Discriminant Analysis via Joint Euler Transform and $\ell_{2,1}$-norm

Shuangli Liao, Quanxue Gao, Zhaohua Yang, Fang Chen, Feiping Nie, and Jungong Han

***Abstract*—Linear Discriminant analysis (LDA) has been widely used for face recognition. However, when identifying faces in the wild, the existence of outliers that deviate significantly from the rest of data can arbitrarily skew the desired solution. This usually deteriorates LDA's performance dramatically, thus preventing it from mass deployment in real-world applications. To handle this problem, we propose an effective distance metric learning method based LDA, namely Euler LDA-L21 (e-LDA-L21). e-LDA-L21 is carried out in two stages, in which each image is mapped into a complex space by Euler transform in the first stage and the $\ell_{2,1}$-norm is adopted as the distance metric in the second stage. This not only reveals nonlinear features but also exploits the geometric structure of data. To solve e-LDA-L21 efficiently, we propose an iterative algorithm, which is a closed-form solution at each iteration with convergence guaranteed. Finally, we extend e-LDA-L21 to Euler 2DLDA-L21 (e-2DLDA-L21) which further exploits the spatial information embedded in image pixels. Experimental results on several face databases demonstrate its superiority over the state-of-the-art algorithms.**

***Index Terms*—Linear Discriminant Analysis, Two-dimensional Linear Discriminant analysis, Dimensionality reduction, Euler transform, $\ell_{2,1}$-norm.**

I. Introduction

FINDING an effective representation for image has been a fundamental problem in the fields of pattern recognition and machine learning. During the last few decades, many approaches have been developed for this task, among which principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2] turn out to be two representatives. PCA is a unsupervised method used for extracting the most representative features, while LDA is a supervised one capable of extracting the most discriminative features. Not surprisingly, LDA generally performs better when conducting data classification task.

In the domain of image analysis, it is essential to reshape the original two-dimensional image into a high-dimensional vector prior to applying the aforementioned methods. This may impair the spatial structure information embedded in the pixels of image. To tackle this problem, many image matrix-based dimensionality reduction methods have been developed, most of which follow the idea of directly estimating the scatter matrices from image matrices and then solving the projection matrix by optimizing the criterion function. Such matrix based methods include two-dimensional PCA (2DPCA) [3], multilinear PCA [4], two-dimensional LDA (2DLDA) [5], tensor LDA [6], and direction tensor independent component analysis [7]. Despite the impressive results obtained in many cases, these approaches struggle with the outliers existed in the real-life applications. The reason behind is that all of the above approaches characterize the geometric structure of data by squared ℓ_2-norm, which is not robust in the sense that outlying measurements would arbitrarily skew the acquired solution from the desired solution [8] [9].

Most existing works have demonstrated that distance metric learning based methods can effectively improve the robustness of algorithms against outliers [10]–[18]. Team Bischof proposed KISSME metric learning [10], which is based on a statistical inference perspective. Martinel *et al.* [11] proposed a Kernelized Saliency-based method for multiple metic learing. Li *et al.* [12] integrated the distance metric to the SVM decision fuction. Alternatively, Liao *et al.* [13] incorporated the dimension reduction and metric learning, and further used the PSD constraints [14] to enhance the robustness of the learned metric. To handle the small sample size (SSS) problem, Zhang *et al.* [15] proposed a method by matching samples in a discriminative null space of the training data. It is noted that there are some other methods based on temporal model [16] and video [17].

Compared to squared ℓ_2-norm, nuclear-norm [19] [26] and ℓ_1-norm [20]–[26] are more robust to outliers for dimensionality reduction. Using nuclear-norm as the distance metric, Zhang *et al.* [26]. proposed nuclear-norm based 2DPCA (N-2DPCA) that seeks the projection matrix by minimizing the nuclear-norm based reconstruction error. Similarly, many subspace learning methods employ ℓ_1-norm as the distance metric in order to obtain more robust projection vectors. For example, L1-PCA [20] employed ℓ_1-norm to calculate the reconstruction error, which was minimized afterwards via a greedy algorithm. To further speed up the L1-PCA algorithm, Kwak [21] proposed to seek the projection vectors by maximizing the ℓ_1-norm variance. This method is called PCA-L1, which was extended later to 2DPCA with L1-norm maximization (2DPCA-L1) [22]. Nie *et al.* proposed a non-greedy iterative method to solve PCA-L1 [23], which was extended to non-

Manuscript received Jan, 2016; revised ********; accepted ********. This work is supported by National Natural Science Foundation of China under Grant 61773302, the 111 Project of China (B08038), and Shenzhen Fundamental Research fund under Grant JCYJ20160530141902978.

Corresponding author: Q. Gao (e-mail: xd_ste_pr@163.com) and Z. Yang (yangzh@buaa.edu.cn).

The authors wish it to be known that, in their opinion, the first and third authors should be regarded as joint First Authors

S. Liao, Q. Gao and F. Cheng are with State Key Laboratory of Integrated Services Networks, Xidian University, 710071, Xi'an CHina

Z. Yang is with the School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing, China

F. Nie is with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, 710069, Xi'an China.

J. Han is School of Computing and Communications, Lancaster University, United Kingdom.

greedy 2DPCA-L1 [24]. To encode the discriminant information better, some ℓ_1-norm based LDA methods have been developed for image classification [27]–[30]. Two of the most popular methods are LDA-L1[28] and l1-norm based kernel LDA [29], where the former used l1-norm to calculate the within-class scatter matrix and between-class scatter matrix, while the latter employed l1-norm as the distance metric in kernel LDA. Liu *et al.* [31] and Wang *et al.* [32] proposed non-greedy algorithms to solve ℓ_1-norm based discriminant analysis methods, which can optimize the criterion function. To well exploit discriminant spatial structure, Li *et al.* [33] proposed L1-2DLDA for image classification.

However, ℓ_1-norm subspace methods do not well characterize the geometric structure due to the fact that the solution of ℓ_1-norm subspace methods relates nothing to scatter matrices that characterize the geometric structure of data [34]–[36]. Moreover, ℓ_1-norm based subspace techniques cannot provide the subspace with join-feature spareseness [37]–[41]. To handle this problem, Ding *et al.* [35] developed $\ell_{2,1}$-norm and used it to estimate the reconstruction error of data. Based on it, a novel method R1-PCA, which seeks low-dimension space by minimizing $\ell_{2,1}$-norm reconstruction error, was proposed to improve the robustness of PCA. Nie *et al.* [38] showed that $\ell_{2,1}$-norm helps select useful features from high-dimensional data. Inspired by these works, $\ell_{2,1}$-norm has been widely used as the regularization term in the criterion function. For example, Wong *et al.* [40] proposed $\ell_{2,1}$-norm based tensor feature selection for image analysis. Gui *et al.* [37] used $\ell_{2,1}$-norm as the regularized term in subspace learning and proposed a joint feature extraction and selection method for data classification. Since all of them still used squared Euclidean distance to measure the similarity between data, the flexibility and robustness of the algorithms are adversely affected. Moreover, for image classification, most existing discriminant methods still measure the similarity in the pixel space, thus leading the methods to be sensitive to illumination and outliers. Finally, the aforementioned robust discriminant methods do not reveal non-linear features which can actually help improve the robustness of algorithms.

