Knowledge-Based Systems 149 (2018) 34–46

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Local feature based multi-view discriminant analysis

Peng Hu, Dezhong Peng*, Jixiang Guo, Liangli Zhen

Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China

ARTICLE INFO

Article history:
Received 31 May 2017
Revised 29 December 2017
Accepted 3 February 2018
Available online 3 February 2018

Keywords:
Feature-based multi-view discriminant analysis
Local features
Common space
View-similarity constraint
Cross-view recognition
Heterogeneous recognition

ABSTRACT

In many real-world applications, an object can be represented from multiple views or styles. Thus, it is important to design algorithms that are able to recognize objects from distinct views. To the end, a large number of approaches have been proposed to achieve the heterogeneous recognition tasks through the use of local features. However, most of them only focus on binary views and thus cannot be applied to multi-view analysis. In this paper, we propose a novel local feature based multi-view discriminant analysis approach (FMDA). The proposed approach consists of three steps: First, the input images are represented using representation matrices and local feature descriptor (LFD) matrices of their overlapping patches, where the representation matrices are the linear coefficients of the LFDs for different views. In this way, it brings two advantages, i.e., addressing the small sample size (SSS) problem and preserving the discriminative information while reducing the redundant information in the LFD matrices. Second, the multi-view discriminant representation and feature projections are learned by projecting the LFDs of different views into a common space using the Fisher criterion. Finally, a simple but effective view-similarity constraint is proposed to adaptively learn the relationships between different views. To verify the effectiveness of the proposed method, extensive experiments are carried out on the FERET, CAS-PEAL-R1, CUFSF and HFB databases comparing with some state-of-the-art methods.

1. Introduction

In many computer vision applications, the same object can be observed at various viewpoints or even by heterogeneous sensors. For example, a person can be described by several facial images depicting different poses, expressions, lighting conditions and even heterogeneities, such as in [1–4]. Moreover, there are an increasing number of applications in which it is necessary to match images captured from various viewpoints or heterogeneous sensors; such a task is usually termed heterogeneous recognition or cross-view recognition [5–11]. However, such multi-view images may be best described in different spaces because of the large gaps between them. Therefore, traditional methods, in which all samples are regarded as being in the same space, may miss a great deal of information that could be used to achieve better performance.

To address the problem described above, numerous techniques for heterogeneous recognition have been proposed in recent decades [6–21]. Among these methods, [10,22–24] have achieved states of art in multi-view analysis under different settings. Although impressive results have been obtained some challenge issues have still remained. First, some works [7,12,14,25–28] have focused only on binary views and thus are not applicable to multi-view analysis. Alternatively, they may be adopted for multi-view analysis if the multi-view problem is transformed into the binary case, although this may be inefficient and does not fully utilize the relations among views. Second, some methods [8,13,16,17] treat multiple views as independent views to be mapped into a common latent space, without considering the similarities between the views, which may result in a loss of some useful discriminant information between views. Moreover, most of them do not use the local features of the images, which have been proven to serve as an effective basis for heterogeneous recognition by virtue of the excellent robustness and strong discriminant power of such features [6,9,18–21]. Thus, these multi-view methods cannot sufficiently and effectively utilize the information available in the data. Third, some of these methods [6,9,18–21] successfully employ local feature descriptors (LFDs, see Appendix A) to solve the heterogeneous recognition problem, but they can only address the binary view problem and do not consider the similarities among views. Furthermore, many types of local features have been widely used in many face recognition systems because of their excellent robustness and strong discriminant power [9,20,29–33], and each of these LFDs has its own characteristics and advantages compared with the others [34]. However, some of these approaches cannot be easily extended to other LFDs since they are designed based on

* Corresponding author.
E-mail address: pengdz@scu.edu.cn (D. Peng).

https://doi.org/10.1016/j.knosys.2018.02.008
0950-7051/© 2018 Elsevier B.V. All rights reserved.

specific LFDs, which may cause the performance of these methods to be limited on the LFDs.

In addition to the above-mentioned limitations, the small sample size (SSS) problem [35] is a well-known problem in subspace face recognition applications in general [36,37]. For many applications, such as face recognition, all scatter matrices in question may be singular because of the SSS problem [38], which is also known as the undersampling or singularity problem [38,39]. Many methods, such as local-feature-based discriminant analysis (LFDA) [6], Fisherface [40], random sampling linear discriminant analysis [36], two-dimensional principal component analysis [41], and two-dimensional linear discriminant analysis [38,42], have been proposed to address this problem. These methods can be classified into three main classes based on their approaches to solving the SSS problem. In methods of the first type, principal component analysis (PCA) is first used to preprocess the high-dimensional data into a low-dimensional feature space, and then, objective methods are applied in the low-dimensional PCA subspace [10,17,40,43]. However, the abandoned eigenfaces with small eigenvalues may also possess some discriminant information that would be useful for subsequent processing. In methods of the second type, multiple low-dimensional data sets are obtained by downsampling the high-dimensional data and then processed using objective methods to obtain several projections for each low-dimensional set [6,36]. However, many projections must be calculated in these methods, and not all data are directly involved in the projection computations; thus, they may not yield the best solutions. In methods of the third type, a two-dimensional trick is applied to address the SSS problem [38,41,42]. Every image is represented by a two-dimensional matrix instead of a one-dimensional vector. Then, all data are used to learn the most suitable space. However, if the dimensionality of the data is so high that the numbers of rows and columns are much greater than the sample size, then this two-dimensional optimization procedure is still subject to the SSS problem.

In this paper, we propose a novel method in which the LFDs of samples are used to perform multi-view discriminant analysis. First, each image is split into many overlapping patches, and then, the LFDs are computed from these patches to extract as much redundant discriminative information as possible and reduce the dimensionality of the data. To maximize the utilization of the discriminative information contained in these LFDs, a representation matrix is proposed to sample the LFDs in feature-based multi-view discriminant analysis (FMDA). Thus, two types of projections, the multi-view discriminant representation and the feature projections, should be calculated through multi-view discriminant analysis. Moreover, there are evidently certain latent relationships between the different views; however, it is difficult to formulate these relationships under complex conditions. We therefore propose a simple but effective view-similarity constraint to incorporate the characteristics of these relationships. With this constraint, the multi-view projections can be adapted to better consider the relative characteristics of different views. Fig. 1 shows the framework of the proposed method. Generally speaking, FMDA is a multi-view method that employs local features; FMDA is not implemented based on any specific LFD, and any LFD can be easily applied in FMDA, even when using simply the raw data of the patches. In addition, FMDA can avoid the SSS problem since the dimensionality of the LFDs of the patches can be substantially smaller than the sample size, and the sizes of the feature matrices are flexible because of the ability to use different LFDs as well as different patch sizes and step lengths. Overall, with the implementation of local feature representations and the view-similarity constraint, an improvement in accuracy is achieved compared with the tested/evaluated methods.

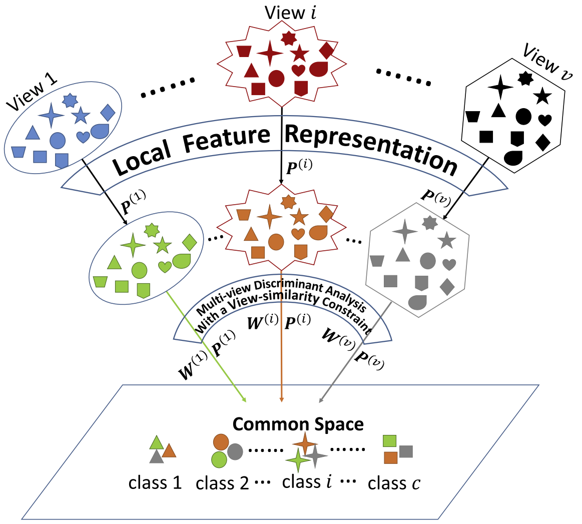

Fig. 1. The FMDA framework. The local features are first extracted from the samples and used to represent the corresponding images with representation matrices. These representative features extracted from the different views are then projected into a common discriminant space with a view-similarity constraint. In this figure, the different shapes denote different classes, and the different colors represent different distinct views.

The main contributions of this work can be summarized as follows:

- To learn more useful discriminative information for face recognition, we proposed a feature-based multi-view learning method based on the Fisher criterion.
- Different local regions of a facial image play different roles in recognition. To utilize such a difference, we proposed a method to learn representation for faces using the linear combination of the LFDs wherein the linear coefficients on regions reflect its importance. Thus, the discriminant of our method is further improved.
- A view-similarity constraint is proposed to incorporate the relationships between different views. With the help of this constraint, the multi-view projections can better reflect and incorporate the relationship of different views into our objective function.

The remainder of this paper is organized as follows. Section 2 introduces related works, Section 3 details the proposed method, Section 4 evaluates the proposed method on four different databases, and Section 5 concludes the paper.

Notations: **lower-case bold letters** represent column vectors and **upper-case bold letters** denote matrices. $\boldsymbol{A}^T$ denotes the transpose of the matrix $\boldsymbol{A}$. Table 1 summarizes some notations used throughout the paper.

2. Related works

2.1. Multi-view Canonical Correlation Analysis (MCCA)

In [16], the authors discussed generalizations of Canonical Correlation Analysis (CCA) [12] to analysis of multiple sets of variables, which is termed as the Multi-view Canonical Correlation Analysis (MCCA) or Multi-set Canonical Correlation Analysis. MCCA [16] attempts to find a set of linear transforms $\boldsymbol{w}^{(i)}|_{i=1}^{v}$ to respectively project the samples of v views $\{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(i)}, \cdots, \boldsymbol{X}^{(v)}\}$ to a common space such that the correlations among the projections of samples from all views are mutually maximized, where v is the number of views, $\boldsymbol{w}^{(i)}$ is the transform of ith view and $\boldsymbol{X}^{(i)} \in \mathbb{R}^{g_i \times n}$ is the data matrix of the ith view with n samples of g_i dimension.

Table 1
Some notations used throughout the paper.

Notation	Definition
v	The number of views
s	The size of the square patches
g	The dimension of the selected LFDs
q	The number of the patches for a giving image
δ	The step length of patch moving
c	The number of classes
d_1	The desired reduced dimensionality of $\boldsymbol{W}^{(k)}$
d_2	The desired reduced dimensionality of $\boldsymbol{P}^{(k)}$
n	The number of all samples from all views in all classes
$\boldsymbol{X}^{(i)}$	The data matrix of the ith view
$\boldsymbol{\Phi_{ij}}$	The patch of ith row and jth column for an image
$\boldsymbol{\varphi_{ij}}$	The local descriptor vector of $\boldsymbol{\Phi_{ij}}$
$\boldsymbol{\Gamma} = [\boldsymbol{\varphi}_{11}, \cdots, \boldsymbol{\varphi}_{MN}]$	A feature matrix for an image
$\boldsymbol{\Gamma}_{ij}^{(k)}$	The feature matrix of jth image in ith class from the kth view
$\boldsymbol{W}^{(k)} = \left[\boldsymbol{w}_1^{(k)}, \cdots, \boldsymbol{w}_{d_1}^{(k)}\right]$	The linear transform matrix of the kth view
$\boldsymbol{P}^{(k)} = \left[\boldsymbol{p}_1^{(k)}, \cdots, \boldsymbol{p}_{d_2}^{(k)}\right]$	The representation matrix of the kth view

The linear transforms can be obtained by maximizing the sum of correlations among each pair of views as follows:

$$\begin{aligned} &\max_{\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(v)}} \sum_{i<j}^{v} \boldsymbol{w}^{(i)^T} \boldsymbol{X}^{(i)} \boldsymbol{X}^{(j)^T} \boldsymbol{w}^{(j)} \\ &s.t. \quad \boldsymbol{w}^{(i)^T} \boldsymbol{X}^{(i)} \boldsymbol{X}^{(i)^T} \boldsymbol{w}^{(i)} = 1, i = 1, 2, \cdots, v. \end{aligned} \tag{1}$$

Using the Lagrange multiplier method, Eq. (1) can be easily solved as an eigenvalue problem [16]. Like CCA, the samples must have pairwise relationships among the all views and MCCA is also an unsupervised method.

