Multi-view Laplacian Eigenmaps Based on Bag-of-Neighbors For RGB-D Human Emotion Recognition

Shenglan Liu[a,b], Shuai Guo[a], Wei Wang[b], Hong Qiao[c], Yang Wang[a], Wenbo Luo[d,*]

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

Abstract

Human emotion recognition is an important direction in the field of biometric and information forensics. However, most existing human emotion research are based on single RGB view. In this paper, we introduce a RGB-D video-emotion dataset and a RGB-D face-emotion dataset for research. To our best knowledge, this may be the first RGB-D video-emotion dataset. We propose a new supervised nonlinear multi-view laplacian eigenmaps (MvLE) approach and a multi-hidden-layer out-of-sample network (MHON) for RGB-D human emotion recognition. To get better representations of RGB view and depth view, MvLE is used to map the training set of both views from original space into the common subspace. As RGB view and depth view lie in different spaces,

*Corresponding author
Email addresses: liusl@mail.dlut.edu.cn (Shenglan Liu), guoshuaiabc@mail.dlut.edu.cn (Shuai Guo), wangwei@dlut.edu.cn (Wei Wang), hong.qiao@ia.ac.cn (Hong Qiao), wangyang521@mail.dlut.edu.cn (Yang Wang), luowb@lnnu.edu.cn (Wenbo Luo)

[1]This study was funded by National Natural Science Foundation of People's Republic of China (No. 61672130, 61602082, 91648205, 31871106), 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), the LiaoNing Revitalization Talents Program (No.XLYC1086006).

a new distance metric bag of neighbors (BON) used in MvLE can get the similar distributions of the two views. Finally, MHON is used to get the low-dimensional representations of test data and predict their labels. MvLE can deal with the cases that RGB view and depth view have different size of features, even different number of samples and classes. And our methods can be easily extended to more than two views. The experiment results indicate the effectiveness of our methods over some state-of-art methods.
Keywords: Human emotion recognition, MvLE, BON, MHON, RGB-D

1. Introduction

Human emotion recognition is an emerging and important area in the fields of biometric and information forensics, where there has been many significant researches. Existing researches on human emotion recognition mainly focus on single view methods, such as physiological signals emotion recognition [1, 2, 3], image-based face emotion recognition [4, 5], speech emotion recognition [6, 7], and video emotion recognition [8]. However, in many scenes of human biometric recognition, people can be observed at various viewpoints, even by different sensors. Recently, multi-view emotion recognition gets more attention, combinations of pre-trained models [9], features [10], expressions in face, speech and body [11] come to be important methods of human-emotion recognition. Emotion-recognition is more and more related with multi-view learning methods.

Multi-view learning, which is also known as data fusion or data integration, has three main categories: (1) co-training, (2) multiple kernel learning, and (3) subspace learning. Assuming that input views are generated from a latent subspace, subspace learning is usually used in the task of classification and clustering. Here we introduce some subspace learning methods that are usually used in the task of classification [12, 13, 14, 15].

Two-view unsupervised methods. Canonical correlation analysis (CCA) [16] may be the most typical method of subspace learning. CCA attempts to find

two linear transforms for each view such that the cross correlation between two views are maximized. In [17], a nonlinear version of CCA was provided. Kernel canonical correlation analysis (KCCA) [18] is another improved version of CCA which introduces kernel method and regularization technique. Fukumizu et al. [19] provided a theoretical justification for KCCA. To recognize faces with various poses, partial least squares (PLS) was proposed in [20] , which can be thought as a balance of projection variance and correlation.

Two-view supervised methods. Correlation discriminant analysis [21] (CDA) is a supervised extension of CCA in correlation measure space, which considers the correlation of between-class and within-class samples. Inspired by linear discriminant analysis (LDA) [22], Discriminative canonical correlation analysis (DCCA) [23] proposed by Tae-Kyun Kim et al. maximizes the within-class correlations and minimizes the between-class correlations from different views. A regularized two-view equivalent of fisher discriminant analysis (MFDA) is derived in [24] by employing the category information. In [25], a single optimization termed SVM-2K that combines SVM and KCCA is proposed. Recently, multi-view uncorrelated linear discriminant analysis (MULDA) [26] was proposed by combining uncorrelated LDA [27] and DCCA to preserve both the class structures of each view and the correlations between views.

Multi-view unsupervised methods. Multiview CCA (MCCA) [28] is a multiview extension of CCA, which aims at maximizing the cross correlation of each two views. Multiview spectral embedding (MSE) is a multiview spectral-embedding algorithm [29], which learns a low-dimensional and sufficiently smooth embedding of all views by preserving the locality in the subspace. In [30], low-dimensional patterns are learned from multiple views using principal component analysis (PCA), and a framework of sparse unsupervised subspace learning method is proposed.

Multi-view supervised methods. A multi-view semi-supervised method was proposed in [31] to improve the performance of unknown distribution data, with a modification for the optimization formulation of SVM. In [32], a generic and kernelizable multiview analysis framework (GMA) is proposed for several known

supervised or unsupervised methods. But GMA only considers the intra-view discriminant information. By reproducing kernel Hilbert space, CCA and PCA, mixed kernel canonical correlation analysis (MKCCA) that can be implemented in multi-view learning and supervised learning is proposed. Multi-view discriminant analysis (MvDA) [33] aims at maximizing the between-class variations and minimizing the within-class variations over all views.

The inherent shortage of two-view methods is that it's not easy to extend them to multi-view problems. By using one-versus-one strategy, they have to convert a v-view problem to C_v^2 two-view problems. The main shortage of unsupervised methods is that the label information is not utilized, which may limit their performance in the task of classification. As mentioned above, preserving the local discriminant structure is an important idea. And most of supervised methods above are linear methods that aims at optimizing the correlation of classes or views. We found that human movement or emotion data almost always have some nonlinearities, while most of existing multi-view learning methods perform poorly in nonlinear data as they are linear methods. Although some of them can deal with the nonlinear problems by using kernel functions, but kernel functions take more calculation, and sometimes it's difficult to find a suitable kernel function.