Motivated by the fact that kernel trick can capture the nonlinear features [42]–[44] and combine the superiority of $\ell_{2,1}$-norm, in this paper, we develop a distance metric learning based robust LDA, namely *e*-LDA-L21, for discriminative feature extraction. *e*-LDA-L21 first maps the original images into the complex space by Euler transform, which not only suppresses outliers but also reveals non-linear patterns embedded in data. Afterwards, it uses $\ell_{2,1}$-norm to measure both between-class scatter matrix and within-class scatter matrix in the complex space. Likewise, we further extend this concept to handle 2D data and propose an e-2DLDA-L21 algorithm. Experimental results reveal the effectiveness of our proposed method. In contrast to most existing robust subspace methods, we have the following contributions:

- By analyzing Euclidean distance, ℓ_1-norm, and $\ell_{2,1}$-norm, we have showed that $\ell_{2,1}$-norm can help enlarge the role of small between-class distance and weaken the effect of large between-class distance. Thus, we employ $\ell_{2,1}$-norm as the distance metric in LDA. This helps improve the robustness of LDA against to outliers.
- Our method extracts nonlinear features with Euler transform in LDA. Different from the commonly used kernel function, which maps data into a higher-dimensional hidden space, Euler transform maps data into an explicitly space which has the same dimensionality as that in the original data space. As a result, our method can be easily implemented in real applications. Moreover, we have showed that cosine distance metric not only helps obtain a large margin but also improves the separability between different class images. Thus, our method well encodes nonlinear discriminant features and simultaneously gets a large margin which is important for classification.
- Our method integrates both kernel trick and distance metric learning into the criterion function. It helps further improve the robustness of algorithm. Moreover, the proposed iterative algorithm has a good convergence.

The remainder of this paper is organized as follows. Section 2 reviews the related work. $\ell_{2,1}$-norm-based LDA with Euler representation, namely *e*-LDA-L21, is proposed in Section 3. We extend *e*-LDA-L21 to *e*-2DLDA-L21 in Section 4. Section 5 reports experimental results. Finally, the conclusions are drawn in Section 6.

II. Linear Discriminant Analysis

Assume that we have N training images $\mathbf{X}_j \in \mathbf{R}^{m\times n}(j = 1, 2, ..., N)$, where m and n denote the number of rows and columns of an image, respectively. The given data have c classes and the ith class has n_i samples. $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}$ denote the class mean image of the ith class and the mean image of all image samples, respectively. Denote $\mathbf{x}_j$, $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}$ $\in \mathbf{R}^{M\times 1}(M = m \times n)$ by the vector forms of matrices $\mathbf{X}_j$, $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}$, respectively, i.e., $\mathbf{x}_j = vect(\mathbf{X}_j)$, $\bar{\mathbf{x}}_i = vect(\bar{\mathbf{X}}_i)$, and $\bar{\mathbf{x}} = vect(\bar{\mathbf{X}})$.

A. LDA and 2DLDA

LDA aims to seek the projection matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d] \in \mathbf{R}^{M\times d}$ which minimizes the within-class scatter and simultaneously maximizes the between-class scatter in the low -dimensional space. The objective function of LDA is [2]

$$\mathbf{W}_{opt} = \arg\max_{\mathbf{w}_k{}^T\mathbf{w}_k=1} \frac{tr(\mathbf{W}^T\mathbf{S_b}\mathbf{W})}{tr(\mathbf{W}^T\mathbf{S_w}\mathbf{W})} \tag{1}$$

where $\mathbf{w}_k(k = 1, 2, \cdots, d)$ is the kth column of the matrix $\mathbf{W}$, $tr(\cdot)$ is the trace operator of a matrix, between-class scatter matrix $\mathbf{S_b}$ and within-class scatter matrix $\mathbf{S_w}$ are defined as

$$\mathbf{S_b} = \sum_{i=1}^{c} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \tag{2}$$

$$\mathbf{S_w} = \sum_{i=1}^{c}\sum_{j\in c_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T \tag{3}$$

where c_i denotes the ith class data set.

The objective function (1) is a trace ratio optimization problem and typically nonconvex, and there does not exist a closed-form solution for the general trace ratio problem

(1). Hence, such problems are often transformed into the simpler yet inexact ratio trace problem, which is equivalent to the determinant ratio problem [45]. For the model (1), the corresponding ratio trace (determinant ratio) form is

$$\mathbf{W}_{opt} = \arg\max_{\mathbf{W}} tr\left(\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)^{-1}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)\right) \quad (4)$$

The optimal projection matrix $\mathbf{W}_{opt}$ of the objective function (4) consists of the eigenvectors $\{\mathbf{w}_k\,|k=1,2,\cdots,d\}$ of the Eigen-equation $(\mathbf{S_w})^{-1}\mathbf{S_b}\mathbf{w}_k = \lambda_k\mathbf{w}_k$ corresponding to the first d largest eigenvalues $\{\lambda_k\,|k=1,2,\cdots,d\}$.

Each image needs to be transformed into a 1D vector in LDA-based methods. Thus, they can no longer exploit the spatial information embedded in pixels. To handle this problem, 2DLDA is one of the most representative methods. In order to be consistent with the existing improved methods, such as L1-2DLDA [33], the model of 2DLDA only considers the left projection here. The model is

$$\mathbf{Q}_{opt} = \arg\max_{\mathbf{q}_k{}^T\mathbf{q}_k=1} \frac{tr\left(\mathbf{Q}^T\mathbf{G}_b\mathbf{Q}\right)}{tr\left(\mathbf{Q}^T\mathbf{G}_w\mathbf{Q}\right)} \quad (5)$$

where $\mathbf{Q} = [q_1, q_2, ..., q_d] \in \mathbf{R}^{m\times d}$ and $\mathbf{q}_k(k=1,2,\cdots,d)$ is the k-th column of the matrix $\mathbf{Q}$, between-class scatter matrix $\mathbf{G}_b$ and within-class scatter matrix $\mathbf{G}_w$ are defined as

$$\mathbf{G}_b = \sum_{i=1}^{c} n_i\left(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}\right)\left(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}\right)^T \quad (6)$$

$$\mathbf{G}_w = \sum_{i=1}^{c}\sum_{j\in c_i}\left(\mathbf{X}_j - \bar{\mathbf{X}}_i\right)\left(\mathbf{X}_j - \bar{\mathbf{X}}_i\right)^T \quad (7)$$

By simple algebra, the nominators and denominators of the objective functions (1) and (5) can be respectively rewritten as

$$tr(\mathbf{W}^T\mathbf{S_b}\mathbf{W}) = \sum_{i=1}^{c} n_i\left\|\mathbf{W}^T(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right\|_2^2 \quad (8)$$

$$tr(\mathbf{W}^T\mathbf{S_w}\mathbf{W}) = \sum_{i=1}^{c}\sum_{j\in c_i}\left\|\mathbf{W}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right\|_2^2 \quad (9)$$

$$tr\left(\mathbf{Q}^T\mathbf{G}_b\mathbf{Q}\right) = \sum_{i=1}^{c} n_i\sum_{k=1}^{n}\left\|\mathbf{Q}^T\left(\bar{\mathbf{X}}_i(:,k) - \bar{\mathbf{X}}(:,k)\right)\right\|_2^2 \quad (10)$$

$$tr\left(\mathbf{Q}^T\mathbf{G}_w\mathbf{Q}\right) = \sum_{i=1}^{c}\sum_{j\in c_i}\sum_{k=1}^{n}\left\|\mathbf{Q}^T\left(\mathbf{X}_j(:,k) - \bar{\mathbf{X}}_i(:,k)\right)\right\|_2^2 \quad (11)$$

where $\bar{\mathbf{X}}_i(:,k)$ and $\mathbf{X}_j(:,k)$ denote the kth column of matrices $\bar{\mathbf{X}}_i$ and $\mathbf{X}_j$, respectively.

As can be seen in Eqs. (8), (9), (10) and (11), the large distance will remarkably dominate the solution of the models (1) and (5). Thus, the optimal solution of the objective function (1) and (5) is susceptible to the presence of outliers which deviate significantly from the rest of data. Moreover, squared ℓ_2-norm can suppress the role of small between-class scatter in the criterion function. Thus, traditional LDA technique is unlikely to obtain a large margin, which is important for classification, in the low-dimensional space.

