dataset.

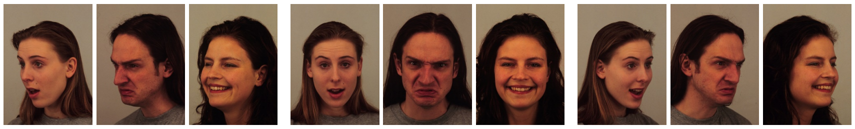

Figure 3: Examples of surprised, angry and happy from three different angles in KDEF dataset.

We extract three features for each selected image, with the 2nd last layer of Xception network [47], the 2nd last layer of Inception network [48], and the 6th last layer of MobileNet [49]. All networks are pretrained on ImageNet [50]. And the dimensions of the features are 2048, 1024 and 2048 respectively.

4.2. Comparison methods and experimental setting

We evaluated the proposed method with as many as two-view and multi-view learning methods that we can compare with, including CCA, PLS, MCCA, MULDA, GMA, MvDA and MvDA-VC. All experiments about the comparison methods are conducted with available codes.

After extracting the features of both datasets, C3D feature of RGB view and Depth view from the RGB-D human video-emotion dataset are used in human emotion recognition across sensors (multi-sensor task); MobileNet feature of half left, half right and straight angles of KDEF dataset are used in human emotion recognition across poses (multi-pose task); and Xception feature, Inception feature, and MobileNet feature of straight view in KDEF dataset are used in human emotion recognition across features (multi-feature task).

If there is no special explanation, the dimension of subspace *dim* of each experiment is set to the minimum of its all views, to preserve as much as energy. Each experiment is randomly repeated for 10 times and the results are averaged. In multi-sensor experiments, the parameter t of Eq. 8 is set to 1000, the number of neighbors K is set to 50; in multi-pose experiments, t is set to 200 and K is set to 300. And ELM [51] is used as the classifier for all experiments. ELM is a classical feedforward network with a single hidden layer, the number of hidden layer node is set to 8000, with *sigmoid* function as the activation function.

4.3. Comparisons and evaluations

4.3.1. Classification accuracy evaluation

In this section, we compare MvLLS with some two-view or multi-view learning method on the three tasks, 2/3 data of each view is randomly selected for training with the others for testing. Table 2 and Table 3 show the average accuracy of each view for both datasets before the subspace learning. Table 4 shows the comparisons between MvLLS and many other methods on human emotion recognition of the multi-sensor, multi-pose and multi-feature task. CCA and PLS could not be used in multi-pose task and multi-feature task as they are two-view learning methods, such that the corresponding spaces in Table 4 are marked as "-". Experimental results indicate that MvLLS performs better than other methods. In the multi-sensor emotion recognition task, average classification accuracy is improved by 2.42% (=48.41%-45.59%) from MvDA-VC, the best performing comparison method. And in the multi-pose emotion recognition task, the average accuracy is improved by 4.38% (=74.02%-69.64%) from MvDA-VC; in the multi-feature emotion recognition task, the average accuracy is improved by 3.57% (75.71%-72.14%) from MvDA-VC. Specially, it can be observed that MvLLS performs better than any single view before learning.

Table 2: The average accuracy of each view for RGB-D video-emotion dataset (%)

Feature	RGB view	Depth view
C3D feature	37.97	31.72

4.3.2. Evaluation of classification performance with different training sizes

To analyse the robustness of the proposed method, we evaluate the average classification accuracy with different training sizes. We take multi-pose and

Table 3: The average recognize accuracy of each view for KDEF dataset (%)

Feature	Half Left	Half Right	Straight
Xception feature	64.11	65.09	72.18
Inception feature	56.50	57.79	68.07
MobileNet feature	66.27	66.07	73.74

Table 4: Comparisons on three tasks in terms of average accuracy (%)

Task	CCA	PLS	MCCA	MULDA	GMA	MvDA	MvDA-VC	MvLLS
Multi-sensor	13.08	40.63	15.15	32.63	35.20	43.65	45.99	48.41
Multi-pose	-	-	35.02	43.21	47.50	61.07	69.64	74.02
Multi-feature	-	-	38.57	46.63	56.79	60.02	72.14	75.71

multi-feature task as examples. Fig. 4 and Fig. 5 show the accuracies with different training sizes for multi-pose and multi-feature task respectively. The training size is setted in change interval of [10, 90] with the step size of 10. And from these tables, we can observe that MvLLS outperforms than all related methods with each training size.