Another fact is that, RGB-D cameras have been widely used in industry and indoor scene. RGB view mainly focus on color difference and changes, but depth view mainly focus on spatial information and depth of field. However, most human emotion recognition researches mainly focus on the RGB view. Compared with the RGB view, depth view is more robust to the influence of environment and other noises. Depth view could reflect the change in distance, concave-convex and structure. The combination of RGB view and depth view has great necessity and importance in human emotion recognition.

In this paper, we use RGB-D cameras (Kinect-2.0) to shoot videos and take images of professional human emotion performances, then we get an a video-emotion dataset and a image-based face-emotion dataset. And we propose a new nonlinear method multi-view laplacian eigenmaps (MvLE) to learn both views,

as well as improving the recognition performance. MvLE is a supervised method where category information and local neighborhood information is introduced. MvLE learns a common subspace for the two views where the "connected" samples stay as close together as possible. As MvLE is a nonlinear and inexplicit method, to adapt to the requirements of large-scale applications, a multi-hidden-layer out-of-sample network (MHON) is proposed based on extreme learning machine (ELM) [34]. MHON could be extended to multi-view problems by adding more input layers. Finally, we evaluate MvLE and MHON on the two human emotion datasets mentioned above, and show both experimentally and theoretically that our framework has a significant improvement compared with some known methods, especially in the video-emotion dataset that has great nonlinearity. Our methods could deal with the cases that different views have different size, even different number of samples or classes.

The major contributions of this paper are summarized as follows:

1. A new multi-view learning method MvLE is proposed to get the low-dimensional representations of training set.
2. A new distance metric BON is introduced to get the similar distributions of different views.
3. A multi-hidden-layer network MHON is proposed to get the low-dimensional representations and predict the labels of test data.
4. Two new RGB-D human-emotion datasets are collected under psychological principles and methods to evaluate the classification performance of proposed method.

In the following, Section 2 reviews some related works of multi-view learning. Section 3 introduces the proposed methods in detail. Section 4 introduces the two human-emotion datasets collected by our own. The experimental results with qualitative and quantitative evaluations are presented in Section 5, followed by a conclusion.

2. Related Works

In this section, we review some existing methods that are related to our works, including LDA, LE, CCA, PLS.

2.1. Notations

Suppose that we are given the samples from many different views, V^i denote the samples of the i^{th} view, which locates in d_i-dimensional vector space, together with labels $Label^i = [label_1^i, label_2^i, \cdots, label_n^i]$. And every $label_k^i \in Label^i$ belongs to the label set $C = \{1, 2, \cdots, c\}$. The multi-view subspace learning methods aim to find a common subspace for various views. Important parameters used in this paper are defined in Table 1.

2.2. Linear Discriminant Analysis

LDA [22] is a linear supervised feature extraction and dimensionality reduction (DR) method of single-view learning. It seeks for a linear transform to map the samples from original space to a low-dimension subspace, such that the between-class variance is maximized and within-class variance is minimized. Let's take V^i as an example:

$$\max_w \frac{w^T S_b w}{w^T S_w w} \tag{1}$$

$$S_w = \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (V_k^{ij} - \mu_j)(V_k^{ij} - \mu_j)^T \tag{2}$$

$$S_b = \sum_{j=1}^{c} n_j (\mu_j - \mu)(\mu_j - \mu)^T \tag{3}$$

$$\tag{4}$$

In Eq. (1), S_b and S_w denote the between-class variance and within-class variance, μ_j denotes the mean of samples in j^{th} class, and μ denotes the mean of all samples in V^i. There are many multi-view learning methods extended from LDA, such as ULDA, MULDA and MvDA.

Table 1: Definitions of Important Parameters

Notation	Description
$V^i \in \mathbb{R}^{d_i \times n_i}$	all n_i samples of i^{th} view
V_k^i	k^{th} sample of i^{th} view
V^{ij}	samples of j^{th} class in i^{th} view
V_k^{ij}	k^{th} sample of j^{th} class in i^{th} view in the subspace
v	the number of views
d_i	dimension of samples in V^i
n	the number of samples of all views
n_i	the number of samples in V^i
n_j	the number of samples of j^{th} class of all views
n_{ij}	the number of samples of j^{th} class in V^i
c	the number of class over all views
W	the weight graph of LE
$W_i \in \mathbb{R}^{d_i \times dim}$	linear transform of the V^i
$w_i \in \mathbb{R}^{d_i}$	basic vector of W_i
$Label^i$	labels of all samples in V^i
$label_k^i$	label of V_k^i
dim	dimension of the common subspace
$Y^i \in \mathbb{R}^{dim \times n_i}$	samples of i^{th} view in the subspace
$y_i \in \mathbb{R}^{dim}$	basic vector of Y_i
Y_k^i	j^{th} sample of i^{th} view in the subspace
I	the identity matrix
$tr(X)$	the trace of symmetric matrix X

2.3. Laplacian Eigenmaps

LE [35] is one of few nonlinear single-view feature extraction and dimensionality reduction methods. Given the n_i samples of V^i, LE constructs a weighted graph W to connect the neighboring samples:

$$W_{ab} = \begin{cases} exp(-\frac{\|V_a^i - V_b^i\|_2^2}{t}), & if\ \|V_a^i - V_b^i\|_2^2 < \epsilon \\ 0, & else \end{cases} \tag{5}$$

With the weighted graph W, LE aims at preserving the local information. Let y_j denotes the low-dimensional representations of j^{th} samples, to choose a good map, the criterion for LE is to minimize the following equation:

$$\begin{gathered} \min_Y \frac{1}{2} \sum_{a,b} \|y_a - y_b\|_2^2 W_{ij} = \min_Y tr(Y^T L Y) \\ s.t.\ Y^T D Y = I \\ D_{kk} = \sum_{j=1}^{n_i} W_{jk},\ L = D - W \end{gathered} \tag{6}$$

In Eq. (6), D is a diagonal matrix, and L is the laplacian matrix. This equation can be solve with lagrange multiplier method and eigenvalue decomposition. LE can obtain the global optima by building a graph incorporating neighborhood information of the view.