B. LDA-L1 and L1-2DLDA

To improve the robustness of discriminant analysis technique, many enhanced discriminant methods have been developed, among which LDA-L1 and L1-2DLDA [33] are two of the most representative methods. The objective function of LDA-L1 is

$$\mathbf{W}_{opt} = \arg\max_{\mathbf{w}_k{}^T\mathbf{w}_k=1} \frac{\sum_{i=1}^{c} n_i\left\|\mathbf{W}^T(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})\right\|_1}{\sum_{i=1}^{c}\sum_{j\in c_i}\left\|\mathbf{W}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right\|_1} \quad (12)$$

L1-2DLDA aims to seek projection matrix $\mathbf{Q}$ by the model (13).

$$\mathbf{Q}_{opt} = \arg\max_{\mathbf{q}_k{}^T\mathbf{q}_k=1} \frac{\sum_{i=1}^{c} n_i\left\|\mathbf{Q}^T\left(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\right\|_{L1}}{\sum_{i=1}^{c}\sum_{j\in c_i}\left\|\mathbf{Q}^T\left(\mathbf{X}_j - \overline{\mathbf{X}}_i\right)\right\|_{L1}} \quad (13)$$

where $\|\bullet\|_{L1}$ denotes the ℓ_1-norm of a matrix which can be defined as $\|\mathbf{X}\|_{L1} = \sum_k \|\mathbf{X}(:,k)\|_1$, $\mathbf{X}(:,k)$ is the kth column of matrix $\mathbf{X}$.

Although the linear discriminant analysis technique based on ℓ_1-norm is robust to outliers, compared with traditional LDA, the solutions of both the models (13) and (12) are irrelevant to the scatter matrices that characterize the geometric structure. Thus, neither LDA-L1 nor L1-2DLDA can well reveal the geometric structure that is crucial for classification. Moreover, it is difficult to solve ℓ_1-norm optimization problem. Consequently, it is unclear whether ℓ_1-norm can help improve the role of small between-class scatter in the models (13) and (12).

III. Euler LDA-L21

A. Motivation

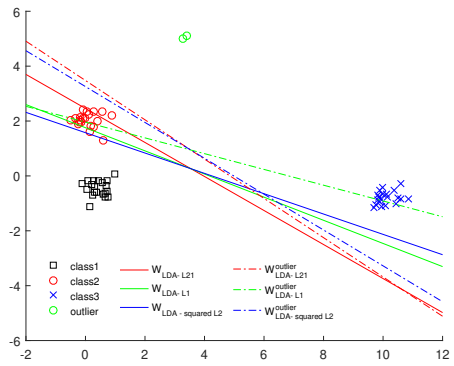
As can be seen from the above analysis, *squared Euclidean distance* actually makes the outlying measurements dominate the solution of the objective function (1). This leads the objective function (1) to be more prone to the outliers. To handle this problem, the distance metric in the model (1) should not only suppress the outliers but also enlarge the role of small between-class scatters. Combining the aforementioned analysis and the recent works [42] [43], we present an efficient and robust LDA for dimensionality reduction. It aims to seek a robust projection matrix $\mathbf{W}$ such that the projected data can well reveal not only discriminant geometric structure of data but also non-linear features.

Prior to formulating the proposed method, we first introduce the definition of $\ell_{2,1}$-norm and cosine distance metric that can be viewed as a kernel function [42] [43], and then analyze their advantages to outliers and discriminant geometric structure.

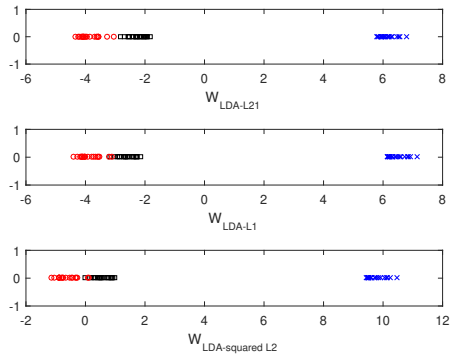
Definition 1 [35]. Given an arbitrary matrix $\mathbf{A} = [\mathbf{A}(i,j)] \in \mathbf{R}^{m\times n}$, the $\ell_{2,1}$-norm of matrix $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{L_{2,1}} = \sum_{j=1}^{n}\sqrt{\sum\nolimits_{i=1}^{m}\mathbf{A}^2(i,j)} = \sum_{j=1}^{n}\|\mathbf{A}(:,j)\|_2 \quad (14)$$

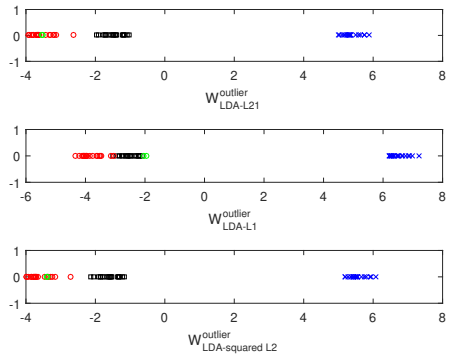
As can be seen in Eq. 14, from the norm point of view, $\ell_{2,1}$-norm of matrix is essentially the ℓ_2-norm which has no



(a)



(b)



(c)

Figure 1. (a) Directions of LDA, LDA-L21 and LDA-L1; (b) projected data of LDA, LDA-L21 and LDA-L1; (c) projected data of LDA, LDA-L21 and LDA-L1 with outliers.

essential difference with squared ℓ_2-norm. Thus, if $\ell_{2,1}$-norm is employed as the distance metric in the criterion function, it can well reveal the geometric structure of data. Moreover, compared with squared ℓ_2-norm, $\ell_{2,1}$-norm can further help enlarge the role of small between-class distance and weaken the effect of large between-class distance in the criterion function. This results in not only the enhanced robustness to outliers but also a large margin in the low-dimensional space, thus improving the performance of algorithm. For example, we randomly produce three classes data points, which are marked with different shapes and each class has 20 data points. The three data classes follow Gaussian distribution with covariance matrices being [0.3 0; 0 0.3] and means being [0.5 -0.5] and [0 2], and [10 -1], respectively. Moreover, we also add two outliers whose color is green. Two outliers belong to the second class. Taking these data points as training samples, we show the directions of LDA, LDA-L21, and LDA-L1, and the corresponding low-dimensional representation in Figure 1. $\mathbf{W}_{LDA}$, $\mathbf{W}_{LDA-L21}$ and $\mathbf{W}_{LDA-L1}$ denote the directions of LDA, LDA-L21 and LDA-L1 respectively when training data do not include outliers. $\mathbf{W}_{LDA}^{outlier}$, $\mathbf{W}_{LDA-L21}^{outlier}$ and $\mathbf{W}_{LDA-L1}^{outlier}$ denote the directions of the aforementioned methods when training data have outliers. As can be seen in Figure 1, when training data are clean, LDA with $\ell_{2,1}$-norm as distance metric obtains a large margin between class 1 and class 2 in the low-dimensional space and well separates all classes in the low-dimensional space, whereas traditional LDA and LDA-L1 do not. When training data include outliers, both our model and LDA can correctly classify all data, while LDA-L1 does not. Moreover, our model achieves a larger margin than LDA. This observation motivates us to employ $\ell_{2,1}$-norm as distance metric in LDA criterion function.

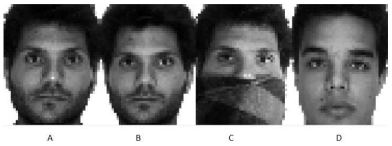


Figure 2. Examples motivating the use of the cosine-based dissimilarity measure. Images from left to right are the original image, a second image of the same subject, an occluded version of the original image and an image of another subject.