2.2. Multi-view Discriminant Analysis (MvDA)

In [10] and [17], a Multi-view Discriminant Analysis (MvDA) approach is proposed to seek a single common discriminant space for multiple views in a non-pairwise manner by jointly learning multiple view-specific linear transforms. The MvDA approach attempts to find the multi-view transforms that project each view into a common space such that samples from the same class are as close to each other as possible, even when they are from different views, and samples from different classes are as far apart from each other as possible, even when they are from the same view. This is achieved as follows:

$$\max_{\boldsymbol{w}^{(1)},\cdots,\boldsymbol{w}^{(v)}} \frac{\boldsymbol{w}^T \boldsymbol{D} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w}}, \tag{2}$$

where $\boldsymbol{w} = [\boldsymbol{w}^{(1)^T}, \cdots, \boldsymbol{w}^{(i)^T}, \cdots, \boldsymbol{w}^{(v)^T}]^T (i = 1, 2, \cdots, v)$, $\boldsymbol{w}^{(i)}$ is the projection of the ith view, v is the number of views, and $\boldsymbol{D}$ and $\boldsymbol{S}$ are the between-class and within-class scatter matrices, respectively, for all views. Since observations from different views share similar data structures, a constraint is introduced in [10] to enforce the view consistency of the multiple linear transforms, resulting in a method called MvDA-VC, as follows:

$$\max_{\boldsymbol{w}^{(1)},\cdots,\boldsymbol{w}^{(v)}} \frac{\boldsymbol{w}^T \boldsymbol{D} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w} + \lambda \sum_{i,j=1}^{v} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2}, \tag{3}$$

where $\boldsymbol{\beta}_i = (\boldsymbol{X}^{(i)^T} \boldsymbol{X}^{(i)})^{-1} \boldsymbol{X}^{(i)^T} \boldsymbol{w}^{(i)}$ and λ is the balance parameter. Under this constraint, the view consistency among the multi-view transforms can be ensured. However, MvDA-VC requires that the number of samples from each view be the same, as in CCA, and the consistency among different views may not be maintained under complex conditions.

2.3. Local-Feature-based Discriminant Analysis (LFDA)

LFDA [6] attempts to address the problem of matching a forensic sketch to a gallery of mug shot images (heterogeneous face recognition). In the LFDA framework, the feature vector of each image is first divided into “slices” of smaller dimensionality, where these slices correspond to concatenations of the feature descriptor vectors from each column of the image patches. Next, discriminant analysis is performed separately on each slice through the following three steps: PCA, within-class whitening, and between-class discriminant analysis. Finally, PCA is applied to the new feature vector to remove redundant information among the feature slices to extract the final feature vector.

In LFDA, both sketches and photos are represented by SIFT feature descriptors and multi-scale local binary patterns (MLBPs). Multiple discriminant projections are then used on the partitioned vectors of the feature-based representation for minimum distance matching. LFDs are successfully used to improve heterogeneous face recognition. However, LFDA can only be applied in the two-view case. Moreover, similar to CCA, the samples must have pairwise relationships between the two views for LFDA.

As summarized above, most multi-view methods do not consider the discriminative information contained in local features, as in [8,10,13,16], or can only address binary-view problems, as in [6,18–20]. Moreover, minimal work has been conducted on multi-view discriminant representation and feature projection learning for multi-view problems. The most closely related work is the coupled discriminant face descriptor (C-DFD) approach [19], in which many image filters and soft sampling matrices are learned to extract features for all non-overlapping regions in an image for each view. However, in the C-DFD method, numerous projections must be calculated for each view, and only binary-view problems can be addressed. Moreover, view-similarity characteristics are not considered in the above methods.

3. Feature-based multi-view discriminant analysis

Image feature descriptors are widely used to represent the distinct characteristics of an image or image region [34]. LFDs have been successfully used in many face recognition applications by virtue of their excellent robustness and strong discriminative power [20]. In the following, we will present how to use LFDs and representation matrices as features to represent a corresponding sample. Then, multi-view discriminant analysis is applied to project the features corresponding to different views into a common space with a view-similarity constraint, as shown in Fig. 1.

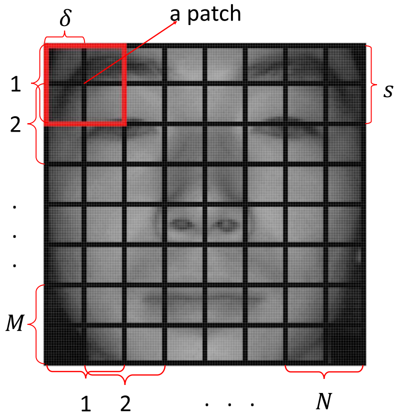

Fig. 2. This figure shows how to slice an $H \times W$ image into an $M \times N$ grid. To simplify the description, we set $\delta = \frac{s}{2}$, as shown in this graph. Each patch is selected by sliding a window (the red square area) over the image with a step length of δ, such that each patch overlaps with its vertical or horizontal neighbors by $s - \delta$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

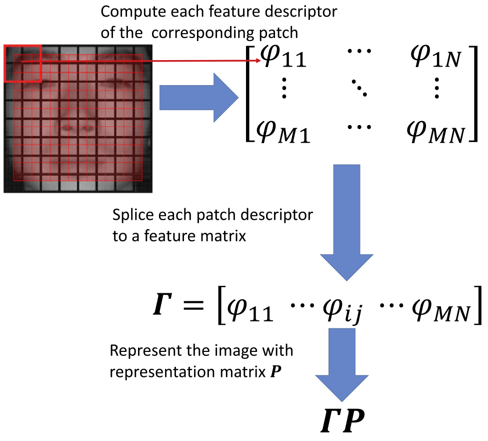
Compute each feature descriptor of the corresponding patch
$\begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1N} \\ \vdots & \ddots & \vdots \\ \varphi_{M1} & \cdots & \varphi_{MN} \end{bmatrix}$
Splice each patch descriptor to a feature matrix
$\Gamma = [\varphi_{11} \cdots \varphi_{ij} \cdots \varphi_{MN}]$
Represent the image with representation matrix P
ΓP

Fig. 3. This figure shows how to represent an image using a feature matrix $\mathbf{\Gamma}$ and a representation matrix $\boldsymbol{P}$. The location of each patch feature vector $\boldsymbol{\varphi}_{ij}$ is the spatial location of the corresponding patch $\mathbf{\Phi}_{ij}$ as defined by the red $M \times N$ grid superimposed on the image. These feature vectors $\boldsymbol{\varphi}_{11}$, $\boldsymbol{\varphi}_{12}$, $\cdots$, $\boldsymbol{\varphi}_{MN}$ are used to linearly represent the image as $\mathbf{\Gamma}\boldsymbol{P}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Local-feature-based representation

In this section, we will introduce how to represent images using their local features and the corresponding representation matrix. First, an image should be segmented into several overlapping regions (or patches), chosen based on two parameters: a patch size s and a step length δ. Every $s \times s$ square patch defined with a δ step is used to compute a local feature vector. More simply, this is similar to slide an $s \times s$ square window across the image with a δ step to extract the feature descriptor of each corresponding window. Obviously, in this way, an image can be split into a grid composed of $M \times N$ overlapping squares, as shown in Fig. 2, where M and N are the total numbers of vertical and horizontal patches, respectively. It is very easy to obtain $M = \lfloor \frac{V-s}{\delta} \rfloor + 1$ and $N = \lfloor \frac{H-s}{\delta} \rfloor + 1$ for a $V \times H$ ($V \geq s$, $H \geq s$) image. Then, we can denote any patch in the image by $\mathbf{\Phi}_{\boldsymbol{ij}}$, where i and j are the row and column numbers, respectively. Moreover, we denote the feature descriptor of $\mathbf{\Phi}_{\boldsymbol{ij}}$ by $\boldsymbol{\varphi}_{ij} = F(\mathbf{\Phi}_{\boldsymbol{ij}})$, where $F(\cdot)$ represents the feature extraction method used. Thus, all $M \times N$ feature vectors can be constructed to represent the corresponding image as shown in Fig. 3. Finally, the feature vectors of all patches can be combined into a feature matrix representing the image features, which is denoted by

$$\mathbf{\Gamma} = \left[\boldsymbol{\varphi}_{11}, \cdots, \boldsymbol{\varphi}_{ij}, \cdots, \boldsymbol{\varphi}_{MN}\right],$$

as shown in Fig. 3. Here, $\mathbf{\Gamma} \in \mathbb{R}^{g \times q}$; $i = 1, \cdots, M$; $j = 1, \cdots, N$; g is the dimensionality of the feature descriptor; and $q = M \times N$. It is clear that the feature matrix includes a substantial amount of redundant information since the columns were obtained from overlapping patches, and each image is linearly represented in terms of its patches. Thus, to preserve as much discriminative information as possible while eliminating redundant information from redundant features, each image can be linearly represented in terms of these local feature vectors as $\mathbf{\Gamma}\boldsymbol{P}$, where $\boldsymbol{P}$ is the representation matrix. In the following, we will show how to use these features to perform multi-view discriminant analysis and to learn multi-view discriminant feature and representation projections.

3.2. Local-feature-based multi-view discriminant analysis

In this section, we will show how to project the features obtained from v views using the method described in the previous section into a common discriminant space by means of their linear feature transforms $\boldsymbol{W}^{(1)}$, $\boldsymbol{W}^{(2)}$, $\cdots$, $\boldsymbol{W}^{(v)}$ and the respective representation transforms $\boldsymbol{P}^{(1)}$, $\boldsymbol{P}^{(2)}$, $\cdots$, $\boldsymbol{P}^{(v)}$. These projections are obtained via the Fisher criterion, which maximizes the between-class variation while minimizing the within-class variation. First, we define $\boldsymbol{X}_{ij}^{(k)}$ as the jth ($j = 1, \cdots, n_i^{(k)}$) sample in the ith ($i = 1, \cdots, c$) class from the kth ($k = 1, \cdots, v$) view, where $n_i^{(k)}$ is the number of samples in the ith class from the kth view and c is the number of classes. The local feature matrix of $\boldsymbol{X}_{ij}^{(k)}$ can be denoted by $\mathbf{\Gamma}_{ij}^{(k)}$; then, the image can be represented as $\boldsymbol{X}_{ij}^{(k)}\boldsymbol{P}^{(k)}$, where $\boldsymbol{P}^{(k)}$ is the representation matrix for the kth view.

To project the samples from the v views into a common latent space using the Fisher criterion, we denote the projected result of $\mathbf{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)}$ by $\boldsymbol{Y}_{ij}^{(k)} = \boldsymbol{W}^{(k)^T}\mathbf{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)}$, where $\boldsymbol{W}^{(k)}$ is the feature projection matrix for the kth view. In the common space, the between-class scatter should be maximized, whereas the within-class scatter should be minimized, as shown in Fig. 1. We can obtain the objective function as follows based on the Fisher criterion:

$$\left[\boldsymbol{W}^{(1)}, \cdots, \boldsymbol{W}^{(v)}; \boldsymbol{P}^{(1)}, \cdots, \boldsymbol{P}^{(v)}\right] = \arg \max_{\substack{\boldsymbol{W}^{(1)}, \cdots, \boldsymbol{W}^{(v)} \\ \boldsymbol{P}^{(1)}, \cdots, \boldsymbol{P}^{(v)}}} \frac{\mathrm{Tr}(\boldsymbol{S}_b)}{\mathrm{Tr}(\boldsymbol{S}_w)}, \tag{4}$$

where $\mathrm{Tr}(\cdot)$ is the trace operator and $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are the between-class and within-class matrices, respectively. Moreover, the between-class matrix is

$$\begin{aligned} \boldsymbol{S}_b &= \sum_{i=1}^{c} n_i |\boldsymbol{M}_i - \boldsymbol{M}|^2 \\ &= \sum_{i=1}^{c} n_i |\boldsymbol{M}_i|^2 - n|\boldsymbol{M}|^2, \end{aligned} \tag{5}$$

and the within-class matrix is

$$\begin{aligned} \boldsymbol{S}_w &= \sum_{i=1}^{c} \sum_{k=1}^{v} \sum_{j=1}^{n_i^{(k)}} |\boldsymbol{Y}_{ij}^{(k)} - \boldsymbol{M}_i|^2 \\ &= \sum_{k=1}^{v} \sum_{i=1}^{c} \sum_{j=1}^{n_i^{(k)}} |\boldsymbol{Y}_{ij}^{k}|^2 - \sum_{i=1}^{c} n_i |\boldsymbol{M}_i|^2, \end{aligned} \tag{6}$$

where $n_i = \sum_k^v n_i^{(k)}$ is the total number of samples in the ith class, $\boldsymbol{M}_i = \frac{1}{n_i} \sum_{k=1}^{v} \sum_{j=1}^{n_i^{(k)}} \boldsymbol{Y}_{ij}^{(k)}$ is the mean of the samples of the ith class, $\boldsymbol{M} = \frac{1}{n} \sum_{k=1}^{v} \sum_{i=1}^{c} \sum_{j=1}^{n_i^{(k)}} \boldsymbol{Y}_{ij}^{(k)}$ is the mean of all samples over all views and all classes, and n is the number of all samples from all views in all classes.