2.4. Canonical Correlation Analysis

CCA [16] is a typical unsupervised two-view subspace learning methods, with normalization as the first step. To get a great low-dimensional common subspace, CCA is usually followed with procedure of dimension-reduction algorithm, such as LDA. CCA aims to find two transforms w_1, w_2 for V^1 and V^2 to project the samples of each view into the common subspace, by maximizing the correlation of the two views in the subspace:

$$\begin{aligned} &\max_{w_1,w_2} w_1^T V^1 V^{2T} w_2 \\ s.t.\ &w_1^T V^1 V^{1T} w_1 = 1, w_2^T V^2 V^{2T} w_2 = 1 \end{aligned} \tag{7}$$

With lagrange multiplier method, we can get w_1 and w_2 by resorting to the eigenvalue decomposition. As an unsupervised method, CCA can be regarded as the two-view extension of PCA [36] . The main limitation of CCA is that V^1 and V^2 must have the same number of samples. In addition, CCA can only deal with the two-view learning case.

2.5. Partial Least Squares

PLS [20] is an unsupervised two-view subspace learning method, which models V^1 and V^2 such that:

$$\begin{aligned} V^1 &= P^T Y^1 + E \\ V^2 &= Q^T Y^2 + F \\ Y^2 &= D Y^1 + H \end{aligned} \tag{8}$$

In Eq. (8), $P \in \mathbb{R}^{dim \times d_1}$, $Q \in \mathbb{R}^{dim \times d_2}$ are the matrices of loadings. $E \in \mathbb{R}^{d_1 \times n_1}$, $F \in \mathbb{R}^{d_2 \times n_2}$ and $H \in \mathbb{R}^{dim \times n_2}$ are the residual matrices. Besides, $D \in \mathbb{R}^{dim \times dim}$ relates the latent scores of V^1 and V^2.

$$\begin{aligned} \max_{w_1,w_2} y^1 y^{2T} &= \max_{w_1,w_2} w_1^T V^1 V^{2T} w_2 \\ s.t.\ w_1^T w_1 &= 1, w_2^T w_2 = 1 \\ V^1 &= P^T Y^1 + E \\ V^2 &= Q^T Y^2 + F \end{aligned} \tag{9}$$

PLS tries to correlate the latent score of V^1 and V^2 as well as capturing the variations of Y^1 and Y^2, while CCA only correlates the latent score. And PLS can be solved with iterative method. Compared with CCA, PLS is a balance between projection variance and correlation.

3. Proposed Method

3.1. Overview

Inspired by the effectiveness of building global optima and preserving local neighborhoods, in this section, we present multi-view laplacian eigenmaps (MvLE) and multi-hidden-layer out-of-sample network (MHON). We introduce the basic idea and formulation of MvLE. As the RGB view and depth view lie in completely different spaces, in MvLE we introduce a new distance metrics called bag of neighbors (BON) to get the similar distributions of the two views. BON is based on the label information of K-nearest neighbors, so that the between-class and within-class discriminant information is included. And MvLE can map the training set of both views from original space into common subspace or latent space,

As the label information of test data is unknown and to be predicted, MvLE need an out-of-sample method to adapt to large-scale applications. Inspired by the work in [37], a multi-hidden-layer out-of-sample network (MHON) is proposed based on ELM [34, 38] to solve the problem of out-of-sample extension and predict the labels of test data. MHON is trained on the training set of RGB-D views and their labels, the input of MHON is the original distributions of RGB-D views, the output of MHON is their labels. In the guiding layer of MHON, the low-dimensional representations of training set got by MvLE is used as the leading information and feed forward. For the test data of RGB-D views, MHON can predict their labels, and the low-dimensional representations of test data can be obtained in the guiding layer.

For a RGB-D human-emotion dataset that is divided into a training set and a test set, MvLE is used to get the low-dimensional representations of the training set, and MHON is used to get the low-dimensional representations of the test set. The process of our methods is shown in Fig. 1.

3.2. MvLE based on Bag of Neighbors

For the classification problem, we take two-view learning as an example. Suppose that matrices $V^1 \in \mathbb{R}^{d_1 \times n_1}$ and $V^2 \in \mathbb{R}^{d_2 \times n_2}$ denote the features of

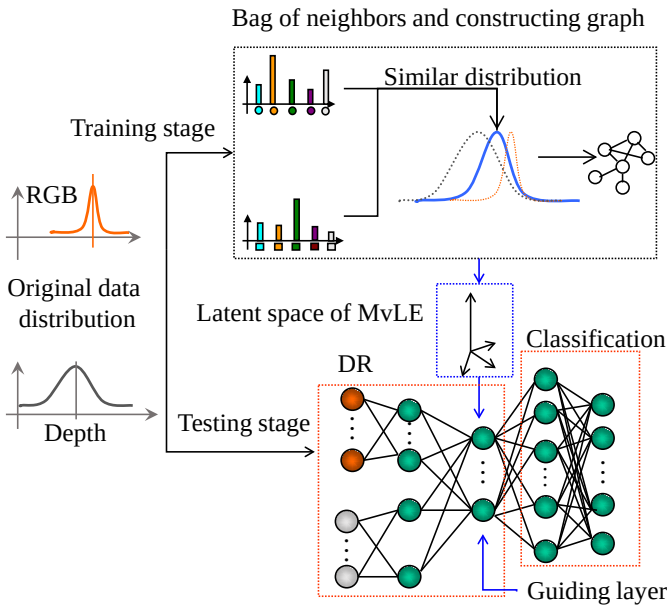

Fig. 1: The process of MvLE and MHON.

RGB view and depth view, c is the number of classes of both views, n_1, n_2 is the size of training set of each view. Note that n_1, d_1 are not necessarily equal to n_2, d_2.