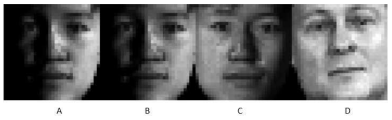


Figure 3. Examples motivating the use of the cosine-based dissimilarity measure. Images from left to right are the original image, a second image of the same subject, an illumination change version of the original image and an image of another subject.

Definition 2 [42] [43]. Given two arbitrary vectors $\mathbf{x}_p$ and $\mathbf{x}_q \in \mathbf{R}^{M\times 1}$, the cosine distance metric between them is

$$d\left(\mathbf{x}_p, \mathbf{x}_q\right) = \sum_{k=1}^{M} \left\{1 - \cos\left(\alpha\pi\left(\mathbf{x}_p\left(k\right) - \mathbf{x}_q\left(k\right)\right)\right)\right\} \qquad (15)$$

where α is an alpha mask [42], which is also known as alpha matte or alpha channel and associates variable transparency with an image. $\mathbf{x}_p\left(k\right)$ and $\mathbf{x}_q\left(k\right)$ are the k-th element of $\mathbf{x}_p$ and $\mathbf{x}_q$, respectively.

TABLE I
COMPARISON OF NORMALIZED DISSIMILARITY MEASURES BETWEEN IMAGES IN FIGURE2.

Dissimilarity metric	A-B	A-C	A-D
Euclidean distance	6.5437	11.5016	8.5694
ℓ_1-norm distance	158.2643	326.4584	275.6062
Cosine-based distance	352.1622	622.0739	851.188

TABLE II
COMPARISON OF NORMALIZED DISSIMILARITY MEASURES BETWEEN IMAGES IN FIGURE 3.

Dissimilarity metric	A-B	A-C	A-D
Euclidean distance	1.2629	16.4548	13.1405
ℓ_1-norm distance	29.6755	450.4374	359.8113
Cosine-based distance	28.0423	1159.577	1240.738

By simple algebra, Eq. (15) becomes

$$\begin{aligned} d(\mathbf{x}_p,\mathbf{x}_q) &= \sum_{k=1}^{M}\{1-\cos\left(\alpha\pi(\mathbf{x}_p(k)-\mathbf{x}_q(k))\right)\} \\ &= \frac{1}{2}\sum_{k=1}^{M}\Big\{\left(\cos\alpha\pi\mathbf{x}_p(k)-\cos\alpha\pi\mathbf{x}_q(k)\right)^2 \\ &\qquad +\left(\sin\alpha\pi\mathbf{x}_p(k)-\sin\alpha\pi\mathbf{x}_q(k)\right)^2\Big\} \\ &= \left\|\tfrac{1}{\sqrt{2}}(e^{i\alpha\pi\mathbf{x}_p}-e^{i\alpha\pi\mathbf{x}_q})\right\|_2^2 \\ &= \|\mathbf{z}_p-\mathbf{z}_q\|_2^2 \end{aligned} \tag{16}$$

where

$$\mathbf{z}_q = \frac{1}{\sqrt{2}}\begin{bmatrix} e^{i\alpha\pi\mathbf{x}_q(1)} \\ \vdots \\ e^{i\alpha\pi\mathbf{x}_q(M)} \end{bmatrix} = \frac{1}{\sqrt{2}}e^{i\alpha\pi\mathbf{x}_q} \tag{17}$$

is called Euler representation of $\mathbf{x}_q$.

Eq. (16) illustrates that cosine distance metric between two images in the pixel space is equivalent to the squared ℓ_2-norm between the corresponding two images with Euler representation.

Let us consider two motivating examples in which different dissimilarity measures are applied to the images that are shown in Figure 2 and Figure 3. Table I and Table II list the Euclidean distance, ℓ_1-norm distance and cosine distance metric between images in Figure 2 and Figure 3, respectively. As can be seen in Figure 2, Figure 3, Table I and Table II, Euclidean distance and ℓ_1-norm distance associate a smaller distance between the original image and an image from a different subject, while the distance between image A and image C, which belong to the same person with occlusion or different illumination, is large. In contrast, the use of the cosine-based measure results in a large distance between images that are from different persons.

Combining the aforementioned analysis, we have the following interesting observations:

- First, the cosine distance metric enlarges the distance between all images, so cosine distance metric can be beneficial to the data classification.
- Second, the distance between the same class images is enlarged a smaller multiple than the distance between the different class images by the cosine distance metric. This clearly shows that cosine distance metric not only helps obtain a large margin but also improves the separability between different class images, compared with Euclidean distance and ℓ_1-norm.

B. Objective function

To improve the robustness of discriminant analysis technique, we present our model with the aforementioned analysis, which employs $\ell_{2,1}$-norm as the distance metric to measure the similarity between data in the Euler space. Specifically, we first map each image $\mathbf{x}_q$ onto the Euler space by Eq. (17) and then use $\ell_{2,1}$-norm to reveal within-class and between-class scatter matrices in the Euler space. Our model is as follows.

$$\mathbf{W}_{opt} = \underset{\mathbf{w}_k{}^H\mathbf{w}_k=1}{\arg\max} \frac{\sum_{i=1}^{c} n_i\left\|\mathbf{W}^H(\bar{\mathbf{z}}_i-\bar{\mathbf{z}})\right\|_2}{\sum_{i=1}^{c}\sum_{j\in c_i}\left\|\mathbf{W}^H(\mathbf{z}_j-\bar{\mathbf{z}}_i)\right\|_2} \tag{18}$$

where $\mathbf{z}_j$, $\bar{\mathbf{z}}_i$ and $\bar{\mathbf{z}}$ are the Euler representation of $\mathbf{x}_j$, $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}$, respectively. They can be calculated by Eq. (17).

The matrix form of model (18), called *e*-LDA-L21, is as follows.

$$\mathbf{W}_{opt} = \underset{\mathbf{w}_k{}^H\mathbf{w}_k=1}{\arg\max} \frac{\left\|\mathbf{W}^H\mathbf{\Phi}_b\right\|_{L_{2,1}}}{\left\|\mathbf{W}^H\mathbf{\Phi}_w\right\|_{L_{2,1}}} \tag{19}$$

where $\mathbf{\Phi}_b = [n_1(\bar{\mathbf{z}}_1-\bar{\mathbf{z}}), n_2(\bar{\mathbf{z}}_2-\bar{\mathbf{z}}), \cdots, n_c(\bar{\mathbf{z}}_c-\bar{\mathbf{z}})]$, $\mathbf{\Phi}_w = [\mathbf{z}_1-\bar{\mathbf{z}}_1, \cdots, \mathbf{z}_N-\bar{\mathbf{z}}_c]$.

By simple algebra, the numerator of Eq. (19) can be rewritten as

$$\begin{aligned} \left\|\mathbf{W}^H\mathbf{\Phi}_b\right\|_{L_{2,1}} &= \sum_{i=1}^{c}\left\|\mathbf{W}^H\mathbf{\Phi}_{\mathbf{b}}(:,i)\right\|_2 \\ &= \sum_{i=1}^{c}\left\|\mathbf{W}^H\mathbf{\Phi}_b(:,i)\right\|_2^2\frac{1}{\left\|\mathbf{W}^H\mathbf{\Phi}_b(:,i)\right\|_2} \\ &= \sum_{i=1}^{c} tr\left(\mathbf{W}^H\mathbf{\Phi}_b(:,i)\mathbf{d}_i\mathbf{\Phi}_b(:,i)^H\mathbf{W}\right) \\ &= tr\left(\mathbf{W}^H\mathbf{\Phi}_b\mathbf{D}\mathbf{\Phi}_b{}^H\mathbf{W}\right) \end{aligned} \tag{20}$$

where $\mathbf{D} = diag(\frac{1}{\|\mathbf{W}^H\mathbf{\Phi}_b(:,1)\|_2}, ..., \frac{1}{\|\mathbf{W}^H\mathbf{\Phi}_b(:,c)\|_2})$.