Because $\boldsymbol{Y}_{ij}^{(k)} = \boldsymbol{W}^{(k)^T}\boldsymbol{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)}$, it is clear that

$$\begin{aligned} \boldsymbol{M}_i &= \tfrac{1}{n_i}\sum_{k=1}^{v}\sum_{j=1}^{n_i^{(k)}}\boldsymbol{W}^{(k)^T}\boldsymbol{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)} \\ &= \tfrac{1}{n_i}\sum_{k=1}^{v}\boldsymbol{W}^{(k)^T}\left(\sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)}\right)\boldsymbol{P}^{(k)} \end{aligned} \tag{7}$$

and

$$\begin{aligned} \boldsymbol{M} &= \tfrac{1}{n}\sum_{k=1}^{v}\sum_{i=1}^{c}\sum_{j=1}^{n_i^{(k)}}\boldsymbol{Y}_{ij}^{(k)} \\ &= \tfrac{1}{n}\sum_{k=1}^{v}\boldsymbol{W}^{(k)^T}\left(\sum_{i=1}^{c}\sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)}\right)\boldsymbol{P}^{(k)}. \end{aligned} \tag{8}$$

We solve the optimization problem defined in Eq. (4) in an iterative manner. In each iteration, either the $\boldsymbol{W}$ variables or the $\boldsymbol{P}$ variables are fixed; the values of the other variables are then computed by optimizing Eq. (4). First, we fix the $\boldsymbol{P}$ to compute the $\boldsymbol{W}$. Thus, Eq. (5) and Eq. (6) can be rewritten as

$$\begin{aligned} \boldsymbol{S}_b &= \sum_{i=1}^{c} n_i|\boldsymbol{M}_i|^2 - n|\boldsymbol{M}|^2 \\ &= \sum_{k,l}^{v}\boldsymbol{W}^{(k)^T}\Bigg(\sum_{i}^{c}\tfrac{1}{n_i}\boldsymbol{S}_i^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}_i^{(l)^T} \\ &\quad -\tfrac{1}{n}\boldsymbol{S}^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}^{(l)^T}\Bigg)\boldsymbol{W}^{(l)} \\ &= \boldsymbol{W}^T\boldsymbol{B}\boldsymbol{W} \end{aligned} \tag{9}$$

and

$$\begin{aligned} \boldsymbol{S}_w &= \sum_{k=1}^{v}\sum_{i=1}^{c}\sum_{j=1}^{n_i^{(k)}}|\boldsymbol{Y}_{ij}^{k}|^2 - \sum_{i=1}^{c} n_i|\boldsymbol{M}_i|^2 \\ &= \sum_{k,l}^{v}\boldsymbol{W}^{(k)}\Bigg(\sum_{i}^{c}\Bigg((k==l)\sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{\Gamma}_{ij}^{(k)^T} \\ &\quad -\tfrac{1}{n_i}\boldsymbol{S}_i^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}_i^{(l)^T}\Bigg)\Bigg)\boldsymbol{W}^{(l)} \\ &= \boldsymbol{W}^T\boldsymbol{C}\boldsymbol{W}, \end{aligned} \tag{10}$$

where $\boldsymbol{S}_i^{(k)} = \sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)}$ is the sum of all samples in the ith class from the kth view; $\boldsymbol{S}^{(k)} = \sum_{i=1}^{c}\boldsymbol{S}_i^{(k)}$ is the sum of all samples from the kth view; $\boldsymbol{W} = [\boldsymbol{W}^{(1)^T}, \boldsymbol{W}^{(2)^T}, \cdots, \boldsymbol{W}^{(v)^T}]^T$ is constructed from the transform matrices for all views; $k == l$ is a Boolean equation, whose value is 1 if $k = l$ and 0 otherwise; and $\boldsymbol{B}$ and $\boldsymbol{C}$ are block matrices with the following forms:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & \cdots & \boldsymbol{B}_{1v} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} & \cdots & \boldsymbol{B}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{B}_{v1} & \boldsymbol{B}_{v2} & \cdots & \boldsymbol{B}_{vv} \end{bmatrix} \tag{11}$$

and

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1v} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{v1} & \boldsymbol{C}_{v2} & \cdots & \boldsymbol{C}_{vv} \end{bmatrix}. \tag{12}$$

Moreover, from Eqs. (9) and (10), each matrix block of $\boldsymbol{B}$ and $\boldsymbol{C}$ should satisfy

$$\boldsymbol{B}_{kl} = \sum_{i=1}^{c}\tfrac{1}{n_i}\boldsymbol{S}_i^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}_i^{(l)^T} - \tfrac{1}{n}\boldsymbol{S}^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}^{(l)^T} \tag{13}$$

and

$$\begin{aligned} \boldsymbol{C}_{kl} &= \sum_{i=1}^{c}\Bigg((k==l)\sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{\Gamma}_{ij}^{(k)^T} \\ &\quad -\tfrac{1}{n_i}\boldsymbol{S}_i^{(k)}\boldsymbol{P}^{(k)}\boldsymbol{P}^{(k)^T}\boldsymbol{S}_i^{(l)^T}\Bigg). \end{aligned} \tag{14}$$

Therefore, under the Fisher criterion, to achieve the goal of maximizing the between-class value and minimizing the within-class value, it is easy to obtain the following objective function:

$$\begin{aligned} \boldsymbol{W}_{opt} &= \arg\max_{\boldsymbol{W}}\tfrac{\mathrm{Tr}(\boldsymbol{S}_b)}{\mathrm{Tr}(\boldsymbol{S}_w)} \\ &= \arg\max_{\boldsymbol{W}}\tfrac{\mathrm{Tr}(\boldsymbol{W}^T\boldsymbol{B}\boldsymbol{W})}{\mathrm{Tr}(\boldsymbol{W}^T\boldsymbol{C}\boldsymbol{W})}. \end{aligned} \tag{15}$$

Since Eq. (15) cannot be solved analytically, ones could relax it as the following tractable determinant ratio problem according to [44]:

$$\boldsymbol{W}_{opt} = \arg\max_{\boldsymbol{W}}\frac{|\boldsymbol{W}^T\boldsymbol{B}\boldsymbol{W}|}{|\boldsymbol{W}^T\boldsymbol{C}\boldsymbol{W}|}, \tag{16}$$

which can be equivalently solved using the following generalized eigenvalue decomposition (GED) problem [10,17,40,44]:

$$\boldsymbol{B}\boldsymbol{w}_k = \tau_k\boldsymbol{C}\boldsymbol{w}_k, \tag{17}$$

where τ_k $(k = 1, 2, \cdots, d_1)$ is the kth largest eigenvalue of the GED with the corresponding eigenvector $\boldsymbol{w}_k$, $\boldsymbol{w}_k$ constitutes the kth column vector of the matrix $\boldsymbol{W}$, and d_1 is the desired dimensionality. Thus, we obtain the desired $\boldsymbol{W}$ by fixing the $\boldsymbol{P}$. Similar to $\boldsymbol{W}$, we can define

$$\boldsymbol{P} = \left[\boldsymbol{P}^{(1)^T}, \cdots, \boldsymbol{P}^{(k)^T}, \cdots, \boldsymbol{P}^{(v)^T}\right]^T,$$

where $\boldsymbol{P}^{(k)}$ is the representation matrix for the kth view. Because $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^T) = \mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A})$, we can easily obtain $\boldsymbol{P}$ in a manner similar to $\boldsymbol{W}$. Then, the columns of $\boldsymbol{P}$ are the eigenvectors corresponding to the top d_2 eigenvalues of the formulation presented below, which is derived through a process similar to that shown for the case of $\boldsymbol{W}$, where d_2 is the desired dimensionality:

$$\boldsymbol{D}\boldsymbol{p}_k = \tau_k\boldsymbol{E}\boldsymbol{p}_k, \tag{18}$$

where τ_k $(k = 1, 2, \cdots, d_2)$ is the kth largest eigenvalue of the GED with the corresponding eigenvector $\boldsymbol{p}_k$, $\boldsymbol{p}_k$ constitutes the kth column vector of the matrix $\boldsymbol{P}$, and $\boldsymbol{D}$ and $\boldsymbol{E}$ are also block matrices, similar to $\boldsymbol{B}$ and $\boldsymbol{C}$. Moreover, each block of these matrices should satisfy the following:

$$\boldsymbol{D}_{kl} = \sum_{i=1}^{c}\frac{1}{n_i}\boldsymbol{S}_i^{(k)^T}\boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)^T}\boldsymbol{S}_i^{(l)} - \frac{1}{n}\boldsymbol{S}^{(k)^T}\boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)^T}\boldsymbol{S}^{(l)} \tag{19}$$

and

$$\begin{aligned} \boldsymbol{E}_{kl} &= \sum_{i=1}^{c}\Bigg((k==l)\sum_{j=1}^{n_i^{(k)}}\boldsymbol{\Gamma}_{ij}^{(k)^T}\boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)^T}\boldsymbol{\Gamma}_{ij}^{(k)} \\ &\quad -\tfrac{1}{n_i}\boldsymbol{S}_i^{(k)^T}\boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)^T}\boldsymbol{S}_i^{(l)}\Bigg). \end{aligned} \tag{20}$$

To summarize, the complete FMDA algorithm is illustrated in Algorithm 1. LDA [40] can be used to further reduce the dimensionality of the common space obtained through FMDA.