Data normalization is the first step. And then, each sample is represented with a bag of neighbors (BON) vector which has a length of c. Let $BON_k^i = [x_1, x_2, \cdots, x_c]$ denotes the BON vector of sample V_k^i, where x_t denotes the number of samples which are labeled as class t in the K-nearest neighbors of sample V_k^i. The K-nearest neighbors depend on Euclidean distance. In terms of V_a^i and V_b^i, the Euclidean distance is defined as follow:

$$distance_{ab}^i = \|V_a^i - V_b^i\|_2 \tag{10}$$

Fig. 2 shows an example of BON vectors. By introducing BON, samples of each view can get similar distributions. Furthermore, let $W \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ denotes the new weight matrix of the proposed method. Actually, W can be

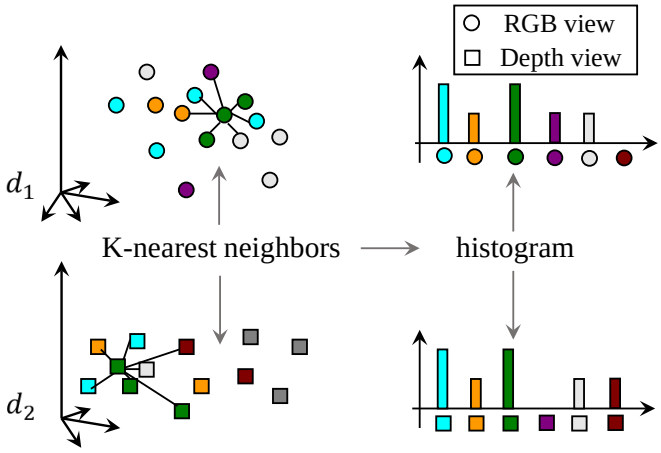

Fig. 2: Bag of neighbors for example.

divided into four parts: $\left[W^{11}, W^{12}; W^{21}, W^{22}\right]$, indicating four inter-view or intra-view similarity measures. And dimensions of them are $n_1 \times n_1$, $n_1 \times n_2$, $n_2 \times n_1$, $n_2 \times n_2$ respectively, as Eq. (11) shows.

$$W = \begin{bmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{bmatrix} \xrightarrow{size} \begin{bmatrix} n_1 \times n_1 & n_1 \times n_2 \\ n_2 \times n_1 & n_2 \times n_2 \end{bmatrix} \tag{11}$$

Compared with traditional laplacian eigenmaps, the weighted matrix W of the proposed method depends on BON vectors got above. Moreover, four parts of W are calculated respectively. Let $Leighb_k^i$ denotes the labels of K-nearest neighbors of V_k^i, and $label_k^i$ denotes the label of V_k^i. For the (a, b) element of W^{ij} above, here $a, b \in \{1, 2\}$. If $label_a^i \in Lneighb_b^j$ and $label_b^j \in Lneighb_a^i$, we would think sample V_a^i is "connected" with sample V_b^j, no matter they are inter-view samples or intra-view samples. Then, BON vectors are naturally used to measure this weight:

$$W_{ab}^{ij} = exp(-\frac{\|BON_a^i - BON_b^j\|_2^2}{t}) \tag{12}$$

In Eq. (12), t is an adjustable constant, which we set as c in the follow-up experiments. And if $label_a^i \notin Lneighb_b^j$ or $label_b^j \notin Lneighb_a^i$, they are not "connected":

$$W_{ab}^{ij} = 0 \tag{13}$$

The new distance metric BON can not only overcome the difference between views, but also introduce category discriminant information. Compared with LDA-based methods that aim at maximizing between-class variance and minimizing within-class variance, such as MvDA and MULDA, MvLE tries to minimize the distance between samples that are "connected". In the proposed method, samples of different classes are almost impossible to be marked as "connected". Accordingly, BON is more insensitive to outliers and noise.

After getting W, subsequent steps are similar to traditional laplacian eigenmaps. Suppose that $Y \in \mathbb{R}^{(n_1+n_2)\times dim}$ denotes the features after fusion, the first n_1 vectors of Y is the low-dimensional representations of V^1, and the last n_2 vectors of Y is the low-dimensional representations of V^2. Let y_a and y_b denote two vectors of Y, v_a and v_b denote the original distribution that corresponding to y_a and y_b. We try to ensure that if v_a and v_b are "connected" and the weight between them is low, y_a and y_b should stay close as well. So the objective function can be written as follow:

$$\begin{aligned} &\min_Y \xi(Y) \ s.t. \ Y^T DY = I \\ \xi(Y) &= \sum_{a,b} \|y_a - y_b\|_2^2 W_{ab} \\ &= \sum_{a,b} (y_a^2 + y_b^2 - 2y_a y_b) W_{ab} \\ &= 2tr(Y^T LY) \end{aligned} \tag{14}$$

The problem boils down to computing eigenvalues and eigenvectors for the generalized eigenvector:

$$
\begin{aligned}
Ly^i =& \lambda D_{ii} y^i \\
D_{ii} =& \sum_{j=1}^{n_1+n_2} W_{ji},\ L = D - W
\end{aligned}
\tag{15}
$$

In Eq. (15), D is a diagonal weight matrix, and L is the Laplacian matrix. Now let $y^1, y^2, \cdots, y^{n_1+n_2}$ ordered by eigenvalues in ascending order denote the solution of Eq. (15). And then Y is given by $[y^2, y^3, \cdots, y^{dim+1}]$, because y^1 corresponds to the smallest eigenvalue which value is 0.

Let Y^1 denotes the first n_1 rows of Y, Y^2 denotes the last n_2 rows of Y. Finally, Y^1 is regarded as features of V_1 after fusion and dimensionality reduction, and Y^2 is regarded as features of V_2 after fusion and dimensionality reduction. In this manner, our method would not be affected by the size of different views.

If more than two views are given, we just need to build the weight graph like the following matrix.

$$
W = \begin{bmatrix}
W^{11} & W^{12} & \cdots & W^{1n} \\
W^{21} & W^{22} & \cdots & W^{2n} \\
\vdots & \vdots & \vdots & \vdots \\
W^{n1} & W^{n2} & \cdots & W^{nn}
\end{bmatrix}
\tag{16}
$$

And after getting Y, the first n_1 rows of Y are regarded as Y^1, the second n_2 rows are regarded as Y^2, $\cdots$, the last n_n rows are regarded as Y^n.