Similarly, the denominator of Eq. (19) can be rewritten as follows:

$$\begin{aligned} \left\|\mathbf{W}^H\mathbf{\Phi}_w\right\|_{L_{2,1}} &= \sum_{j=1}^{N}\left\|\mathbf{W}^H\mathbf{\Phi}_w(:,j)\right\|_2 \\ &= \sum_{j=1}^{N}\left\|\mathbf{W}^H\mathbf{\Phi}_w(:,j)\right\|_2^2\frac{1}{\left\|\mathbf{W}^H\mathbf{\Phi}_w(:,j)\right\|_2} \\ &= tr\left(\mathbf{W}^H\mathbf{\Phi}_w\mathbf{E}\mathbf{\Phi}_w{}^H\mathbf{W}\right) \end{aligned} \tag{21}$$

where $\mathbf{E} = diag(\frac{1}{\|\mathbf{W}^H\mathbf{\Phi}_w(:,1)\|_2}, ..., \frac{1}{\|\mathbf{W}^H\mathbf{\Phi}_w(:,N)\|_2})$.

Substituting Eq. (20), Eq. (21) into Eq. (19), the objective function (19) becomes

$$\mathbf{W}_{opt} = \underset{\mathbf{w}_k{}^H\mathbf{w}_k=1}{\arg\max} \frac{tr\left(\mathbf{W}^H\mathbf{\Phi}_b\mathbf{D}\mathbf{\Phi}_b{}^H\mathbf{W}\right)}{tr\left(\mathbf{W}^H\mathbf{\Phi}_w\mathbf{E}\mathbf{\Phi}_w{}^H\mathbf{W}\right)} \tag{22}$$

In the literature [2] [5] [45], Eq. (22) can be converted to

$$\begin{aligned} \mathbf{W}_{opt} = & \underset{\mathbf{w}_k{}^H\mathbf{w}_k=1}{\arg\min}\ tr\left(\mathbf{W}^H\mathbf{\Phi}_w\mathbf{E}\mathbf{\Phi}_w{}^H\mathbf{W}\right) \\ s.t. \quad & \mathbf{W}^H\mathbf{\Phi}_b\mathbf{D}\mathbf{\Phi}_b{}^H\mathbf{W} = \mathbf{T} \end{aligned} \tag{23}$$

where $\mathbf{T}$ is a constant matrix with all elements being constants.

The model (23) is a constrained optimization problem, which is usually solved by Lagrange multiplier method. The Lagrangian function of the problem (23) is

$$L\left(\mathbf{W}\right) = tr\left(\mathbf{W}^H\mathbf{C}\mathbf{W}\right) - tr\left(\Lambda\left(\mathbf{W}^H\mathbf{B}\mathbf{W} - \mathbf{T}\right)\right) \tag{24}$$

where Λ is a diagonal matrix for enforcing the constraints in Eq. (23), $\mathbf{B} = \mathbf{\Phi}_b\mathbf{D}\mathbf{\Phi}_b{}^H$ and $\mathbf{C} = \mathbf{\Phi}_w\mathbf{E}\mathbf{\Phi}_w{}^H$.

The KKT condition for the optimal solution is that the gradient of $L\left(\mathbf{W}\right)$ must be zero, i.e.,

$$\frac{\partial L\left(\mathbf{W}\right)}{\partial \mathbf{W}} = \mathbf{C}\mathbf{W} - \mathbf{B}\mathbf{W}\Lambda = 0 \tag{25}$$

By simple algebra, we have

$$\mathbf{B}^{-1}\mathbf{C}\mathbf{W} = \mathbf{W}\Lambda \tag{26}$$

The optimal projection matrix $\mathbf{W}$ of the objective function (23) consists of the eigenvectors of the Eigen-equation $\mathbf{B}^{-1}\mathbf{C}\mathbf{w}_k = \lambda_k\mathbf{w}_k (k = 1, \cdots, d)$ corresponding to the first d smallest eigenvalues except zeros. The aforementioned process needs to solve the inverse of matrix $\mathbf{B}$ which may not exist if the data dimensions are very high, and easily causes unstable solution due to the small sample size problem. To handle this problem, we use complex PCA to reduce the dimension of $\mathbf{B}$ and $\mathbf{C}$ in advance. Denote $\mathbf{P}$ by PCA projection matrix, then

$$\mathbf{\Psi}_b = \mathbf{P}^H\mathbf{B}\mathbf{P} \tag{27}$$

$$\mathbf{\Psi}_w = \mathbf{P}^H\mathbf{C}\mathbf{P} \tag{28}$$

Let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d]$, $\mathbf{u}_k$ and λ_k $(k = 1, \cdots, d)$ be the eigenvector and the corresponding eigenvalues of the following eigen-equation, respectively, and ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

$$\left(\mathbf{\Psi}_b\right)^{-1}\left(\mathbf{\Psi}_w\right)\mathbf{u}_k = \lambda_k\mathbf{u}_k \tag{29}$$

then

$$\mathbf{W} = \mathbf{P}\mathbf{U} \tag{30}$$

We summarize the pseudo code of solving the objective function (23), i.e., *e*-LDA-L21 in **Algorithm 1**.

C. Convergence analysis

Corollary 1. In t-th iteration of *Algorithm* 1, we have

$$\begin{aligned} & tr\left(\left(\mathbf{W}^{(t+1)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t+1)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t+1)}\right) \\ & \qquad \leq tr\left(\left(\mathbf{W}^{(t)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t)}\right) \end{aligned} \tag{31}$$

Proof: For t-th iteration, $\mathbf{W}^{(t+1)}$ is the optimal solution of the objective function (23) according to the Eq. (24) and Eq. (25). Thus the following inequality holds.

$$\begin{aligned} & tr\left(\left(\mathbf{W}^{(t+1)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t+1)}\right) \\ & \qquad \leq tr\left(\left(\mathbf{W}^{(t)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t)}\right) \end{aligned} \tag{32}$$

Algorithm 1: *e*-LDA-L21 algorithm

Require:
Initialize $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_N] \in \mathbf{R}^{M\times N}$ as data sample.
Initialize parameter $\alpha = 1.1$, ε=$1e-8$, $\mathbf{W}^{(\mathbf{1})} \in \mathbf{W}^{M\times d}$ which satisfies $\mathbf{w}_k{}^H\mathbf{w}_k = \mathbf{1}$ $(k = 1, 2, \cdots, d)$, $t = 1$.