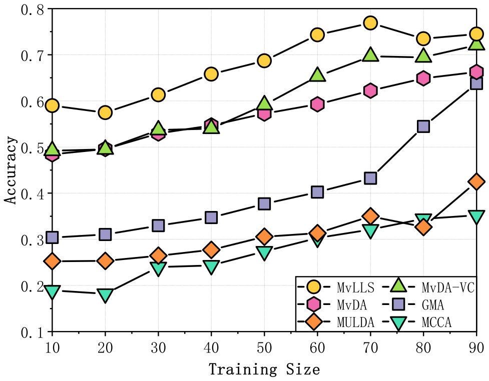

Figure 4: Comparisons of average accuracy for multi-pose task with different training Sizes.

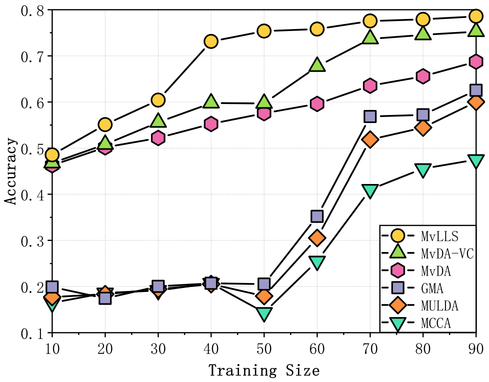

Figure 5: Comparisons of average accuracy for multi-feature task with different training sizes.

4.3.3. Evaluation with different dimensions of the subspace

In this section we compare the performance of each method on different dimensions of the subspace *dim*. We take the multi-pose task and multi-feature task as examples as well. And for the both tasks, *dim* varies in change interval of [100, 1000] with the step size of 100. Fig. 6 and Fig. 7 indicate that MvLLS could get the best classification accuracy with different target dimensions of the subspace.

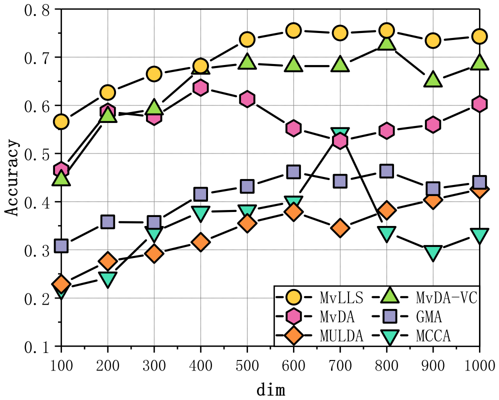

Figure 6: Comparisons of average accuracy for multi-pose task on different target dimensions.

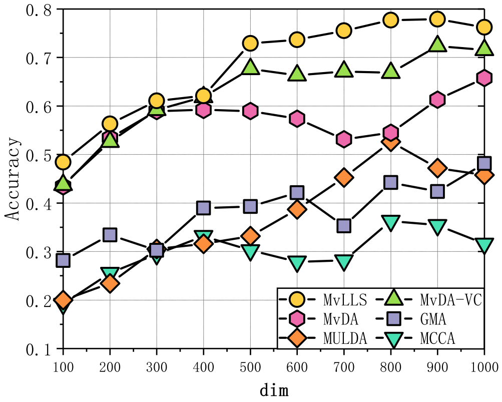

Figure 7: Comparisons of average accuracy for multi-feature task on different target dimensions.

4.3.4. Evaluation of the effectiveness of interactive method

470 To analyse the effectiveness of interactive method used in MvLLS, we conduct experiments on the three tasks to compare the average accuracy between MvLLS and its non-iterative version (called MvLLS$_{\text{noITE}}$). Similarly, 2/3 data of each view is used as the training set with the others as the test set. From Table 8 it can be observed that the average classify accuracy is improved by 475 nearly 1% for each task when iterative method is used.