MvLE builds a global weight graph over all views to incorporate the inter-view and intra-view neighborhood information. The size of global graph in this paper is equal to the number of samples of all views. With the interaction of different views, samples of each view can get appropriate representations in the subspace, which helps to get a better performance in classification. As far as we are concerned, there are few researches focusing on building global graph in multi-view learning. Most of existing methods like CCA, PLS, and MvDA did not make full use of inter-view and intra-view information. Another advantage of global graph is that, the multi-view locality-preserving character of MvLE

makes it relatively insensitive to outliers and noise. But the time complexity and spatial complexity tend to be high as well.

3.3. Multi-hidden-layer Out-of-sample Network (MHON)

As a supervised nonlinear multiview learning method, category information is used in BOW to measure the weight between samples. However, category information is only given for the training dataset, for the test dataset, category information is to be predicted. Another fact is that, MvLE cannot get a linear transform for each view. In [37], a nonlinear manifold learning framework QLLP was proposed by Shenglan Liu et al. , they chose a small subset of original data to learn the explicit mapping function from original data to the low-dimensional coordinates. Manifold learning assumes that high-dimensional input data lie on a low-dimensional manifold. And QLLP preserves the local geometry structure as well as the true manifold structure of original space.

Inspired by their work, here we propose a multi-hidden-layer out-of-sample network (MHON) to get the low-dimension representations of test data and predict their labels, as the Fig. 3 shows. MHON is trained on the original distributions of RGB-D views and their labels. In the guiding layer of MHON, low-dimensional representations of training set is used as the leading information and feed forward. For the test data, the input of MHON is original distributions of RGB-D views, MHON can predict their labels in the last layer, and the low-dimensional representations are got in the guiding layer.

In [39], robust activation function (RAF) is proved to be beneficial to the performance of ELM. As a continuous, monotonic and nonlinear active function, RAF is used in the first hidden-layer of MHON to improve the recognition performance. For the second hidden-layer, we use sigmoid active function to fulfill the task of classification.

If more than two views are given, the input of MHON is the original distributions of all views. Their low-dimensional representations and labels are got in the guiding layer ant output layer respectively.

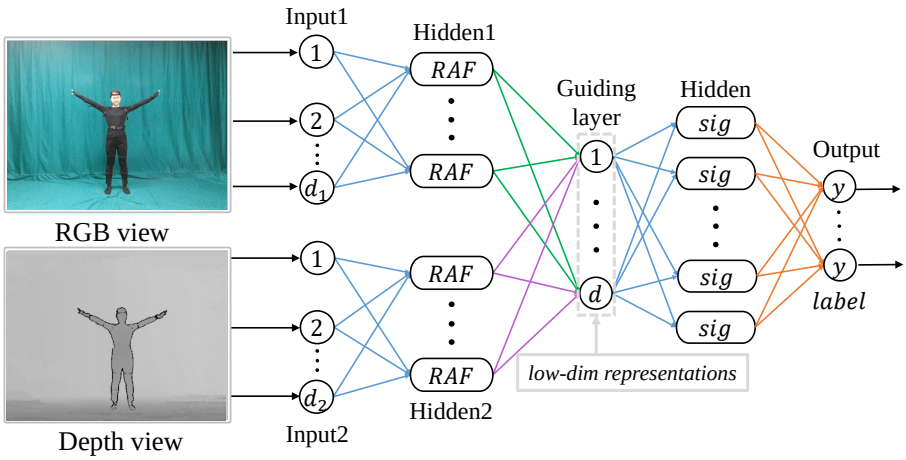

Fig. 3: The overview of MHON.

4. Human Emotion Datasets

The majority of existing human-emotion datasets suffer from two disadvantage: (1) The videos or images in existing datasets could not get rid of the influence of environment. (2) The information provided by a single RGB view seem to be deficient. In this section, we introduce a new RGB-D video-emotion dataset and a new RGB-D face-emotion dataset that are collected at a changeless scene. Compared with RGB view that mainly focus on color difference, depth view has unique advantages by introducing spatial and depth information of the field. The combination of RGB view and depth view would has great necessity and importance to human emotion recognition. As far as we are concerned, there are few RGB-D video-emotion datasets in existence. AFEW [40] is a popular video-emotion dataset composed by videos from movies and reality TV shows. In contrast to AFEW, the video-emotion dataset is designed under psychological principles and well-designed scripts, and there is only one person in an video or image.

4.1. Video-Emotion Dataset

The video-emotion dataset consists of over 4k (4 thousand) clips of RGB videos and 4k clips of depth videos that correspond to each other, and each video has a length of 6 seconds and a resolution of 702×538. It contains the following 7 emotion classes: angry, disgusted, fearful, happy, neutral, sad, and surprised. As a whole-body video-emotion dataset, it also has some significance in human-emotion expression from the view point of psychological [41] . The video-emotion dataset is collected under psychological methods and principles, firstly, we designed a number of 6-seconds length scenes that can show one of the emotions above. For example, jumping and dancing with joy means someone is happy, wiping tears and sobbing means someone is sad. After that, at least 200 people are asked to grade on these scenes. We then know which scenes can show human emotion better. At last, we selected 6 highest score scenes for each emotion as the final scripts.

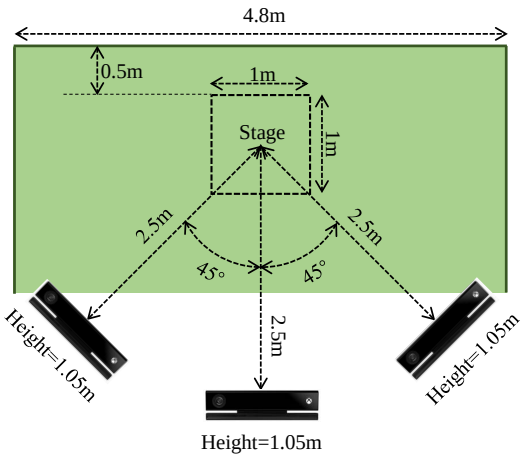

Fig. 4: The scene arrangement of video-emotion dataset.

Furthermore, we employed 24 professional actors to perform these scripts. The background color of the scene is green, and actors perform the scripts at a 1 square meter stage which is centered at the scene. To record their performances, we have 3 Kinect-2.0 cameras shooting RGB-D videos at the same time, which

are placed at front, left, and right of the stage, as Fig. 4 shows. Actors may perform a script more than one time with different body movements. After cutting and editing, we finally get a video-emotion dataset of 7 emotions and 14 hours of RGB-D clips. Fig. 5 shows three examples of this dataset, and each example has 9 discontinuous frames of a RGB clip and a depth clip that correspond to each other.