Ensure:
1. Calculate $\mathbf{z}_j$ by Eq. (17).
2. Calculate the PCA projection matrix $\mathbf{P}$.
3. Calculate $\delta = \left|J(\mathbf{W}^{(t)}) - J(\mathbf{W}^{(t-1)})\right|$, where $J(\mathbf{W}^{(t)}) = \sum_{i=1}^{c}\sum_{j\in c_i}\left\|\left(\mathbf{W}^{(t)}\right)^H\left(\mathbf{z}_j - \bar{\mathbf{z}}_i\right)\right\|_2$

while $\delta \geq \varepsilon$ **do**
4. Calculate $\mathbf{\Psi}_b$ by Eq. (27) and calculate $\mathbf{\Psi}_w$ by Eq. (28).
5. Calculate the eigenvector $\mathbf{U}$ of $\left(\mathbf{\Psi}_b\right)^{-1}\left(\mathbf{\Psi}_w\right)$.
6. Calculate $\mathbf{W} = \mathbf{P}\mathbf{U}$.
7. Update δ.
8. Update $t : t \leftarrow t + 1$.

end while

Output: $\mathbf{W}$

Let $\mathbf{F}^{(t)} = \mathbf{\Phi}_w{}^H\mathbf{W}^{(t)}$, then we get

$$tr\left(\left(\mathbf{F}^{(t+1)}\right)^H\mathbf{E}^{(t)}\mathbf{F}^{(t+1)}\right) \leq tr\left(\left(\mathbf{F}^{(t)}\right)^H\mathbf{E}^{(t)}\mathbf{F}^{(t)}\right) \tag{33}$$

Substituting $\mathbf{E}^{(t)}$ into Eq. (32) then

$$\sum_{j=1}^{N}\frac{\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2^2}{\left\|\mathbf{F}^{(t)}(:,j)\right\|_2} \leq \sum_{j=1}^{N}\left\|\mathbf{F}^{(t)}(:,j)\right\|_2 \tag{34}$$

According to $a^2 + b^2 \geq 2ab$, we can get

$$2\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2 - \left\|\mathbf{F}^{(t)}(:,j)\right\|_2 \leq \frac{\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2^2}{\left\|\mathbf{F}^{(t)}(:,j)\right\|_2} \tag{35}$$

Then

$$\sum_{j=1}^{N}\left(2\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2 - \left\|\mathbf{F}^{(t)}(:,j)\right\|_2\right) \leq \sum_{j=1}^{N}\frac{\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2^2}{\left\|\mathbf{F}^{(t)}(:,j)\right\|_2} \tag{36}$$

Combining Eq. (34) and Eq. (36) yields

$$\sum_{j=1}^{N}\left\|\mathbf{F}^{(t+1)}(:,j)\right\|_2 \leq \sum_{j=1}^{N}\left\|\mathbf{F}^{(t)}(:,j)\right\|_2 \tag{37}$$

Substituting $\mathbf{F}^{(t)}$ and $\mathbf{F}^{(t+1)}$ into Eq. (37) then

$$\sum_{j=1}^{N}\left\|\mathbf{\Phi}_w{}^H\mathbf{W}^{(t+1)}(:,j)\right\|_2 \leq \sum_{j=1}^{N}\left\|\mathbf{\Phi}_w{}^H\mathbf{W}^{(t)}(:,j)\right\|_2 \tag{38}$$

According to Eq. (21), we get

$$\begin{aligned} & tr\left(\left(\mathbf{W}^{(t+1)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t+1)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t+1)}\right) \\ & \qquad \leq tr\left(\left(\mathbf{W}^{(t)}\right)^H\mathbf{\Phi}_w\mathbf{E}^{(t)}\mathbf{\Phi}_w{}^H\mathbf{W}^{(t)}\right) \end{aligned} \tag{39}$$

Eq. (39) shows that the objective of *e*-LDA-L21 in each iteration is monotonically non-increasing. Then combining the solution process in Eq. (24) and Eq. (25), we have that *Algorithm* 1 converges to a local solution of *e*-LDA-L21.

IV. Extension to Euler 2DLDA-L21

A. Objection function

For image recognition, in order to well exploit the spatial information embedded in image pixels, we extend *e*-LDA-L21 to Euler 2DLDA-L21 (*e*-2DLDA-L21) in this section. Prior to formulating *e*-2DLDA-L21, we first introduce the matrix form of cosine distance metric. The matrix form of Eq. (15) is

$$d\left(\mathbf{X}_p,\mathbf{X}_q\right)=\sum_{h,k}\left\{1-\cos\left(\alpha\pi\left(\mathbf{X}_p\left(h,k\right)-\mathbf{X}_q\left(h,k\right)\right)\right)\right\} \tag{40}$$

where $\mathbf{X}_p(h,k)$ and $\mathbf{X}_q(h,k)$ denote the element of the hth-row and kth-column of matrices $\mathbf{X}_p$ and $\mathbf{X}_q$, respectively.

By simple algebra, Eq. (40) becomes

$$\begin{aligned} d\left(\mathbf{X}_p,\mathbf{X}_q\right) &= \sum_{h,k}\left\{1-\cos\left(\alpha\pi\left(\mathbf{X}_p\left(h,k\right)-\mathbf{X}_q\left(h,k\right)\right)\right)\right\} \\ &= \sum_{h,k}\left\|\tfrac{1}{\sqrt{2}}\left(e^{i\alpha\pi\mathbf{X}_p(h,k)}-e^{i\alpha\pi\mathbf{X}_q(h,k)}\right)\right\|_2^2 \\ &= \left\|\mathbf{Z}_p-\mathbf{Z}_q\right\|_F^2 \end{aligned} \tag{41}$$

where

$$\mathbf{Z}_q=\frac{1}{\sqrt{2}}e^{i\alpha\pi\mathbf{X}_q}\in C^{m\times n} \tag{42}$$

is called Euler representation of $\mathbf{X}_q\in R^{m\times n}$.

In *e*-2DLDA-L21, the between-class and within-class scatter matrices are respectively defined as

$$tr\left(\mathbf{Q}^H\mathbf{G}_b\mathbf{Q}\right)=\sum_{i=1}^{c}n_i\left\|\mathbf{Q}^H\left(\bar{\mathbf{Z}}_i-\bar{\mathbf{Z}}\right)\right\|_{L_{2,1}} \tag{43}$$

$$tr\left(\mathbf{Q}^H\mathbf{G}_w\mathbf{Q}\right)=\sum_{i=1}^{c}\sum_{j\in c_i}\left\|\mathbf{Q}^H\left(\mathbf{Z}_j-\bar{\mathbf{Z}}_i\right)\right\|_{L_{2,1}} \tag{44}$$

where $\mathbf{Z}_j$, $\bar{\mathbf{Z}}_i$ and $\bar{\mathbf{Z}}$ are the corresponding Euler representation of $\mathbf{X}_j$, $\bar{\mathbf{X}}_i$ and $\bar{\mathbf{X}}$, which can be obtained by Eq. (42), respectively.

According to the Eq. (43) and Eq. (44), the objective function of *e*-2DLDA-L21 can be rewritten as

$$\mathbf{Q}_{out}=\underset{\mathbf{q}_k^H\mathbf{q}_k=1}{\arg\max}\frac{\left\|\mathbf{Q}^H\mathbf{\Phi}_b\right\|_{L2,1}}{\left\|\mathbf{Q}^H\mathbf{\Phi}_w\right\|_{L2,1}} \tag{45}$$

where $\mathbf{\Phi}_b=\left[n_1\left(\bar{\mathbf{Z}}_1-\bar{\mathbf{Z}}\right),n_2\left(\bar{\mathbf{Z}}_2-\bar{\mathbf{Z}}\right),\cdots,n_c\left(\bar{\mathbf{Z}}_c-\bar{\mathbf{Z}}\right)\right]$, $\mathbf{\Phi}_w=\left[\mathbf{Z}_1-\bar{\mathbf{Z}}_1,\cdots,\mathbf{Z}_j-\bar{\mathbf{Z}}_{c_i},\cdots,\mathbf{Z}_N-\bar{\mathbf{Z}}_c\right]$.

B. Algorithm

Before solving the objective function (45), we first introduce the following equations. Given an arbitrary matrix $\mathbf{A}\in\mathbf{C}^{m\times n}$, according to the definition of L21-norm, we have

$$\begin{aligned} \left\|\mathbf{A}\right\|_{L_{21}} &= \sum_{k=1}^{n}\left\|\mathbf{A}(:,k)\right\|_2 \\ &= \sum_{k=1}^{n}tr(\frac{\mathbf{A}(:,k)\mathbf{A}(:,k)^H}{\left\|\mathbf{A}(:,k)\right\|_2}) \\ &= tr(\mathbf{A}\Lambda\mathbf{A}^H) \end{aligned} \tag{46}$$

where $\Lambda=diag\left(\frac{1}{\left\|\mathbf{A}(:,1)\right\|_2},\frac{1}{\left\|\mathbf{A}(:,2)\right\|_2},...,\frac{1}{\left\|\mathbf{A}(:,\mathrm{n})\right\|_2}\right)$.