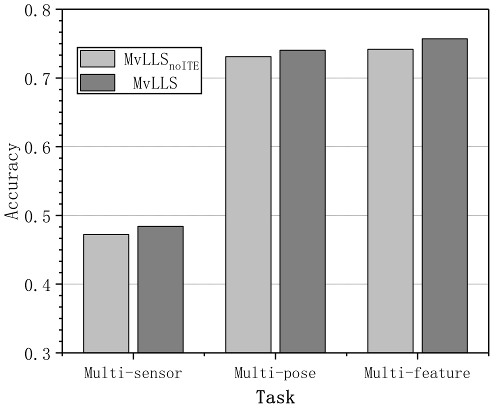

Figure 8: Comparisons of average accuracy between MvLLS and MvLLS$_{\text{noITE}}$

4.3.5. Evaluation of influence of the parameters

The coefficient t and the number of neighbor K in the neighborhood are two important parameters in MvLLS. Taking the multi-sensor task as an example, we conduct experiments to evaluate the influence of the two parameters. As shown in Table 5, K is set as 50 and t varies in the interval of [600, 1400] with the step size of 100. And in Table 6, t is set as 1000 and K varies in the interval of [10, 90] with the step size of 10. We can draw the conclusion that t has little influence on MvLLS, but MvLLS performs worse when the value of K is extremely low. That is due to each sample would not be represented well and not many samples would be connected with few neighbors.

Table 5: Influence of t in the multi-sensor recognition task (%)

	600	700	800	900	1000	1100	1200	1300	1400
Accuracy	48.33	48.41	48.72	48.25	48.88	48.64	48.52	48.71	48.53

Table 6: Influence of K in the multi-sensor recognition task (%)

	10	20	30	40	50	60	70	80	90
Accuracy	43.86	46.53	47.17	48.28	48.41	48.33	48.41	48.64	48.32

4.4. Analysis

Overall, MvLLS has higher average accuracy compared with other methods. The improvements of MvLLS in average accuracy could be boiled down to its great nonlinear learning, scatter discriminant balance and locality protection characters. CCA and GMA perform poorly as the inter-view and intra-view information is not considered. By using iterative approximation, PLS performs much better than CCA. MvDA jointly learns between-class variation and within-class variation across all views, while the entirety of each view and locality is not taken into consideration. And MvDA-VC has a significant improvement than MvDA by adding view-consistency. The proposed method MvLLS takes advantages of PLS, category discriminant information and locality information, then get better performance in average accuracy.

MvLLS shows great robustness in the experiments. The proposed method performs well when training size varies, this may attribute to the good properties of the global graph GLWP and out-of-sample extension method WLPE. GLWP makes full use of category information and locality of the training set, and WLPE gets new samples involved in known samples of the subspace. MvLLS performs well when *dim* is extremely low, which means the subspace got by MvLLS is more representative and typical. And as MvLLS is optimized with interactive method, the performance is proved essentially stable with different parameters.

5. Conclusion

In this paper, a flexible and extensible nonlinear method multi-view laplacian least square (MvLLS) is proposed for multi-view human-emotion recognition. With the global laplacian weighted graph (GLWP), MvLLS records the local structures, introduces the category discriminant information, and protects the local information. MvLLS finds a common subspace across all views where the connected samples stay close to each other as well. MvLLS is optimized with interactive method, and the weighted local preserving embedding (WLPE) is the out-of-sample extension of MvLLS. Experimental results verified the effectiveness of the proposed method.

In the future, we will work to reduce the time complexity of MvLLS, and evaluate MvLLS with more datasets.

Acknowledgments

This study was funded by National Natural Science Foundation of People's Republic of China (No. 61672130, 61602082, 91648205), the National Key Scientific Instrument and Equipment Development Project (No. 61627808), the Development of Science and Technology of Guangdong Province Special Fund

Project Grants (No. 2016B090910001). All the authors declare that they have no conflict of interest.

References

[1] H. Rosenberg, S. McDonald, M. Dethier, R. P. Kessels, R. F. Westbrook, Facial emotion recognition deficits following moderate–severe traumatic brain injury (tbi): Re-examining the valence effect and the role of emotion intensity, Journal of the International Neuropsychological Society 20 (10) (2014) 994–1003.