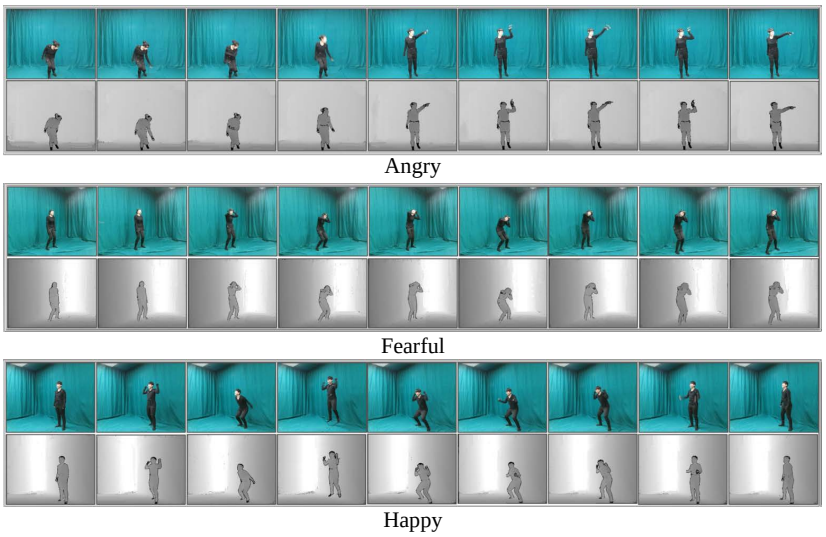

Fig. 5: Three examples in video-emotion RGB-D dataset.

4.2. Face-Emotion Dataset

The face-emotion dataset includes about 1k RGB face emotion images and 1k depth face emotion images that correspond to each other. Less than the video-emotion dataset mentioned above, the face-emotion dataset has 6 emotion classes: angry, afraid, happy, neutral, sad, surprised. We get 69 volunteers to perform all these emotions with facial expressions from 5 different viewpoints, which are front, up, down, left, and right.

In addition, a Kinect-2.0 camera is used to take RGB-D images of the facial emotion. To crop out the background information of the scene, we use Kinect-2.0 to detect the position of head and neck of the actor in the image. Then we

draw a square centered in the position of head in each image, and the width of which depends on the distance between head and neck. With this square, we crop out the background and get the facial emotion images. So that every volunteer have 30 RGB images and 30 depth images taken, the resolution of which is about 150×110. At last, we get a face emotion dataset of 6 emotion classes, 1k RGB images, and 1k depth images.

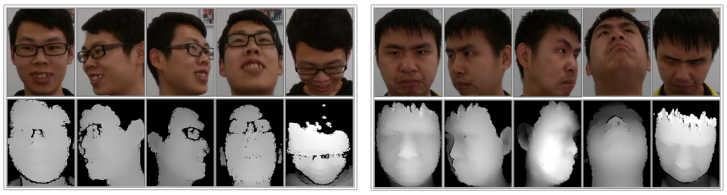

Fig. 6: Two examples in face-emotion RGB-D dataset.

5. Experiments

In this section, we evaluate the proposed method on the video-emotion dataset and face-emotion dataset introduced above. Firstly, we extract C3D features of the video-emotion dataset, and VGG16 [42] features of the face-emotion dataset. By introducing intra-class variance S_W and inter-class variance S_B, we illustrate that both C3D features and VGG16 features are nonlinear features. After that, the quantitative comparisons of average accuracy are presented between our new method, CCA-LDA, PLS, GMA, MvDA and MvDA-VC.

5.1. Features Extraction and nonlinearity analysis

Taking videos as sequences of frames, 3-Dimensional convolutional neural networks (3D-CNN) can capture the spatial and temporal dimensions along with discriminative information. As a popular and effective method for spatiotemporal feature learning and video analyzing, 3D-CNN has been widely used in many researches [43, 44] . In addition, C3D-1.0 [45, 46, 44] trained on

UCF101 [47] is a modified version of BVLC_caffe to support 3D-CNN. We make a fine-tune on C3D-1.0 and extract features of our video-emotion dataset. Besides, convolution neural network has been proved to be extreme useful in image classification, we use the classical network VGG16 [42] to extract features of our face-emotion image dataset.

To evaluate the effectiveness and nonlinearity of the features extracted above, we calculate average intra-class variance S_W and inter-class variance S_B for RGB-D features of the two datasets, which are defined as below:

$$S_W = \frac{1}{c}\sum_{i=1}^{c}\sum_{x\in X_i}\frac{1}{n_i - 1}\|x - \mu_i\|_2^2 \tag{17}$$

$$S_B = \frac{1}{n-1}\sum_{i=1}^{c} n_i\|\mu_i - \mu\|_2^2 \tag{18}$$

where μ denotes the mean of all samples, μ_i denotes mean of samples in class i, N_i denotes the neighborhood of sample i, and α_{ij} denotes the angle between x_j and its orthogonal projection. S_W and S_B measure how far the within-class and between-class samples spread out of their mean. If the value of S_W and S_B is high, within-class samples and between-class samples would be very different. Therefore, the nonlinearity tend to be fine with high values of S_W, or a low value of S_B. We calculate S_W and S_B of RGB data and depth data for C3D features and VGG16 features in Table 2, with RGB data and depth data individually.

Table 2: S_W and S_B of features of two datasets

Variance	Video-Emotion Dataset		Face-Emotion Dataset	
	RGB	depth	RGB	depth
S_W	0.8324	0.8948	0.1431	0.1814
S_B	0.0093	0.0052	0.5970	0.5257

As seen, C3D features of video-emotion dataset have a high value of S_W and an extremely low value of S_B, on the contrary of face-emotion dataset. This

shows that C3D features of video-emotion dataset have great nonlinearity, but the nonlinearity of VGG16 features of face-emotion image dataset is not very well. Furthermore, C3D and VGG16 features of video-emotion and face-emotion dataset are visualized using t-SNE [48] in Fig. 7, in which samples of each class are denoted in color-coded figures. Fig. 7 confirm the analyses of nonlinearity above.