According to the Eq. (46), we have

$$\left\|\mathbf{Q}^H\mathbf{\Phi}_b\right\|_{L2,1}=\sum_{k=1}^{c\times n}\left\|\mathbf{Q}^H\mathbf{\Phi}_b\left(:,k\right)\right\|_2=tr\left(\mathbf{Q}^H\mathbf{\Phi}_b\mathbf{D}{\mathbf{\Phi}_b}^H\mathbf{Q}\right) \tag{47}$$

$$\left\|\mathbf{Q}^H\mathbf{\Phi}_w\right\|_{L2,1}=\sum_{k=1}^{N\times n}\left\|\mathbf{Q}^H\mathbf{\Phi}_w\left(:,k\right)\right\|_2=tr\left(\mathbf{Q}^H\mathbf{\Phi}_w\mathbf{E}{\mathbf{\Phi}_w}^H\mathbf{Q}\right) \tag{48}$$

where $\mathbf{D}=diag\left(\frac{1}{\left\|\mathbf{Q}^H\mathbf{\Phi}_b(:,1)\right\|_2},\cdots,\frac{1}{\left\|\mathbf{Q}^H\mathbf{\Phi}_b(:,c\times n)\right\|_2}\right)$, $\mathbf{E}=diag\left(\frac{1}{\left\|\mathbf{Q}^H\mathbf{\Phi}_w(:,1)\right\|_2},\cdots,\frac{1}{\left\|\mathbf{Q}^H\mathbf{\Phi}_w(:,N\times n)\right\|_2}\right)$.

Substituting Eq. (47), Eq. (48) into Eq. (45), the objective function of *e*-2DLDA-L21 becomes

$$\mathbf{Q}_{opt}=\underset{\mathbf{q}_k{}^H\mathbf{q}_k=1}{\arg\max}\frac{tr\left(\mathbf{Q}^H\mathbf{\Phi}_b\mathbf{D}{\mathbf{\Phi}_b}^H\mathbf{Q}\right)}{tr\left(\mathbf{Q}^H\mathbf{\Phi}_w\mathbf{E}{\mathbf{\Phi}_w}^H\mathbf{Q}\right)} \tag{49}$$

Eq. (49) can be converted to

$$\begin{aligned} &\mathbf{Q}_{opt}=\underset{{q_k}^Hq_k=1}{\arg\min}tr\left(\mathbf{Q}^H\mathbf{\Phi}_w\mathbf{E}{\mathbf{\Phi}_w}^H\mathbf{Q}\right) \\ &s.t.\quad \mathbf{Q}^H\mathbf{\Phi}_b\mathbf{D}{\mathbf{\Phi}_b}^H\mathbf{Q}=\mathbf{T} \end{aligned} \tag{50}$$

where $\mathbf{T}$ is a matrix whose elements are constants.

With reference to the solution process of *e*-2DLDA-L21 in Eq. (24), Eq. (25) and Eq. (26), we can derive the Lagrangian function corresponding to the model (50), and then let the result be equal to zero. Finally, we can get

$${\mathbf{\Psi}_b}^{-1}\mathbf{\Psi}_w\mathbf{Q}=\mathbf{Q}\Lambda \tag{51}$$

where $\mathbf{\Psi}_b=\mathbf{\Phi}_b\mathbf{D}{\mathbf{\Phi}_b}^H$ and $\mathbf{\Psi}_w=\mathbf{\Phi}_w\mathbf{E}{\mathbf{\Phi}_w}^H$.

The optimal projection matrix $\mathbf{Q}$ of the objective function (50) consists of the eigenvectors of the Eigen-equation ${\mathbf{\Psi}_b}^{-1}\mathbf{\Psi}_w\mathbf{q}_k=\lambda_k\mathbf{q}_k(k=1,\cdots,d)$ corresponding to the first d smallest eigenvalues except zero.

Finally, we summarize the pseudo code of solving the objective function (50) in **Algorithm 2**.

Algorithm 2: *e*-2DLDA-L21 algorithm

Initialize:

Initialize parameter α, ε, $\mathbf{Q}^{(\mathbf{1})}\in\mathbf{Q}^{m\times d}$ which satisfies ${\mathbf{q}_k}^H\mathbf{q}_k=\mathbf{1}(k=1,2,\cdots,d)$, $t=1$.

Iteration:

1. Calculate $\mathbf{Z}_j\ (j=1,2,\cdots,N)$ by Eq. (42).
2. Calculate $\mathbf{\Psi}_w$ and $\mathbf{\Psi}_b$.
3. Calculate $\mathbf{Q}^t$ by Eq. (51) and select d eigenvectors corresponding to the first d smallest eigenvalues except zero as $\mathbf{Q}^t$.
4. $t=t+1$, go to until converges.

Output: $\mathbf{Q}^t$

V. Experimental results

In this section, we validate our approaches in five face datasets (Extended Yale B, AR, CMU PIE, LFWcrop and SUFR-W) and compare them with the some representative methods such as LDA-L1 [28], ILDA-L1 [30], Wang's method [27], KISSME [10], XQDA [13], and DNS [15], 2DPCA [3], 2DPCA-L1 [24], 2DLDA [5], L1-2DLDA [33]. In our experiments, we use 1-nearest neighbor (1NN) for classification. For

Figure 4. Some samples of one person in the Extend Yale B database.The second row is noised images. The third row is face images+object images (outliers).

Figure 5. Some samples of one person in the CMU PIE database.The second row is noised images. The third row is face images+object images (outliers).

all 1D methods, we empirically set the maximum number of projection vectors to the number of input-data class minus 1, and set the parameter α in *e*-LDA-L21 to 1.0 for AR, CMU PIE, LFWcrop and SUFR-W databases, [0.1,0.3,0.7,0.7] for four groups of experiments on the Extended Yale B database. For all 2D methods, we set the maximum number of projection vectors to 25, and set the parameter α in *e*-2DLDA-L21 to 1.0 for all databases. All the experiments are performed on the windows-7 operating system (Intel Core i6-4770 CPU @ 3.40 GHz 8 GB RAM).

A. Databases

The Extended Yale B database [46] includes 2414 face images that were sampled from 38 persons under frontal-view with different illuminations. In the Extended Yale B dataset, most classes (person) have 64 images except for the 11th, 12 th, 13th, 14th, 15th, 16th and 17th that have 60, 59, 60, 63, 62, 63, and 63 images, respectively. Each image was resized to be 32×32 pixels. 14 images per person were randomly selected and placed two types of noise, respectively. One is the black and white dots. The ratio of the pixels of noise to number of image pixels is intervenient 0.05 to 0.15. Another is the 16×16 pixels block of object images. Thus, we got two new galleries for the experiments. Figure 4 show some images of one person, where the second and third rows denote the first and second types of noised images, respectively.

Figure 6. Some samples in the LFWcrop database.

Figure 7. Some samples in the SUFR-W database.

AR dataset [47] contains over 4000 color face image of 126 people, including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by two weeks). Each session contained 13 color images, which include 6 images with occlusion and 7 full facial images with different facial expressions and lighting conditions. In this dataset, we converted each color image to gray image by Matlab function rgb2gray and then normalized it to 50×40 pixels [3].

The CMU PIE dataset [48] has 2856 frontal-face images sampled from 68 persons (classes) with various illumination. In the PIE dataset, each image was resized to 32×32 pixels. 10 images per person were randomly selected and placed the two same types of noise as that in the Extend Yale B database. Thus, we got two galleries for the experiments. Figure 5 shows some images of one person, where the second and third rows denote the first and second types of noised images, respectively.