[2] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, C. Pal, Recurrent neural networks for emotion recognition in video, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 467–474.

[3] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 445–450.

[4] B. Schuller, G. Rigoll, M. Lang, Hidden markov model-based speech emotion recognition, in: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, Vol. 2, IEEE, 2003, pp. II–1.

[5] A. M. Bhatti, M. Majid, S. M. Anwar, B. Khan, Human emotion recognition and analysis in response to audio music using brain signals, Computers in Human Behavior 65 (2016) 267–275.

[6] M. Murugappan, S. Murugappan, Human emotion recognition through short time electroencephalogram (eeg) signals using fast fourier transform (fft), in: Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on, IEEE, 2013, pp. 289–294.

[7] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 423–426.

[8] A. Dhall, R. Goecke, J. Joshi, J. Hoey, T. Gedeon, Emotiw 2016: Video and group-level emotion recognition challenges, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 427–432.

[9] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, 2013, pp. 543–550.

[10] Z. Tong, W. Zheng, C. Zhen, Z. Yuan, J. Yan, K. Yan, A deep neural network driven feature learning method for multi-view facial expression recognition, IEEE Transactions on Multimedia 18 (12) (2016) 2528–2536.

[11] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, B. Schuller, Enhanced semi-supervised learning for multimodal emotion recognition, in: IEEE International Conference on Acoustics, 2016.

[12] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.

[13] S. Akaho, A kernel method for canonical correlation analysis, arXiv preprint cs/0609071.

[14] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Conference on Data Mining and Data Warehouses (SiKDD 2010), 2010, pp. 1–4.

[15] A. Sharma, D. W. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch.

[16] R. Rosipal, L. J. Trejo, Kernel partial least squares regression in reproducing kernel hilbert space, Journal of machine learning research 2 (Dec) (2001) 97–123.

[17] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-taylor, S. Szedmak, Two view learning: Svm-2k, theory and practice, in: Advances in neural information processing systems, 2006, pp. 355–362.

[18] S. Szedmak, J. Shawe-Taylor, Synthesis of maximum margin and multiview learning using unlabeled data, Neurocomputing 70 (7-9) (2007) 1254–1264.

[19] H. Liu, L. Liu, T. D. Le, I. Lee, S. Sun, J. Li, Nonparametric sparse matrix decomposition for cross-view dimensionality reduction, IEEE Transactions on Multimedia 19 (8) (2017) 1848–1859.

[20] J. Li, H. Yong, B. Zhang, M. Li, L. Zhang, D. Zhang, A probabilistic hierarchical model for multi-view and multi-feature classification, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[21] R. O. Duda, P. E. Hart, Pattern classification and scene analysis, A Wiley-Interscience Publication, New York: Wiley, 1973.

[22] Y. Ma, S. Lao, E. Takikawa, M. Kawade, Discriminant analysis in correlation similarity measure space, in: Proceedings of the 24th international conference on Machine learning, ACM, 2007, pp. 577–584.

[23] A. Sharma, A. Kumar, H. Daume, D. W. Jacobs, Generalized multiview analysis: A discriminative latent space, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2160–2167.

[24] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE transactions on pattern analysis and machine intelligence 38 (1) (2016) 188–194.

[25] S. Sun, X. Xie, M. Yang, Multiview uncorrelated discriminant analysis, IEEE transactions on cybernetics 46 (12) (2016) 3272–3284.

[26] Z. Jin, J. Y. Yang, Z. S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, Pattern Recognition 34 (7) (2001) 1405–1416.

[27] G. Cao, A. Iosifidis, K. Chen, M. Gabbouj, Generalized multi-view embedding for visual recognition and cross-modal retrieval, IEEE transactions on cybernetics 48 (9) (2018) 2542–2555.

[28] Y. H. Yuan, Y. Li, X. B. Shen, Q. S. Sun, J. L. Yang, Laplacian multiset canonical correlations for multiview feature extraction and image recognition, Multimedia Tools & Applications 76 (1) (2017) 731–755.