3.3. FMDA with a view-similarity constraint (FMDAvs)

Intuitively, certain latent relationships will exist between different views because the multi-view samples depict the same subject. However, it is difficult to formulate these relationships under complex conditions such as differing poses, illumination conditions, and expressions. Therefore, it might not be easy to convert from a view to another, as in [10]. To simplify the representation,

Algorithm 1 Feature-based Multi-view Discriminant Analysis (FMDA).

```
Input: A set of abstracted local feature matrices, $\{\boldsymbol{\Gamma}_{ij}^{(k)}, i = 1, \cdots, c,\ k = 1, \cdots, v,\ j = 1, \cdots, n_i^{(k)}\}$, where $\boldsymbol{\Gamma}_{ij}^{(k)} \in \mathbb{R}^{g \times q}$, $g$ is the dimensionality of the LFDs, and $q$ is the number of patches used to compute the local features for each image. The desired reduced dimensionalities $d_1$ and $d_2$ and the number of iterations $T$ are set in advance.
Initialize: $\boldsymbol{W} = \boldsymbol{I}$ and $\boldsymbol{P} = \boldsymbol{I}$.
for $t = 1, 2, \cdots, T$ do
    Compute the between- and within-class block scatter matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ using Eqs. (13) and (14) by fixing $\boldsymbol{P}$.
    Solve the generalized eigenvalue problem defined in Eq. (17) and obtain the eigenvectors $\boldsymbol{W}'$ with the $d_1$ largest eigenvalues.
    $\boldsymbol{W} \leftarrow \boldsymbol{W}'$
    Calculate the between- and with-class block scatter matrices $\boldsymbol{D}$ and $\boldsymbol{E}$ using Eqs. (19) and (20) by fixing $\boldsymbol{W}$.
    Solve the generalized eigenvalue problem defined in Eq. (18) and obtain the eigenvectors $\boldsymbol{P}'$ with the $d_2$ largest eigenvalues.
    $\boldsymbol{P} \leftarrow \boldsymbol{P}'$
end for
Output: The projections $\boldsymbol{W}^{(k)} \in \mathbb{R}^{g \times d_1}$ and $\boldsymbol{P}^{(k)} \in \mathbb{R}^{q \times d_2}$, where $d_1$ and $d_2$ are the desired reduced dimensionalities.
```

we use the similarities between different views to describe their relationships. Moreover, for simplicity, the similarities between different views can be described in terms of the Euclidean distances between the corresponding projections. Therefore, we first define the distance between the projections of the kth and lth views as

$$dis(\boldsymbol{W}^{(k)}, \boldsymbol{W}^{(l)}) = \| \boldsymbol{W}^{(k)} - \boldsymbol{W}^{(l)} \|_F^2, \tag{21}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\boldsymbol{W}^{(k)}$ and $\boldsymbol{W}^{(l)}$ are the projection matrices for the kth and lth views, respectively. Then, we can obtain the total distances between all pairs of views from $\boldsymbol{W}$ as follows:

$$\begin{aligned} s_1 &= \sum_{k=1}^{v} \sum_{l=k+1}^{v} dis(\boldsymbol{W}^{(k)}, \boldsymbol{W}^{(l)}) \\ &= (v-1) \sum_{k}^{v} \mathrm{Tr}\Big(\boldsymbol{W}^{(k)^T} \boldsymbol{W}^{(k)}\Big) - \sum_{k \neq l}^{v} \mathrm{Tr}\Big(\boldsymbol{W}^{(k)^T} \boldsymbol{W}^{(l)}\Big) \\ &= \mathrm{Tr}\Big(\boldsymbol{W}^T \boldsymbol{\Phi} \boldsymbol{W}\Big), \end{aligned} \tag{22}$$

where $\boldsymbol{\Phi}$ is a block matrix, as in Eqs. (11) and (12):

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} & \cdots & \boldsymbol{\Phi}_{1v} \\ \boldsymbol{\Phi}_{21} & \boldsymbol{\Phi}_{22} & \cdots & \boldsymbol{\Phi}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Phi}_{v1} & \boldsymbol{\Phi}_{v2} & \cdots & \boldsymbol{\Phi}_{vv} \end{bmatrix}. \tag{23}$$

Thus, it is very easy to obtain the klth block of $\boldsymbol{\Phi}$ as follows:

$$\boldsymbol{\Phi}_{kl} = \begin{cases} (v-1)\boldsymbol{I}, & k = l \\ -\boldsymbol{I}, & \text{otherwise}, \end{cases} \tag{24}$$

where $\boldsymbol{I}$ is the identity matrix. Similarly, it is very easy to obtain the total distances between all views for $\boldsymbol{P}$ as follows:

$$s_2 = \mathrm{Tr}\Big(\boldsymbol{P}^T \boldsymbol{\Phi} \boldsymbol{P}\Big). \tag{25}$$

Clearly, these distances can be used to formulate a constraint to increase or decrease the deviations between the transforms. In this way, the transform matrices can be made more consistent with the relationships between the views; we call this consistency the view similarity. Thus, under the view-similarity constraint, the objective function given in Eq. (4) can be easily rewritten as

$$\begin{aligned} &[\boldsymbol{W}_{opt}; \boldsymbol{P}_{opt}] = \arg\max_{\boldsymbol{W};\ \boldsymbol{P}} \tfrac{\mathrm{Tr}(\boldsymbol{S}_b)}{\mathrm{Tr}(\boldsymbol{S}_w)+\kappa} \\ &s.t.\ \kappa = \sum_{k}^{2} \lambda_k s_k + \sum_{k=1}^{v} \eta_k \|\boldsymbol{W}^{(k)}\|_F^2 + \sum_{k=1}^{v} \gamma_k \|\boldsymbol{P}^{(k)}\|_F^2, \end{aligned} \tag{26}$$

where λ_1, λ_2, $\eta_1, \cdots, \eta_v$, and $\gamma_1, \cdots, \gamma_v$ are real numbers and $\|\boldsymbol{W}^{(k)}\|_F^2$ and $\|\boldsymbol{P}^{(k)}\|_F^2$ are the shrinkage constraints, also known as the Tikhonov regularizers [45,46], which help to improve the generalizability of the solutions. The LDA model with Tikhonov regularization is usually referred to as Regularized Discriminant Analysis [47,48]. Moreover, λ_1 and λ_2 control the balance of the view similarity. When their values are positive, the transform matrices for the different views will be closer together. When their values are negative, these matrices will be farther apart. Clearly, a value of 0 will not work. Therefore, Eqs. (17) and (18) can be easily rewritten as

$$\boldsymbol{B}\boldsymbol{w}_k = \tau_k(\boldsymbol{C} + \lambda_1 \boldsymbol{\Phi} + \boldsymbol{\Psi})\boldsymbol{w}_k \tag{27}$$

and

$$\boldsymbol{D}\boldsymbol{p}_k = \tau_k(\boldsymbol{E} + \lambda_2 \boldsymbol{\Phi} + \boldsymbol{\Upsilon})\boldsymbol{p}_k, \tag{28}$$

where $\boldsymbol{\Psi}$ and $\boldsymbol{Y}$ are block matrices, as in Eqs. (11), (12) and (23):

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} & \cdots & \boldsymbol{\Psi}_{1v} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} & \cdots & \boldsymbol{\Psi}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Psi}_{v1} & \boldsymbol{\Psi}_{v2} & \cdots & \boldsymbol{\Psi}_{vv} \end{bmatrix} \tag{29}$$

and

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Upsilon}_{11} & \boldsymbol{\Upsilon}_{12} & \cdots & \boldsymbol{\Upsilon}_{1v} \\ \boldsymbol{\Upsilon}_{21} & \boldsymbol{\Upsilon}_{22} & \cdots & \boldsymbol{\Upsilon}_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Upsilon}_{v1} & \boldsymbol{\Upsilon}_{v2} & \cdots & \boldsymbol{\Upsilon}_{vv} \end{bmatrix}. \tag{30}$$

Moreover, from Eq. (26), it is easy to obtain the klth blocks of $\boldsymbol{\Psi}$ and $\boldsymbol{Y}$ as follows:

$$\boldsymbol{\Psi}_{kl} = \begin{cases} \eta_k \boldsymbol{I}, & k = l \\ \boldsymbol{0}, & \text{otherwise} \end{cases} \tag{31}$$

and

$$\boldsymbol{\Upsilon}_{kl} = \begin{cases} \gamma_k \boldsymbol{I}, & k = l \\ \boldsymbol{0}, & \text{otherwise.} \end{cases} \tag{32}$$

Therefore, $\boldsymbol{W}$ and $\boldsymbol{P}$ can be obtained by solving the eigenvalue problems defined in Eqs. (27) and (28), respectively, by fixing the other set of variables, as in Algorithm 1.

3.4. Time complexity analysis

Suppose the computational cost of LFD method $F(\cdot)$ is denoted by $O(\rho(s))$. In the training stage, our method takes $O(\rho(s)nq)$ to extract LFDs from n training samples, and $O\big(cn + T\big(gq(cv^2 + n)(d_1 + d_2 + g + q) + v^3g^3 + v^3q^3\big)\big)$ to compute the projections $\boldsymbol{W}$ and $\boldsymbol{P}$ for T iterations, where $O\big(v^3g^3\big)$ is the time complexity of Eqs. (17) and (27), $O\big(v^3q^3\big)$ is the cost of Eqs. (18) and (28), and the left $O\big(cn + Tgq(cv^2 + n)(d_1 + d_2 + g + q)\big)$ is the computational cost of $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Upsilon}$ for T iterations. Therefore, the complexity for training is $O(\rho(s)nq + cn + T\big(gq(cv^2 + n)(d_1 + d_2 + g + q) + v^3g^3 + v^3q^3\big)\big)$ in terms of T iterations. Similarly, for the feature extraction, it takes $O(\rho(s)q + d_1q(g + d_2))$ for each image, where $O(\rho(s)q)$ is the time cost for the LFD extraction regarding to a single image and $O(d_1q(g + d_2))$ is used to obtain the projection.

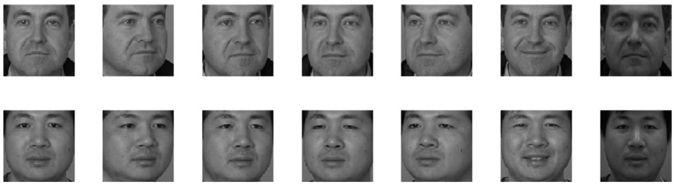

Fig. 4. The first two subjects in the subset of the FERET database used in this experiment. From left to right, the columns show the images of the 0°, 25°, 15°, −15°, and −25° poses, followed by the smiling and dim images.

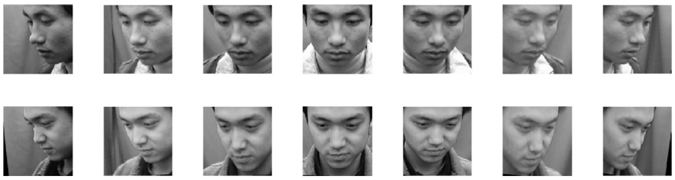

Fig. 5. The first two subjects in the subset of the CAS-PEAL-R1 PD database used in this experiment. From left to right, the yaw values of the images in each column are −45°, −30°, −15°, 0°, 15°, 30°, and 45°.

4. Experiments

In this section, FMDA is evaluated on four different datasets for four different heterogeneous face recognition tasks, i.e., face recognition across poses, illumination conditions and expressions; face recognition across poses; photo vs. sketch face recognition; and visual (VIS) vs. near-infrared (NIR) face recognition.

4.1. Datasets

1. The *Facial Recognition Technology Database (FERET)* [1] is used to evaluate the performance of the proposed method for face recognition across different poses, illumination conditions and expressions. The subset used in this experiment is simply a subset of the FERET database. This dataset includes images of 200 subjects, each in 5 poses (−25°, −15°, 0° (neutral), 15°, and 25°), with 2 expressions (neutral and smiling for the 0° pose) and under 2 illumination conditions (neutral and dim for the 0° pose). This subset was further divided into two parts to serve as the training set and the test set. For tuning the parameters of all methods, the first 140 subjects (140×7 images) were selected as the training set, and the remaining 60 subjects (60×7 images) were used as the test set. Then the mean and standard deviations of the 10 times running results were reported to evaluate the performance of the methods by fixing the tuned parameters. For each time, 140 subjects were randomly selected as the training set, and the remaining subjects were used as the test set. In addition, all images were cropped to 80×80 pixels, as shown in Fig. 4.
2. The *CAS-PEAL-R1 Database* [2] is applied to evaluate the performance of the proposed method across different poses. CAS-PEAL-R1 contains 21,832 face images of 1040 individuals across 21 poses [49]. All images were captured under ambient illumination conditions and with the subjects showing neutral expressions. Only one image exists for each subject in each pose. The 21 poses are sampled from a pose space with 7 discrete yaw values and 3 discrete pitch values. For this experiment, we chose a subset of the database in which the yaw values range from −45° to +45° and the pitch value is approximately +30°, namely, the PD set, which contains 7×939 images of 939 individuals. This subset was further divided into two parts to serve as the training set and the test set. For tuning the parameters of all methods, the first 500 subjects (500×7 images) were selected as the training set, and the remaining 439 subjects (439×7 images) were used as the test set. Then the mean and standard deviations of the 10 times running results were reported to evaluate the performance of the methods by fixing the tuned parameters. For each time, 500 subjects were randomly selected as the training set, and the remaining subjects were used as the test set. In addition, all images were aligned and cropped to 80×80 pixels according to the provided eye coordinates. Fig. 5 shows cropped face examples from the selected subset.