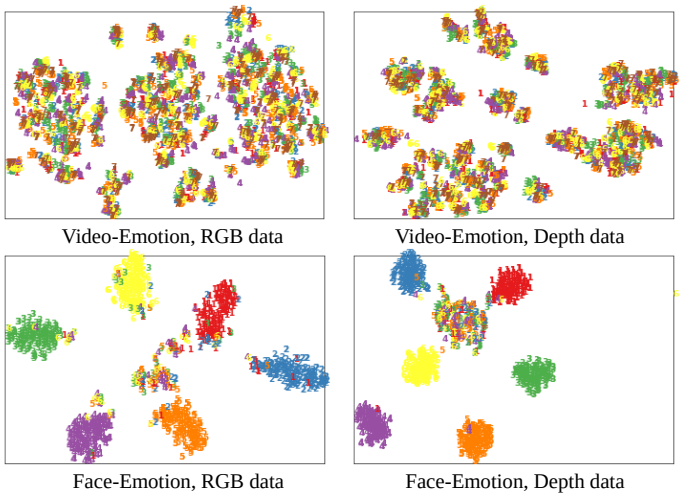

Fig. 7: Visualizing with t-SNE.

5.2. Comparisons And Analyses

Firstly, we calculate the recognition accuracy of original RGB and depth features individually as the references for both datasets, as Table 5 shows. Then we compare the proposed methods with several state-of-art methods, *i.e.* PLS, GMA, MvDA-VC, MvDA and CCA-LDA, which use ELM as the classifier. All these methods use 2/3 data as training set, and the rest as test data. The experiments are randomly repeated for 5 times, and average accuracy are shown as Table 3 and Table 4. Fig. 8 and 9 show the change curve of different dimensions of these methods on our two datasets. For each figure, the left panel is on RGB data, the right panel is on the depth data. The dotted lines denote

the recognition accuracy of RGB or depth data before multi-view learning.

Table 3: Evaluation of Video-Emotion Dataset on different dimensions

Methods	Feature	300	250	200	150	100	50
CCA-LDA	RGB	0.1293	0.1485	0.1489	0.1491	0.1496	0.1493
	depth	0.1309	0.1477	0.1491	0.1486	0.1490	0.1486
PLS	RGB	0.3716	0.3767	0.3746	0.3809	0.3697	0.3563
	depth	0.3201	0.3215	0.3180	0.3040	0.3129	0.3038
GMA	RGB	0.1771	0.1742	0.1834	0.1590	0.1538	0.1617
	depth	0.1502	0.1464	0.1427	0.1389	0.1353	0.1542
MvDA-VC	RGB	0.3812	0.3700	0.3636	0.3142	0.3333	0.2727
	depth	**0.3262**	0.3070	0.2927	0.2807	0.2624	0.2033
MvDA	RGB	0.3427	0.3549	0.3166	0.3325	0.3038	0.2384
	depth	0.3111	0.3086	0.2927	0.2656	0.2376	0.2081
MvLE	RGB	**0.3917**	**0.3734**	**0.3892**	**0.4100**	**0.3949**	**0.3868**
	depth	0.3260	**0.3258**	**0.3322**	**0.3214**	**0.3244**	**0.3086**

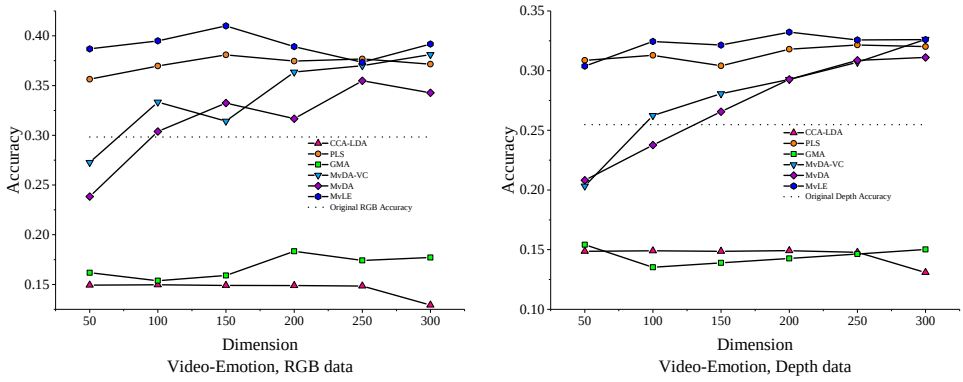

Fig. 8: Change curve of different target dimension for RGB-D video-emotion dataset.

Table 4: Evaluation of face-emotion dataset on different dimensions

Methods	Feature	300	250	200	150	100	50
CCA-LDA	RGB	0.7240	0.7363	0.7454	0.7559	0.7710	0.7642
	depth	0.7179	0.7300	0.7392	0.7491	0.7561	0.7636
PLS	RGB	**0.8587**	0.8591	0.8591	0.8569	0.8555	0.8390
	depth	0.7832	0.7800	0.7872	0.7930	0.8034	0.8108
GMA	RGB	0.8535	0.8574	0.8552	0.8567	0.8565	**0.8625**
	depth	0.8166	0.8237	0.8234	0.8239	0.8217	0.8244
MvDA-VC	RGB	0.8565	**0.8612**	0.8581	0.8572	0.8581	0.8601
	depth	0.8215	0.8203	0.8275	0.7877	0.8082	0.7877
MvDA	RGB	0.8432	0.8456	0.8492	0.8492	0.8540	0.8565
	depth	0.8263	0.8287	0.8251	0.8275	0.8082	0.8251
MvLE	RGB	0.8583	0.8589	**0.8637**	**0.8589**	**0.8613**	0.8616
	depth	**0.8306**	**0.8328**	**0.8335**	**0.8323**	**0.8323**	**0.8316**