[29] L. Han, X.-Y. Jing, F. Wu, Multi-view local discrimination and canonical correlation analysis for image classification, Neurocomputing 275 (2018) 1087–1098.

[30] T. Xia, D. Tao, T. Mei, Y. Zhang, Multiview spectral embedding, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40 (6) (2010) 1438–1446.

[31] Wang, Fengshan, Zhang, Daoqiang, A new locality-preserving canonical correlation analysis algorithm for;multi-view dimensionality reduction, Neural Processing Letters 37 (2) (2013) 135–146.

[32] J. Wu, Z. Lin, W. Zheng, H. Zha, Locality-constrained linear coding based bi-layer model for multi-view facial expression recognition, Neurocomputing 239 (C) (2017) 143–152.

[33] Y. H. Yuan, Q. S. Sun, Graph regularized multiset canonical correlations with applications to joint feature extraction, Pattern Recognition 47 (12) (2014) 3907–3919.

[34] W. Liu, X. Yang, D. Tao, J. Cheng, Y. Tang, Multiview dimension reduction via hessian multiset canonical correlations, Information Fusion 41 (2017) S1566253517300519.

[35] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 945–953.

[36] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.

[37] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation 15 (6) (2003) 1373–1396.

[38] I. Jolliffe, Principal component analysis, in: International encyclopedia of statistical science, Springer, 2011, pp. 1094–1096.

[39] J. Trygg, S. Wold, Orthogonal projections to latent structures (o-pls), Journal of Chemometrics: A Journal of the Chemometrics Society 16 (3) (2002) 119–128.

[40] S. Liu, S. Guo, H. Qiao, Y. Wang, B. Wang, W. Luo, M. Zhang, K. Zhang, B. Du, Multi-view Laplacian Eigenmaps Based on Bag-of-Neighbors For RGBD Human Emotion Recognition, arXiv e-printsarXiv:1811.03478.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[44] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402.

[45] D. Lundqvist, A. Flykt, A. Öhman, The karolinska directed emotional faces (kdef), CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet 91 (1998) 630.

[46] D. Lundqvist, J. Litton, The averaged karolinska directed emotional faces, Stockholm: Karolinska Institute, Department of Clinical Neuroscience, Section Psychology.

[47] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[49] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[51] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1-3) (2006) 489–501.

Lin Feng received the B.S. degree in electronic technology from Dalian University of Technology, China, in 1992, the M.S. degree in power engineering from Dalian University of Technology, China, in 1995, and the PhD degree in mechanical design and theory from Dalian University ofTechnology, China, in 2004. He is currently a professor and doctoral supervisor in the School of Innovation and Entrepreneurship, Dalian University of Technology, China. His research interests include intelligent image processing, robotics, data mining,and embedded systems.

Shuai Guo received the B.S. degree in the School of Computer Science and Technology from Dalian University of Technology, in 2017. Currently, he is working toward the M.S. degree in the School of Innovation and Entrepreneurship, Dalian University of Technology. His research interests include multi-view learning, dimensionality reduction, image and video learning.

Zhan-Bo Feng is working toward the B.S. degree in the School of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests include reinforcement learning, robot control and collaboration.

Yi-Hao Li is working toward the B.S. degree in the School of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests include computer vision and deep learning.

Yang Wang is a M.S. degree candidate in the School of Innovation and Entrepreneurship, Dalian University of Technology. His research interests include information retrieval, computer vision and machine learning.

Sheng-Lan Liu received the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, China, in 2015. Currently, he is an associate professor with the School of Innovation and Entrepreneurship, Dalian University of Technology, China. His research interests include manifold learning, human perception computing. Dr. Liu is currently the editorial board member of Neurocomputing.

Hong Qiao (SM'06) received the B.Eng. degree in hydraulics and control and the M.Eng. degree in robotics and automation from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree in robotics control from De Montfort University, Leicester, U.K., in 1995. She was an Assistant Professor with the City University of Hong Kong, Hong Kong, and a Lecturer with the University of Manchester, Manchester, U.K., from 1997 to 2004. She is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include robotics, machine learning, and pattern recognition.