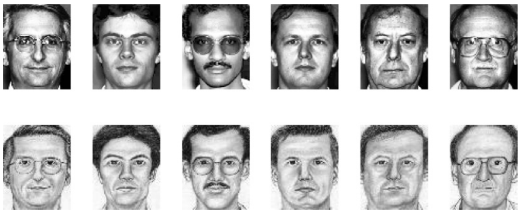

Fig. 6. The photos and sketches of six subjects in the CUFSF database. The first row shows the photos of the subjects, and the second row shows the corresponding sketches.

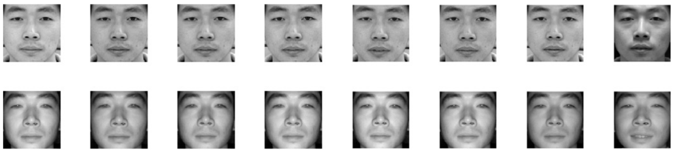

Fig. 7. The VIS and NIR face images of a subject in the HFB database. The first row shows the 8 VIS images of the subject, and the second row shows the 8 NIR images of the subject.

3. The *CUHK Face Sketch FERET (CUFSF) Database* [3,5] is employed to evaluate the performance of the proposed method for photo vs. sketch face recognition. CUFSF includes images corresponding to 1194 persons from the FERET database [1]. For each person, the database includes a face photo with lighting variations and a sketch with shape exaggeration drawn by an artist while viewing that photo. For tuning the parameters of all methods, the first 500 subjects were selected for training, and the remaining 694 subjects were used for testing. Then the mean and standard deviations of the 10 times running results were reported to evaluate the performance of the methods by fixing the tuned parameters. For each time, 500 subjects were randomly selected as the training set, and the remaining subjects were used as the test set. Moreover, all images were cropped to 80×64 pixels, as shown in Fig. 6.
4. The *Heterogeneous Face Biometrics (HFB)* [4] database is used to evaluate the performance of FMDA for VIS vs. NIR heterogeneous face recognition. This database consists of images of 200 subjects, each corresponding to 8 VIS and 8 NIR face images. Each face image was aligned and cropped to 128×128 pixels according to the position of the eyes as provided by the authors. Fig. 7 shows several cropped example images from the HFB dataset. For tuning the parameters of all methods, the first 100 subjects were chosen as the training data, and the remaining 100 subjects were used as the test set. Then the mean and standard deviations of the 10 times running results were reported to evaluate the performance of the methods by fixing the tuned parameters. For each time, 100 subjects were randomly selected as the training set, and the remaining subjects were used as the test set.

4.2. Experimental setting

All images from FERET were cropped to 80 × 80 pixels, as shown in Fig. 4; all images from CAS-PEAL-R1 were aligned and cropped to 80 × 80 pixels using a standard protocol, as shown in Fig. 5; all images from CUFSF were cropped to 80 × 64 pixels using a standard protocol, as shown in Fig. 6; and all samples from HFB were cropped to 128 × 128 pixels as specified by the authors of [4], as shown in Fig. 7. The proposed method was then evaluated by comparison with the most closely related existing algorithms: PCA [50], Multiset CCA (MCCA) [16], GMA [8], PLS [7], LDA [40], SIFT [33], SIFT+LDA [9], HOG [31], HOG+LDA [9], C-DFD [19], MvDA [17], MvDA-VC [10], CDFL [21], LRDE [11], GSS-SL [28], and C-DFD [19]. The MATLAB codes for GMA,[1] PLS,[1] MvDA,[2] MvDA-VC[2] and GSS-SL[3] as implemented by the original authors were downloaded from the Internet. The SIFT and HOG algorithms were obtained from [51]. The MCCA and LDA algorithms were implemented by ourselves. For a comprehensive comparison, we use the raw data, SIFT and HOG in experiments. The corresponding dimensions are s^2, 128 and 124, where s denotes size of the square local region.

In our experiments, we empirically set s as 16 or 15, δ as 4 or 5, and $T = 3$. For raw data, d_1 was determined to give the best result from 5 to 60 with an interval of 5. For HOG and SIFT, the d_1 was chosen from 50 to 100 with an interval of 10. Moreover, the value ranges of d_2, λ_1, λ_2, η_k and γ_k are [20, 200] with an interval of 10, $[10^{-3}, 3]$, $[10^{-3}, 3]$, $[10^{-5}, 0.3]$ and $[10^{-5}, 0.3]$, respectively. To reduce the effort for tuning parameters, we use the tuned d_1 and d_2 of FMDA for FMDAvs and seek an optimal combination of λ_1, λ_2, η_k and γ_k by

$$a_i = \begin{cases} a_{i-1} + 10^{\lfloor \lg(a_{i-1}) \rfloor} & (a_{i-1} < end \quad \text{and} \quad i = 1, 2, \cdots) \\ start & i = 0 \end{cases},$$

where a_i indicates the value of ith possible choice for the parameter $a = \{\lambda_1, \lambda_2, \eta_k, \gamma_k\}$, $[start, end]$ denotes the value range of the corresponding parameter, $\lg(\cdot)$ is the logarithm operator to base 10 and $\lfloor \cdot \rfloor$ is the rounded down operator. When the number of views is large (on the FERET and PEAL data sets), it is a daunting task to tune all η_k and γ_k $(k = 1, 2, \cdots, v)$. Thus, we simply use the same value for η_k, as well as γ_k.

To reduce the dimensionality of the data and avoid the SSS problem, PCA [50] was first applied to reduce the dimensionality of all of the data to obtain the best performances of the MCCA, GMA, PLS, LDA, MvDA and MvDA-VC methods. Then, the reduced training data were used to obtain the transform matrices produced by each of the above methods. The obtained projections were then applied to the reduced test data to obtain the final test set, which was used to compute the best accuracies across all dimensionalities in terms of the rank-1 recognition rate. The rank-1 testing was conducted in a pairwise manner; i.e., the images from one view were used as the gallery, and the images from another view were used as the probe. The means of all of the pairwise accuracies were then used to compare the performances of all methods on the corresponding dataset. It is notable that GMA and GSS-SL only can be used on HFB and fail to work on FERET, CAS-PEAL-R1 and CUFSF, since there is only one sample per class per view in these dataset.

4.3. Parameter analysis

In this section, we investigate the influence of parameters on the performance of our method on the HFB data set with SIFT. For each investigation, we change the value of one parameter and fix the others. Fig. 8a and b show the recognition rate of FMDAvs versus different values of λ_1 and λ_2. We can see that the recognition rate is significantly improved after considering the view-similarity constraint. Fig. 8c–f show the influence of η_1, η_2, γ_1 and γ_2, respectively. From the results, we can see that these shrinkage constraints give slight improvements in the recognition rate. However, they have much smaller influence on the results than the view-similarity does. Therefore, when the number of views is large, all η_k and γ_k $(k = 1, 2, \cdots, v)$ could be set to be the same as did in following experiments.

4.4. Face recognition across poses, illumination conditions and expressions

The face recognition performance across different poses, illumination conditions and expressions (PIE) was evaluated on the FERET database by considering each pose, illumination condition or expression as one view for the calculation of the pairwise accuracies. The samples in FERET correspond to seven different views, thus leading to $7 \times 6 = 42$ evaluations of the rank-1 recognition rate. All 42 results were averaged to obtain the mean accuracy, as shown in Fig. 9. In the figure, methods in boldface are significantly better than the others, according to the t-test with a significance level at 0.05. From the results, we can see that SIFT+LDA and HOG+LDA achieve better performances compared with traditional LDA by virtue of the incorporation of the local features. However, when the view differences are disregarded, their performances are worse than those of the multi-view methods. It is clear that the proposed method achieves better PIE performance compared with the other methods on the FERET database. The proposed method achieves improvements over MvDA-VC by 6.68% , over MCCA by 7.84% and over HOG+LDA by 15.27% , indicating that our method is a good multi-view learning method for cross-view recognition. Moreover, we also compare the performances achieved by our method when using different local features. As seen, the performances with SIFT and HOG are better than those achieved when using the raw patch data since these LFDs may contain more discriminative information than the raw data. However, our method can also efficiently extract more discriminative information from the raw patch data than can the other methods. Even when using raw data, our method achieves improvements over MvDA-VC by 5.48%, over MCCA by 6.64% and over HOG+LDA by 14.07% . Furthermore, when the view similarity is considered (denoted by FMDAvs), the performances of FMDA and FMDA+LDA are further improved, as shown in this figure.

To enable more detailed performance comparisons, some of the cross-view recognition results are shown in Table 6; these results are the rank-1 recognition accuracies achieved when the 0° view is used as the gallery and the remaining 6 views are used as the probes. As seen, our method performs the best, achieving significant improvements of up to 6% for some views, e.g., 25°. Therefore, our method is more effective than the other methods for the PIE recognition task.

4.5. Face recognition across poses

Face recognition across different poses was evaluated on the CAS-PEAL-R1 database by considering each pose as one view for the calculation of the pairwise accuracies. Similar to FERET, the samples from CAS-PEAL-R1 correspond to seven views, thus leading to $7 \times 6 = 42$ evaluations of the rank-1 recognition rate. All 42 results were averaged to obtain the mean accuracy, as shown in Fig. 10. In the figure, methods in boldface are significantly better than the others, according to the t-test with a significance level at 0.05. From the results, we can see that SIFT+LDA and

[1] The code is available at: https://www.cs.umd.edu/~bhokaal/Research.htm.

[2] The code is available at: http://sourcedb.ict.cas.cn/cn/jssrck/201511/t20151119_4470344.html.

[3] The code is available at: http://people.ucas.edu.cn/~bpma.

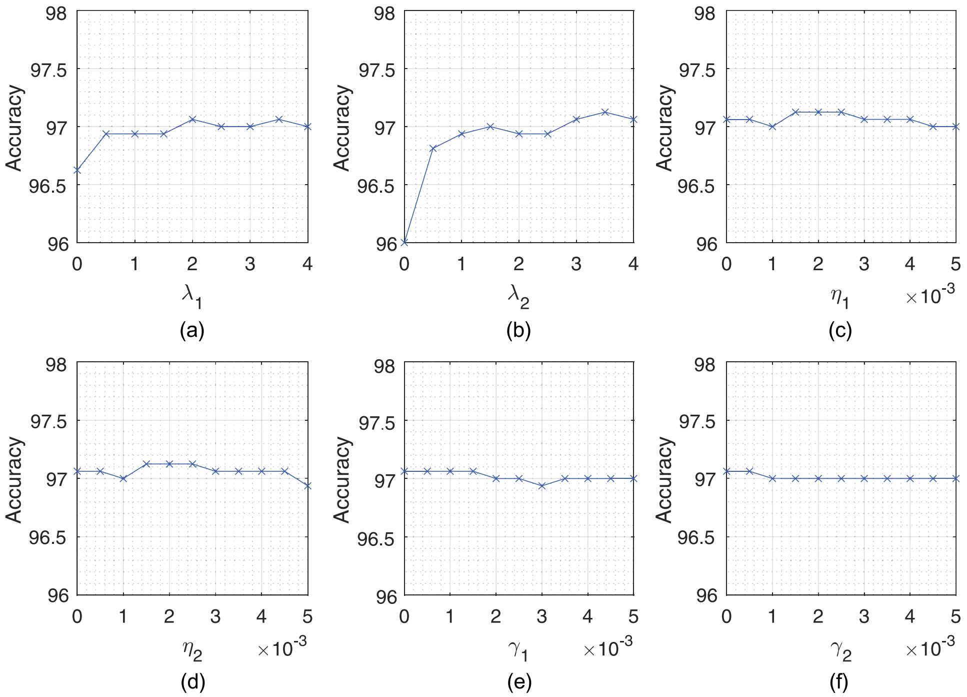

Fig. 8. Rank-one recognition rate of FMDAvs on HFB versus different values of λ_1, λ_1, η_1, η_2, γ_1 and γ_2, respectively.