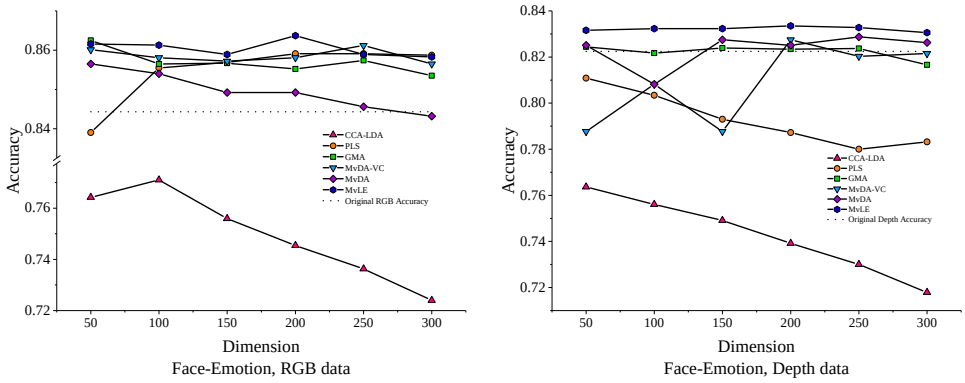

Fig. 9: Change curve of different target dimension for RGB-D face-emotion dataset.

Table 5: Average Accuracy of Each Feature Before Fusion

Datasets	RGB	depth
video-emotion	0.2983	0.2547
face-emotion	0.8443	0.8224

5.3. Discussions

Experimental results indicate that for both datasets, MvLE could not only performs better but also more stably as dimension decreases. For video-emotion dataset that has great nonlinearity, LDA based methods CCA-LDA and GMA perform poorly. Their recognition accuracy after multi-view learning is much lower than before. MvDA-VC and MvDA perform better with a higher target dimension, but when the target dimension decreases, the accuracy decreases quickly. This can be ascribed to their ignorance of inter-view and inner-view discriminant information. PLS tries to correlate the latent score of original space, as well as minimizing the variations of views in the common subspace. As a result, PLS gets a better and more stable performance, but PLS is difficult to extend to multi-view learning. By building global weighted graph and introducing the category discriminant information, the nonlinear method MvLE performs not only better, but also more stably when target dimension decreases. In case of 50 dimensions, the improvement of proposed method over MvDA-VC and MvDA is as much as 10.05 percent. But for face-emotion dataset that has poor nonlinearity, MvLE just has a weak advantage over other methods.

6. Conclusion

In this paper, we propose a new nonlinear supervised multi-view learning method named MvLE and its out-of-sample extension MHON to perform the RGB-D human emotion recognition. MvLE can map the training set of RGB-D data to a common subspace, and MHON is used to get the low-dimensional representations of test data. The new distance metric method BON can not only

overcome the difference between views, but also introduce the category discriminant information. Moreover, we introduced a video-emotion RGB-D dataset and a face-emotion RGB-D dataset to evaluate the proposed method. The experiment results indicate the effectiveness of our method in the two datasets. In the future, we can apply MvLE and MHON to other machine learning tasks.

Acknowledgment

Thanks for Bin Wang, Mingming Zhang, Bixuan Du, Keye Zhang, Shaohua Chen and Bin Zhan in the contribution of data acquisition of the video-emotion dataset.

References

[1] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, in: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 491–497.

[2] 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.

[3] 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.

[4] 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.

[5] A. Daros, K. Zakzanis, A. Ruocco, Facial emotion recognition in borderline personality disorder, Psychological Medicine 43 (9) (2013) 1953–1963.

[6] 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.

[7] S. Lalitha, A. Madhavan, B. Bhushan, S. Saketh, Speech emotion recognition, in: Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on, IEEE, 2014, pp. 1–4.

[8] 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.

[9] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al., Emonets: Multimodal deep learning approaches for emotion recognition in video, Journal on Multimodal User Interfaces 10 (2) (2016) 99–111.

[10] H. Kaya, F. Gürpınar, A. A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image and Vision Computing 65 (2017) 66–75.

[11] T. Bänziger, D. Grandjean, K. R. Scherer, Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)., Emotion 9 (5) (2009) 691.

[12] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7-8) (2013) 2031–2038.

[13] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7-8) (2013) 2031–2038.

[14] L. Zhang, D. Zhang, Visual understanding via multi-feature shared learning with global consistency, IEEE Transactions on Multimedia 18 (2) (2016) 247–259.

[15] L. Zhang, W. Zuo, D. Zhang, Lsdt: Latent sparse domain transfer learning for visual adaptation, IEEE Transactions on Image Processing 25 (3) (2016) 1177–1191.

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

[17] F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of machine learning research 3 (Jul) (2002) 1–48.

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

[19] K. Fukumizu, F. R. Bach, A. Gretton, Statistical consistency of kernel canonical correlation analysis, Journal of Machine Learning Research 8 (Feb) (2007) 361–383.

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

[21] 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.

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

[23] T.-K. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: European Conference on Computer Vision, Springer, 2006, pp. 251–262.

[24] T. Diethe, D. R. Hardoon, J. Shawe-Taylor, Multiview fisher discriminant analysis, in: NIPS workshop on learning from multiple sources, 2008.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, J. Jiang, Sparse unsupervised dimensionality reduction for multiple view data, IEEE Transactions on Circuits and Systems for Video Technology 22 (10) (2012) 1485.

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

[32] 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.

[33] 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.

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

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

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

[37] S. Liu, J. Wu, L. Feng, S. Luo, D. Yan, Quasi-curvature local linear projection and extreme learning machine for nonlinear dimensionality reduction, Neurocomputing 277 (2018) 208–217.

[38] L. Zhang, D. Zhang, Robust visual knowledge transfer via extreme learning machine based domain adaptation, IEEE Transactions on Image Processing 25 (10) (2016) 4959–4973.

[39] S. Liu, L. Feng, Y. Xiao, H. Wang, Robust activation function and its application: Semi-supervised kernel extreme learning method, Neurocomputing 144 (2014) 318–328.

[40] 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.

[41] B. De Gelder, Why bodies? twelve reasons for including bodily expressions in affective neuroscience, Philosophical Transactions of the Royal Society of London B: Biological Sciences 364 (1535) (2009) 3475–3484.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.