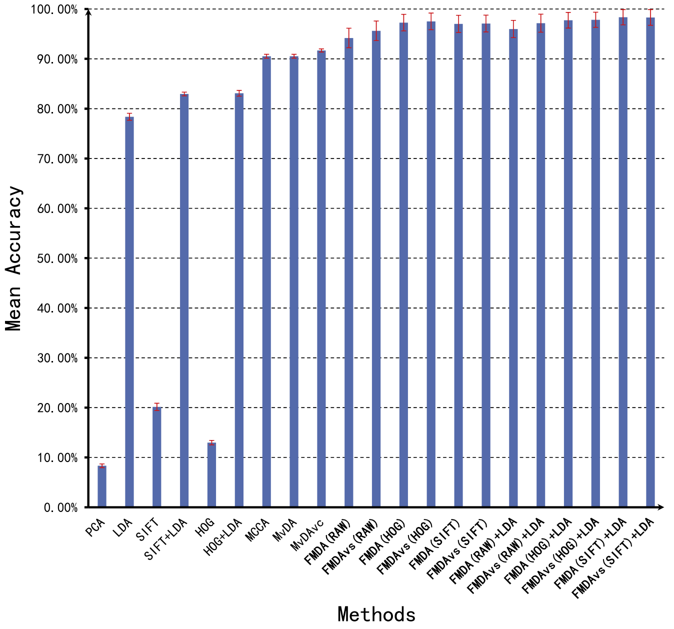

Fig. 9. Evaluation of the PIE recognition performance on the FERET dataset in terms of the mean accuracy.

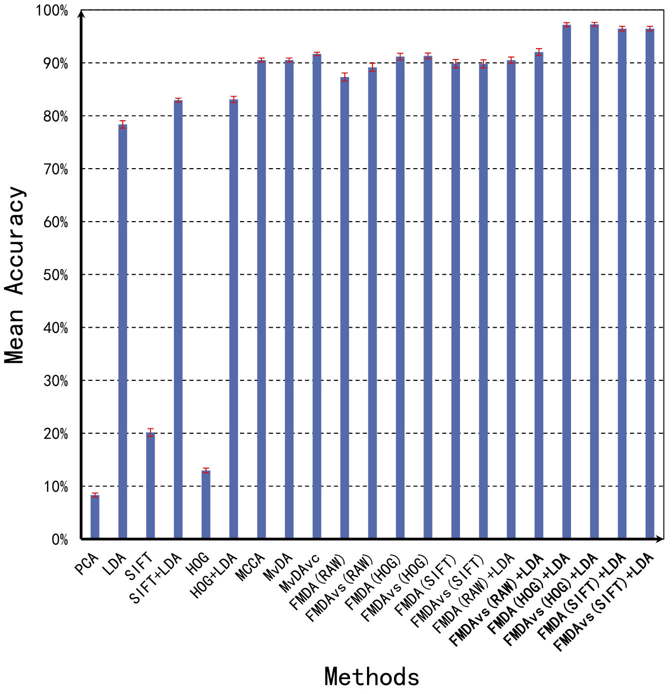

Fig. 10. Comparison of the best mean recognition accuracy rates on the CAS-PEAL-R1 dataset for the cross-pose face recognition task.

HOG+LDA achieve better performances compared with traditional LDA by virtue of the incorporation of the local features. However, when the view differences are disregarded, their performances are worse than those of the multi-view methods. It is clear that the proposed method achieves better performance across poses compared with the other methods on the CAS-PEAL-R1 database. The proposed method achieves improvements over MvDA-VC by 5.60%, over MCCA by 6.76% and over HOG + LDA by 14.19%, indicating that our method is a good multi-view learning method for cross-view recognition. Moreover, we also compare the performances achieved by our method using different local features. As seen, the performances with SIFT and HOG are better than those achieved using the raw patch data since these LFDs may contain more discriminative information than the raw data. However, our method can also efficiently extract more discriminative information from the raw patch data than can the other methods. Even with the raw data, our method achieves improvements over MvDA-VC by 0.37%, over MCCA by 1.53% and over HOG + LDA by 8.96% . When the relationships between views are disregarded, the performance of FMDA with the raw data is worse than that of MvDA-VC, but

Table 2
Evaluation of the PIE recognition performance on the FERET dataset (with the 0° view as the training set). Results in boldface are significantly better than the other methods, according to the *t*-test with a significance level at 0.05.

Method	25°	15°	−15°	−25°	smile	dim	Average
PCA [50]	27.83 ± 4.65	53.33 ± 4.97	61.50 ± 3.96	29.67 ± 5.37	81.67 ± 5.09	72.83 ± 4.23	54.47 ± 2.48
LDA [40]	88.67 ± 4.57	91.83 ± 3.37	93.83 ± 4.23	91.00 ± 3.26	95.83 ± 3.36	89.00 ± 5.34	91.69 ± 2.78
SIFT [33]	34.83 ± 4.93	60.00 ± 4.84	71.83 ± 4.74	36.67 ± 6.43	86.17 ± 4.01	88.67 ± 3.91	63.03 ± 1.29
SIFT+LDA [9]	90.67 ± 4.53	95.50 ± 2.94	94.67 ± 3.22	92.83 ± 4.31	96.50 ± 2.77	95.83 ± 3.36	94.33 ± 2.23
HOG [31]	38.17 ± 8.48	78.67 ± 5.08	87.50 ± 5.17	51.00 ± 5.40	88.17 ± 2.54	88.33 ± 2.94	71.97 ± 2.24
HOG+LDA [9]	87.83 ± 5.78	95.33 ± 3.12	96.83 ± 2.14	89.50 ± 3.34	98.67 ± 1.89	94.67 ± 3.22	93.81 ± 2.65
MCCA [16]	92.50 ± 2.97	93.67 ± 3.58	94.00 ± 4.17	92.67 ± 2.74	91.50 ± 3.55	86.33 ± 4.43	91.78 ± 2.62
MvDA [17]	92.50 ± 2.97	93.67 ± 3.58	94.00 ± 4.17	92.67 ± 2.74	91.50 ± 3.55	86.33 ± 4.43	91.78 ± 2.62
MvDA-VC [10]	92.17 ± 3.43	93.33 ± 3.51	94.00 ± 3.44	92.33 ± 3.44	91.67 ± 2.83	87.67 ± 4.25	91.86 ± 2.75
FMDA (RAW)	**94.67 ± 2.92**	96.67 ± 2.94	96.50 ± 2.77	**96.83 ± 2.54**	96.17 ± 2.36	96.17 ± 3.34	**96.17 ± 2.53**
FMDAvs (RAW)	**96.83 ± 2.42**	**97.00 ± 2.58**	97.17 ± 2.49	**97.00 ± 2.70**	96.67 ± 2.61	96.33 ± 3.75	**96.83 ± 2.60**
FMDA (HOG)	**96.33 ± 2.46**	**97.00 ± 2.58**	98.00 ± 1.72	**97.17 ± 2.23**	98.33 ± 1.92	**97.8 ± 2.09**	**97.44 ± 2.00**
FMDAvs (HOG)	**96.83 ± 2.42**	**97.17 ± 2.49**	98.00 ± 1.72	**97.50 ± 2.12**	98.50 ± 2.00	97.67 ± 2.38	**97.61 ± 2.08**
FMDA (SIFT)	**96.17 ± 2.73**	96.67 ± 2.61	98.00 ± 1.72	**97.33 ± 2.11**	98.00 ± 1.72	97.33 ± 2.38	**97.25 ± 2.06**
FMDAvs (SIFT)	**96.33 ± 2.46**	**96.83 ± 2.66**	98.00 ± 1.72	**97.33 ± 2.11**	98.00 ± 1.72	97.33 ± 2.51	**97.31 ± 2.04**
FMDA (RAW)+LDA	**96.83 ± 2.14**	**97.33 ± 2.11**	98.17 ± 1.83	**97.17 ± 2.73**	97.67 ± 2.11	96.67 ± 2.36	**97.31 ± 1.91**
FMDAvs (RAW)+LDA	**97.50 ± 1.80**	**97.33 ± 2.25**	**98.67 ± 1.31**	**97.17 ± 2.49**	98.67 ± 2.05	**98.17 ± 2.00**	**97.92 ± 1.82**
FMDA (HOG)+LDA	**98.17 ± 2.00**	**98.17 ± 2.00**	98.17 ± 2.28	**97.67 ± 1.79**	98.17 ± 2.28	**97.50 ± 1.80**	**97.97 ± 1.72**
FMDAvs (HOG)+LDA	**98.00 ± 1.53**	**98.00 ± 1.53**	98.17 ± 2.28	**97.33 ± 2.25**	98.00 ± 1.89	**97.83 ± 1.77**	**97.89 ± 1.62**
FMDA (SIFT)+LDA	**98.50 ± 1.66**	**98.00 ± 2.05**	**98.50 ± 1.66**	**97.83 ± 1.77**	98.50 ± 1.66	**98.50 ± 1.83**	**98.31 ± 1.58**
FMDAvs (SIFT)+LDA	**98.50 ± 1.66**	**98.17 ± 2.14**	98.00 ± 1.72	**97.67 ± 2.11**	98.67 ± 1.72	**98.67 ± 1.72**	**98.28 ± 1.71**

Table 3
Evaluation of the face recognition performance across different poses on the CAS-PEAL-R1 dataset (with the 0° view as the training set) Results in boldface are significantly better than the other methods, according to the *t*-test with a significance level at 0.05.

Method	−45°	−30°	−15°	15°	30°	45°	Average
PCA [50]	2.26 ± 0.47	5.26 ± 0.69	19.84 ± 2.45	20.21 ± 1.37	3.64 ± 0.56	2.44 ± 0.46	8.94 ± 0.71
LDA [40]	57.45 ± 1.93	88.66 ± 1.71	99.04 ± 0.37	98.95 ± 0.42	92.69 ± 0.93	75.03 ± 2.42	85.30 ± 0.67
SIFT [33]	10.73 ± 1.90	25.76 ± 2.32	49.48 ± 3.39	50.36 ± 2.17	23.26 ± 2.36	11.21 ± 1.33	28.47 ± 1.15
SIFT+LDA [9]	65.17 ± 1.82	93.30 ± 0.85	99.66 ± 0.25	99.45 ± 0.19	96.47 ± 0.64	82.94 ± 1.59	89.50 ± 0.41
HOG [31]	2.44 ± 0.54	6.31 ± 1.23	29.20 ± 4.02	30.64 ± 1.97	6.15 ± 0.96	2.78 ± 0.58	12.92 ± 0.87
HOG+LDA [9]	61.91 ± 2.96	92.89 ± 0.90	99.86 ± 0.16	99.98 ± 0.07	97.49 ± 0.58	82.64 ± 1.62	89.13 ± 0.50
MCCA [16]	89.32 ± 1.04	97.40 ± 0.66	99.86 ± 0.12	99.25 ± 0.47	97.18 ± 0.65	86.47 ± 0.68	94.91 ± 0.24
MvDA [17]	89.32 ± 1.04	97.40 ± 0.66	99.86 ± 0.12	99.25 ± 0.47	97.18 ± 0.65	86.47 ± 0.68	94.91 ± 0.24
MvDA-VC [10]	90.30 ± 1.35	97.52 ± 0.54	99.86 ± 0.16	99.23 ± 0.31	97.59 ± 0.66	88.15 ± 0.70	95.44 ± 0.29
FMDA (RAW)	86.45 ± 1.00	94.92 ± 0.78	99.48 ± 0.19	98.61 ± 0.29	94.08 ± 1.16	82.64 ± 1.06	92.70 ± 0.47
FMDAvs (RAW)	88.27 ± 1.27	95.65 ± 1.03	99.64 ± 0.27	98.84 ± 0.33	94.49 ± 0.98	85.60 ± 1.11	93.75 ± 0.59
FMDA (HOG)	**91.46 ± 1.11**	96.90 ± 0.74	99.75 ± 0.13	99.04 ± 0.32	96.72 ± 0.70	88.95 ± 0.89	95.47 ± 0.41
FMDAvs (HOG)	**91.44 ± 0.97**	96.95 ± 0.82	99.79 ± 0.07	99.07 ± 0.33	96.90 ± 0.72	**89.38 ± 0.78**	95.59 ± 0.36
FMDA (SIFT)	88.63 ± 1.63	96.04 ± 0.77	99.38 ± 0.39	98.91 ± 0.4	95.63 ± 0.88	87.29 ± 1.35	94.31 ± 0.63
FMDAvs (SIFT)	88.59 ± 1.49	95.97 ± 0.73	99.41 ± 0.42	98.91 ± 0.41	95.56 ± 0.91	87.22 ± 1.40	94.27 ± 0.60
FMDA (RAW)+LDA	89.75 ± 0.79	96.90 ± 0.43	99.79 ± 0.23	99.32 ± 0.26	96.54 ± 0.91	86.08 ± 1.14	94.73 ± 0.28
FMDAvs (RAW)+LDA	**91.57 ± 0.84**	97.15 ± 0.34	99.79 ± 0.17	99.29 ± 0.36	97.24 ± 0.56	88.29 ± 0.97	95.56 ± 0.24
FMDA (HOG)+LDA	**97.27 ± 0.67**	**99.54 ± 0.26**	**100.00 ± 0.00**	99.95 ± 0.10	**99.64 ± 0.24**	**96.17 ± 0.64**	**98.76 ± 0.22**
FMDAvs (HOG)+LDA	**97.27 ± 0.50**	**99.59 ± 0.26**	**100.00 ± 0.00**	99.95 ± 0.10	**99.64 ± 0.34**	**96.45 ± 0.65**	**98.82 ± 0.18**
FMDA (SIFT)+LDA	**96.10 ± 0.86**	**98.70 ± 0.42**	99.95 ± 0.10	99.84 ± 0.11	**99.27 ± 0.44**	**95.44 ± 0.94**	**98.22 ± 0.37**
FMDAvs (SIFT)+LDA	**96.01 ± 0.80**	**98.70 ± 0.42**	99.95 ± 0.10	99.84 ± 0.11	**99.27 ± 0.50**	**95.38 ± 1.05**	**98.19 ± 0.38**

the performance of FMDAvs+LDA is better than that of MvDA-VC. Furthermore, when the view similarity is considered, the performances of FMDA and FMDA+LDA are improved, as shown in this figure. Therefore, the view-similarity constraint is a powerful tool for adapting the relationships between views.

To enable more detailed performance comparisons, some of the cross-view recognition results are shown in Table 3; these results are the rank-1 recognition accuracies achieved when the 0° view is used as the gallery and the remaining 6 views are used as the probes. As seen, our method performs the best, achieving significant improvements of up to 8% for certain views, e.g., 45°. Therefore, our method is more effective than the other methods for the cross-pose recognition task. (Table 2)

4.6. Photo vs. sketch face recognition

Face recognition between photos and sketches was evaluated on the CUFSF database. The photos and sketches were considered as different views for the computation of the pairwise accuracies. The results are shown in Fig. 11 and Table 4. To present the recognition accuracies for different dimensionalities on the tuning set in this figure, for the FMDA and FMDAvs methods, the dimensions of their results were first further reduced to 100 by PCA, and LDA was applied to the reduced data for comparison with the other methods. For the other methods, the dimensions of the raw data or local features were first reduced to 100 by PCA, and these methods were then applied to the reduced data. From the results, we can see that the proposed method achieves better performance in photo vs. sketch face recognition compared with the other methods. The proposed method achieves improvements over LRDE by 20.37, MvDA by 25.48%, over MCCA by 29.09%, over SIFT+LDA by 14.19% and over CDFL by 4.7%, indicating that our method is a good multi-view learning method for photo vs. sketch recognition. With the view-similarity constraint, FMDAvs and FMDAvs+LDA achieve better performances than those of FMDA and FMDA+LDA in terms of the recognition rate, as shown in Fig. 11 and Table 4. From the results, we can see that the LFDs provide more discriminative in-

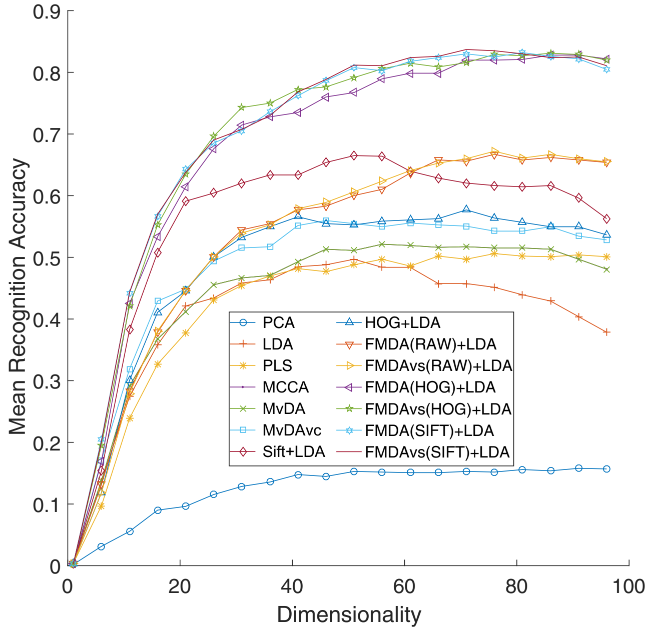

Fig. 11. Photo vs. sketch face recognition. This figure shows the mean accuracies of the different methods for various reduced dimensionalities in increments of 5.

Table 4
Photo vs. sketch face recognition rates. Results in boldface are significantly better than the other methods, according to the *t*-test with a significance level at 0.05.

Method	Photo-Sketch	Sketch-Photo	Average
PCA [50]	15.52 ± 1.27	16.46 ± 1.32	15.99 ± 1.09
LDA [40]	50.10 ± 1.56	54.41 ± 1.39	52.26 ± 1.36
SIFT [33]	23.01 ± 1.83	33.47 ± 1.05	28.24 ± 0.92
SIFT+LDA [9]	70.09 ± 1.70	74.35 ± 1.00	72.22 ± 1.09
HOG [31]	27.10 ± 1.34	39.39 ± 1.79	33.25 ± 1.31
HOG+LDA [9]	55.09 ± 1.78	57.16 ± 1.42	56.12 ± 1.50
PLS [7]	52.15 ± 1.80	59.02 ± 1.89	55.58 ± 1.72
MCCA [16]	58.46 ± 1.66	56.18 ± 2.21	57.32 ± 1.66
MvDA [17]	59.87 ± 1.07	61.99 ± 1.40	60.93 ± 1.04
MvDA-VC [10]	59.09 ± 2.15	62.54 ± 1.42	60.81 ± 1.55
CDFL [21][a]	81.3	–	–
LRDE [11][a]	65.94	66.13	66.04
FMDA (RAW)	61.04 ± 0.74	66.17 ± 1.04	63.60 ± 0.77
FMDAvs (RAW)	61.21 ± 1.36	66.71 ± 1.60	63.96 ± 1.35
FMDA (HOG)	74.83 ± 1.72	**79.18 ± 1.91**	**77.00 ± 1.75**
FMDAvs (HOG)	79.39 ± 1.53	**82.25 ± 1.58**	**80.82 ± 1.46**
FMDA (SIFT)	**82.25 ± 1.40**	**83.78 ± 1.37**	**83.01 ± 1.13**
FMDAvs (SIFT)	**82.32 ± 1.49**	**83.78 ± 1.39**	**83.05 ± 1.19**
FMDA (RAW)+LDA	64.35 ± 1.41	66.33 ± 1.02	65.34 ± 0.95
FMDAvs (RAW)+LDA	65.65 ± 0.72	66.70 ± 1.10	66.17 ± 0.87
FMDA (HOG)+LDA	**84.18 ± 1.47**	**85.07 ± 0.77**	**84.63 ± 1.01**
FMDAvs (HOG)+LDA	**85.27 ± 1.05**	**86.14 ± 1.34**	**85.71 ± 1.06**
FMDA (SIFT)+LDA	**85.84 ± 1.08**	**86.82 ± 1.35**	**86.33 ± 1.02**
FMDAvs (SIFT)+LDA	**85.99 ± 0.90**	**86.83 ± 1.22**	**86.41 ± 0.89**

[a] The results are reported in the original papers.

formation than do the raw data for photo vs. sketch face recognition. When LFDs are used, LDA and CDFL can achieve good performances. Moreover, even when using only the raw patch data, our method achieves improvements over LRDE by 0.13%, MvDA by 5.24% and over MCCA by 8.85% . Therefore, our method can extract more useful information for face recognition from either these LFDs or the raw data than can the other methods.

4.7. Visual vs. near-Infrared face recognition

The performance of FMDA for the VIS vs. NIR face recognition task was evaluated on the HFB database. Difference of Gaussians (DoG)-based preprocessing [52] was applied to each face image to reduce the effects of illumination variations, as in [19,20]. As in the photo vs. sketch face recognition experiment, the VIS and NIR images were considered as different views for the calculation of the pairwise accuracies, which are presented in Table 5. From the results, we can see that the proposed method achieves better performance than the other methods for VIS vs. NIR face recognition. The proposed method achieves improvements over C-DFD by 4.22%, over MvDA-VC by 45.32, over SIFT+LDA by 10.97% and over HOG+LDA by 8.97%, indicating that our method is a good multi-view learning method for VIS vs. NIR recognition. Moreover, even when using only the raw patch data, our method achieves improvements over C-DFD by 0.08%, over MvDA-VC by 40.83%, over MvDA by 41.30%, over SIFT+LDA by 6.48% and over HOG+LDA by 4.48% . Therefore, our method can extract more useful information for face recognition from either these LFDs or the raw data than can the other methods. We can see that some discriminative information may be lost when LDA is applied following FMDA or FMDAvs; however, LDA can reduce the higher-dimensional results of FMDA and FMDAvs while retaining most of the discriminative information. In addition, the view-similarity constraint leads to higher VIS vs. NIR face recognition performances of FMDAvs and FMDAvs+LDA compared with FMDA and FMDA+LDA regardless of whether the raw data or the SIFT or HOG features are chosen.

Table 5
VIS vs. NIR face recognition rates. Results in boldface are significantly better than the other methods, according to the *t*-test with a significance level at 0.05.

Method	NIR-VIS	VIS-NIR	Average
PCA [50]	6.14 ± 1.32	8.35 ± 1.18	7.24 ± 1.06
LDA [40]	58.19 ± 3.76	61.59 ± 4.41	59.89 ± 3.80
SIFT [33]	79.64 ± 1.56	61.71 ± 6.42	70.67 ± 3.43
SIFT+LDA [9]	83.49 ± 2.05	85.71 ± 1.32	84.60 ± 1.55
HOG [31]	53.40 ± 5.85	51.13 ± 5.86	52.26 ± 4.96
HOG+LDA [9]	85.88 ± 2.23	87.33 ± 1.94	86.60 ± 1.81
PLS [7]	31.60 ± 2.27	34.57 ± 2.83	33.09 ± 2.39
MCCA [16]	44.65 ± 5.00	46.70 ± 5.92	45.67 ± 5.34
GMA [8]	41.64 ± 4.87	43.93 ± 5.75	42.78 ± 4.68
MvDA [17]	49.92 ± 5.03	49.64 ± 6.20	49.78 ± 5.48
MvDA-VC [10]	49.17 ± 6.89	51.32 ± 5.19	50.25 ± 5.89
GSS-SL [28]	35.41 ± 4.16	39.46 ± 5.74	37.44 ± 4.71
C-DFD [19][a]	–	92.20	–
FMDA (RAW)	**87.72 ± 1.46**	89.76 ± 2.40	**88.74 ± 1.59**
FMDAvs (RAW)	**89.89 ± 1.12**	**92.28 ± 1.73**	**91.08 ± 1.17**
FMDA (HOG)	**92.24 ± 2.27**	**93.97 ± 1.35**	**93.11 ± 1.66**
FMDAvs (HOG)	**94.01 ± 1.65**	**95.94 ± 1.64**	**94.97 ± 1.49**
FMDA (SIFT)	**93.85 ± 1.21**	**95.41 ± 1.39**	**94.63 ± 1.01**
FMDAvs (SIFT)	**94.71 ± 0.91**	**96.42 ± 1.25**	**95.57 ± 0.78**
FMDA (RAW)+LDA	76.61 ± 2.86	81.65 ± 3.98	79.13 ± 3.30
FMDAvs (RAW)+LDA	82.81 ± 1.11	86.79 ± 3.16	84.80 ± 2.01
FMDA (HOG)+LDA	**89.64 ± 2.47**	**92.41 ± 2.43**	**91.03 ± 2.28**
FMDAvs (HOG)+LDA	**92.28 ± 2.31**	**94.34 ± 1.39**	**93.31 ± 1.66**
FMDA (SIFT)+LDA	**92.60 ± 1.31**	**95.76 ± 0.85**	**94.18 ± 0.83**
FMDAvs (SIFT)+LDA	**94.39 ± 1.32**	**96.36 ± 1.25**	**95.38 ± 1.02**

[a] The results are reported in the original papers.

4.8. Computational cost

In this section, we compared the computational efficiency of different multi-view methods on the HFB dataset. All the computational time is calculated on a PC with 3.2 GHz i5-3470 CPU and 12GB RAM in MATLAB. Table 6 reports the computational time cost of FMDAvs with SIFT, MCCA [16], MvDA-VC [11], C-CBFD [19] and C-DFD [18]. For each method, we repeat it 10 time and compute the average time for training and feature extraction. From the result, we can see that FMDAvs is a more efficient than C-CBFD and C-DFD in terms of the training and feature extraction time. The reason is that our method does not need to compute so many pro-

Table 6
Time cost of different multi-view methods.

Time (s)	MCCA [16]	MvDA-VC [10]	C-CBFD [20]	C-DFD [19]	FMDAvs
training	9.82	9.97	9996.58	5304.45	346.69
feature extraction	0.0013	0.0013	0.22	3.32	0.028

jections and cluster centers like them, thus leading to lower time cost.

5. Conclusion

In this paper, we propose a novel multi-view analysis method. To achieve flexible local feature abstraction, we propose a local feature representation approach in which an image is represented by the features of its patches and the representation matrix for the corresponding view. To find the common discriminant latent feature space, we propose a feature-based multi-view analysis method for calculating the multi-view discriminative representation and feature projections in that space. To better consider the relationships between different views, we propose a view-similarity constraint that makes the multi-view projections more consistent with the characteristics of the relationships between the views. Using FMDA, high-dimensional multi-view data can be effectively projected into a common space of lower dimensionality, as demonstrated by our experiments. From the experimental results, we can see that our method exhibits superiority over other methods on four heterogeneous face recognition tasks. Moreover, recent advances in deep learning have shown that deep networks can typically exploit more useful information to achieve better performance [20,53–56]. In the future, we plan to investigate how our method might benefit from deep learning or other feature learning methods. Moreover, we also hope to investigate our method's performance in a semi-supervised scenario.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under grants U1435213 and 61172180, and Chengdu International Cooperation Project under grants 2016-GH02-00048-HZ and 2015-GH02-00041-HZ.

Appendix A

LFDs compute interest regions as features, which have shown effective in various tasks including object recognition and texture recognition [18,34]. The popular LFDs are SIFT [30], HOG [31], LBP [32] and Gabor [29], which are designed in terms of different assumptions.

- Each SIFT descriptor is a 3D histogram of gradient location and orientation, where the location is quantized into a 4×4 grid and the gradient angle is quantized into eight orientations. Thanks to the localization strategy, SIFT could be invariant to uniform scaling, orientation, illumination changes, and partially invariant to affine distortion.
- HOG counts occurrences of gradient orientation in localized portions of an image. The major difference between HOG and SIFT is that the former is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. The major advantage of HOG is the invariance to geometric and photometric transformations.
- LBP describes the neighboring changes around the central point, which is a simple yet effective way to represent faces. Thanks to the binary pattern, LBP could be invariant to monotone transformation and robust to illumination changes to some extent.
- Gabor wavelets incorporate specific spatial frequency, spatial locality, and selective orientation into the extracted features. In consequence, it is robust to illumination and expression changes [19,29].

To make LFDs adaptive to data distribution, some recent works proposed to learn LFDs in a data-driven way. For example, [19–21] proposed LBP-like descriptors to solve heterogeneous recognition problems through learning and clustering instead of handcrafted ways adopted by traditional LBPs. Despite the improvement in performance, the major disadvantage of these methods [19,20] is the high computational complexity since they need to compute multiple projections for each patch of a view, as well as cluster the corresponding centers.

References

[1] P. Jonathon Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104, doi:10.1109/34.879790.

[2] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale chinese face database and baseline evaluations, IEEE Trans. Syst., Man, Cybern. Part A 38 (1) (2008) 149–161, doi:10.1109/TSMCA.2007.909557.

[3] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1955–1967, doi:10.1109/TPAMI.2008.222.

[4] S.Z. Li, Z. Lei, M. Ao, The HFB face database for heterogeneous face biometrics research, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 2009, pp. 1–8. doi:10.1109/CVPR.2009.5204149.

[5] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 513–520, doi:10.1109/CVPR.2011.5995324.

[6] B. Klare, Z. Li, A.K. Jain, Matching forensic sketches to mug shot photos, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 639–646, doi:10.1109/TPAMI.2010.180.

[7] A. Sharma, D.W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 593–600, doi:10.1109/CVPR.2011.5995350.

[8] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 2160–2167, doi:10.1109/CVPR.2012.6247923.

[9] T.I. Dhamecha, P. Sharma, R. Singh, M. Vatsa, On effectiveness of histogram of oriented gradient features for visible to near infrared face matching, in: International Conference on Pattern Recognition, 2014, pp. 1788–1793.

[10] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194, doi:10.1109/TPAMI.2015.2435740.

[11] J. Li, Y. Wu, J. Zhao, K. Lu, Low-rank discriminant embedding for multiview learning, IEEE Trans. Cybern. 47 (11) (2017) 3516–3529, doi:10.1109/TCYB.2016.2565898.

[12] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3) (1936) 321–377, doi:10.1093/biomet/28.3-4.321.

[13] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data, IEEE Trans. Image Process. 11 (3) (2002) 293–305, doi:10.1109/83.988962.

[14] S. Akaho, A kernel method for canonical correlation analysis, arXiv preprint arXiv:0609071 (4) (2006) 1–7. 0609071.

[15] Z. Lei, S.Z. Li, Coupled spectral regression for matching heterogeneous faces, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1123–1128, doi:10.1109/CVPRW.2009.5206860.

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

[17] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: European Conference on Computer Vision, 2012, pp. 808–821.

[18] B.F. Klare, A.K. Jain, Heterogeneous face recognition using kernel prototype similarities, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1410–1422, doi:10.1109/TPAMI.2012.229.

[19] Z. Lei, M. Pietikainen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 289–302, doi:10.1109/TPAMI.2013.112.

[20] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056, doi:10.1109/TPAMI.2015.2408359.

[21] Y. Jin, J. Lu, Q. Ruan, Coupled discriminative feature learning for heterogeneous face recognition, IEEE Trans. Inf. Forensics Secur. 10 (3) (2015) 640–652, doi:10.1109/TIFS.2015.2390414.

[22] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, Q. He, Multi-task multi-view learning for heterogeneous tasks, in: In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014), Shanghai, China, November 3–7, 2014, pp. 441–450.

[23] J. He, C. Du, F. Zhuang, X. Yin, Q. He, G. Long, Online bayesian max-margin subspace multi-view learning, in: In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), New York, USA, July 9–15, 2016, pp. 1555–1561.

[24] C. Du, C. Du, J. Li, W. Zheng, B. Lu, H. He, Semi-supervised bayesian deep multi-modal emotion recognition, arXiv preprint arXiv:1704.07548 (2017).

[25] T.-K. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: Computer Vision - ECCV 2006, European Conference on Computer Vision, Graz, Austria, May 7–13, 2006, Proceedings, Springer-Verlag Berlin, BERLIN, 2006, pp. 251–262, doi:10.1007/11744078_20.

[26] 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, 2007, pp. 577–584, doi:10.1145/1273496.1273569.

[27] T. Sun, S. Chen, J. Yang, P. Shi, A novel method of combined feature extraction for recognition, in: 8th IEEE International Conference on Data Mining, in: IEEE International Conference on Data Mining, Dec 15–19, 2008, pp. 1043–1048, doi:10.1109/icdm.2008.28.

[28] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Trans. Multimedia 20 (1) (2018) 128–141, doi:10.1109/TMM.2017.2723841.

[29] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476, doi:10.1109/TIP.2002.999679.

[30] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110, doi:10.1023/B:VISI.0000029664.99615.94.

[31] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, volume I, 2005, pp. 886–893, doi:10.1109/CVPR.2005.177.

[32] T. Ahonen, A. Hadid, M. Pietikäinen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041, doi:10.1109/TPAMI.2006.244.

[33] C. Liu, J. Yuen, A. Torralba, J. Sivic, W.T. Freeman, SIFT flow: dense correspondence across different scenes, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 5304 LNCS, 2008, pp. 28–42, doi:10.1007/978-3-540-88690-7-3.

[34] K. Mikolajczyk, K. Mikolajczyk, C. Schmid, C. Schmid, A performance evaluation of local descriptors, IEEE Trans Pattern Anal Mach Intell 27 (10) (2005) 1615–1630, doi:10.1109/TPAMI.2005.188.

[35] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners 13 (3) (1991) 252–264. arXiv: 1202.3819v2, doi: 10.1109/34.75512.

[36] X. Wang, X. Tang, Random sampling for subspace face recognition, Int. J. Comput. Vis. 70 (1) (2006) 91–104.

[37] X. Peng, C. Lu, Z. Yi, H. Tang, Connections between nuclear norm and frobenius norm based representation, IEEE Trans. Neural Networks (2016) 1–7. doi: 10.1109/TNNLS.2016.2608834.

[38] J. Ye, R. Janardan, Q. Li, Two-dimensional linear discriminant analysis, in: Proc. 18th Ann. Conf. Neural Informatiojn processing Systems (NIPS), 2004, pp. 1569–1576.

[39] W.J. Krzanowski, P. Jonathan, W.V. Mccarthy, M.R. Thomas, Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data, Appl. Stat. 44 (1) (1995) 101–115, doi:10.2307/2986198.

[40] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720. arXiv: 1011.1669v3, doi: 10.1109/34.598228.

[41] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (1) (2004) 131–137, doi:10.1109/TPAMI.2004.1261097.

[42] H. Kong, L. Wang, E.K. Teoh, J.G. Wang, R. Venkateswarlu, A framework of 2d fisher discriminant analysis: application to face recognition with small number of training samples, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, 2005, pp. 1083–1088.

[43] X. Peng, H. Tang, L. Zhang, Z. Yi, S. Xiao, A unified framework for representation-based subspace clustering of out-of-sample and large-scale data., IEEE Trans. Neural Networks 27 (12) (2016) 2499–2512.

[44] H. Wang, S. Yan, D. Xu, X. Tang, T.S. Huang, Trace ratio vs. ratio trace for dimensionality reduction (2007) 1–8.

[45] A.N. Tikhonov, The regularization of incorrectly posed problems, Proc. USSR Acad. Sci. 153 (1) (1963) 1624–1627.

[46] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning data mining, inference and prediction, Mathe. Intell. 27 (2) (2009) 83–85, doi:10.1007/b94608.

[47] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (405) (1989) 165–175, doi:10.1080/01621459.1989.10478752.

[48] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–7, doi:10.1109/ICCV.2007.4408856.

[49] C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition, ACM Trans. Intell. Syst. Technol. 7 (3) (2016) 37.

[50] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.

[51] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms, 2008, (http://www.vlfeat.org/).

[52] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Trans. Image Process. 19 (6) (2010) 1635–1650, doi:10.1109/TIP.2010.2042645.

[53] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554. 1111.6189v1, doi: 10.1162/neco.2006.18.7.1527.

[54] H. Lee, P. Pham, Y. Largman, A. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks., Nips (2009) 1–9. arXiv:1301.3605v3, doi:10.1145/1553374.1553453.

[55] J. Wu, J. Long, M. Liu, Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm, Neurocomputing 148 (2015) 136–142, doi:10.1016/j.neucom.2012.10.043.

[56] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, CVPR (2016) 4847–4855, doi:10.1109/CVPR.2016.524.