LFWcrop dataset [49] is a cropped version of the Labeled Faces in the Wild (LFW) dataset. In the vast majority of images, almost all of the backgrounds are omitted. The extracted area was then scaled to 64x64 pixels. Figure 6 shows some images of one person in the LFWcrop. The cropped faces in LFWcrop exhibit real-life conditions, including misalignment, scale variations, in-plane as well as out-of-plane rotations.

SUFR-W dataset [49] is a new unconstrained natural image dataset which contains both grayscale and color images. In this dataset, we converted each color image to gray image and then normalized it to 64x64 pixels. Figure 7 shows some images of one person in the SUFR-W.

B. Experiments for 1-D methods

In the AR dataset, we randomly select 13 images per person for training and the remaining images for testing, and then respectively use the aforementioned four 1-d methods to extract features. We repeat this process 10 times. Table III lists the top average recognition accuracy with corresponding standard deviation (std) and average running time with corresponding standard deviation (std).

In the Extended Yale B dataset, we conduct two group experiments for each type of noised image, respectively. In the first group experiment, we randomly choose 32 images per person, which include 18 noise-free images and 14 noised images, for training, and the remaining images are viewed as testing images. In the second group experiment, we randomly choose 32 images including 7 noised images per person for training and the remaining images for testing. All of

the aforementioned experiments are repeated 10 times. Table IV lists the top average recognition accuracy and standard deviation.

In the CMU PIE dataset, we do the same experiments as those in the Extended Yale B databaset for each type of noised image. In the first group experiment, 21 images per class, including 11 noise-free images and 10 noised images are randomly selected for training, and the remaining images for testing. In the second group experiment, we randomly select 21 images (5 noised and 16 noise-free images) per person for training and the remaining images are viewed as testing data. All of the aforementioned experiments are repeated 10 times in the experiments. Table V lists the top average recognition accuracy and standard deviation.

In the LFWcrop dataset, we choose person who has more than 20 photos but less than 100 photos as the sub-database, which contains 57 classes and 1883 images. For each person, we randomly choose ninety percent of all images for training, and the remaining images for testing. We repeat this process 10 times. Table VI lists the top average recognition accuracy and standard deviation.

In the SUFR-W dataset, we choose person who has more than 50 photos but less than 60 photos as the sub-database, which contains 54 classes and 2921 images. For each person, we randomly choose ninety percent of all images for training, and the remaining images for testing. We repeat this process 10 times. Table VI lists the top average recognition accuracy and standard deviation.

Figure 8 plots the average classification curves of four methods versus the number of projection vectors on the AR database, Extended Yale B database and CMU PIE database. Figure 9 plots the average classification curves of four methods versus the number of projection vectors on the LFWcrop and SUFR-W databases.

TABLE III
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) WTIH CORRESPONDING STANDARD DEVIATION(STD) AND AVERAGE TRAINING TIME WITH CORRESPONDING STANDARD DEVIATION (STD)ON THE AR DATABASE.

Methods	accuracy	std	running time	std
e-LDA-L21	**98.87**	**0.46**	**9.4465**	**0.0757**
ILDA-L1	96.65	0.56	75.6425	1.7764
LDA-L1	92.72	1.37	14.2829	0.5127
Wang's	97.52	0.51	46.4280	1.9850
KISSME	97.86	0.38	8.3278	0.6236
XQDA	98.46	0.29	8.7592	0.4930
DNS	97.78	0.64	6.5290	2.5289

As can be seen in the aforementioned experiments, we have

- LDA-L1 is inferior to ILDA-L1, this is because that LDA-L1 solves the optimal projection vectors by greedy strategy and the obtained solution does not best optimize the corresponding trace ratio objective function, while ILDA-L1 avoids this problem by non-greedy algorithm. Wang's method is superior to the other two ℓ_1-norm based methods. The reason may be that Wang's method takes into account the relationship between projection vectors which is important for classification. LDA-L1 and ILDA-L1 are inferior to *e*-LDA-L21. This is due to the fact that LDA-L1 and ILDA-L1 measure the similarity in the image pixel space, which is sensitive to outliers and illumination. Another reason is that they do not well characterize both the geometric structure of data and nonlinear features.
- KISSME considers two independent generation processes for observed commonalities of similar and dissimilar pairs which utilize the statistical characteristics of the data itself. XQDA performs better than the aforementioned four algorithms. Here, XQDA is used in the single-view scenario, which is equivalent to the quadratic discriminant analysis (QDA) algorithm. The quadratic discriminant analysis algorithm is similar to the linear discriminant analysis algorithm. The difference is that

TABLE IV
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE EXTENDED YALE B DATASET

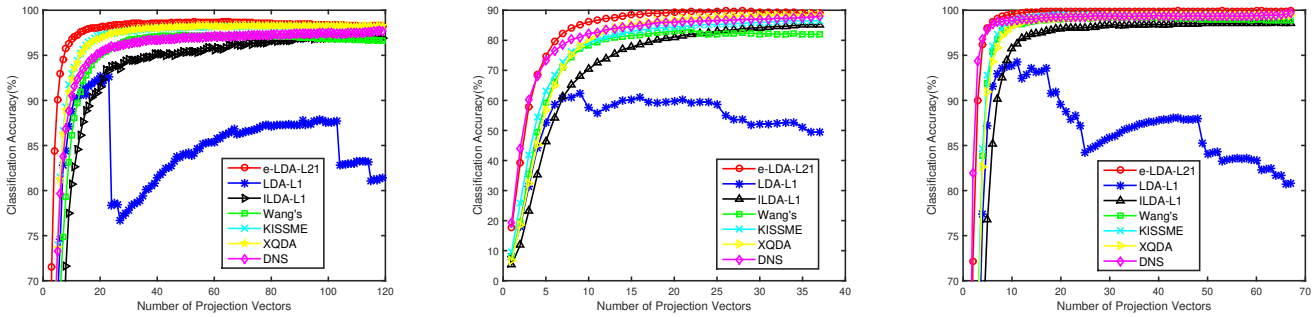
Methods	Black and white dots		Image outliers	
	$Exp1-1$	$Exp1-2$	$Exp2-1$	$Exp2-2$
e-LDA-L21	**90.10±0.69**	**87.90±0.77**	**90.40±0.62**	**87.49±1.12**
ILDA-L1	85.18±0.78	85.40±0.87	85.36±1.08	82.13±1.09
LDA-L1	69.99±2.47	71.12±2.04	67.53±6.53	68.12±6.41
Wang's	85.64±0.59	85.68±0.76	85.10±5.09	84.08±1.41
KISSME	85.74±0.93	86.95±0.64	85.93±0.55	83.69±0.95
XQDA	87.97±0.61	86.81±0.74	88.73±0.73	86.65±1.02
DNS	90.04±0.89	85.01±0.95	88.03±1.41	83.28±1.31

TABLE V
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE CMU PIE DATASET

Methods	Black and white dots		Image outliers	
	$Exp1-1$	$Exp1-2$	$Exp2-1$	$Exp2-2$
e-LDA-L21	**99.99±0.02**	**99.69±0.17**	**99.94±0.09**	98.20±0.29
ILDA-L1	99.05±0.38	98.20±0.34	98.70±0.63	95.22±0.43
LDA-L1	89.43±1.26	89.10±0.58	90.04±1.43	88.50±1.30
Wang's	99.28±0.39	98.51±0.27	99.38±0.48	97.32±0.43
KISSME	99.01±0.30	99.15±0.16	99.29±0.44	97.03±0.47
XQDA	99.22±0.36	98.87±0.23	99.55±0.21	**98.51±0.33**
DNS	99.75±0.21	99.18±0.23	99.68±0.34	96.43±0.57

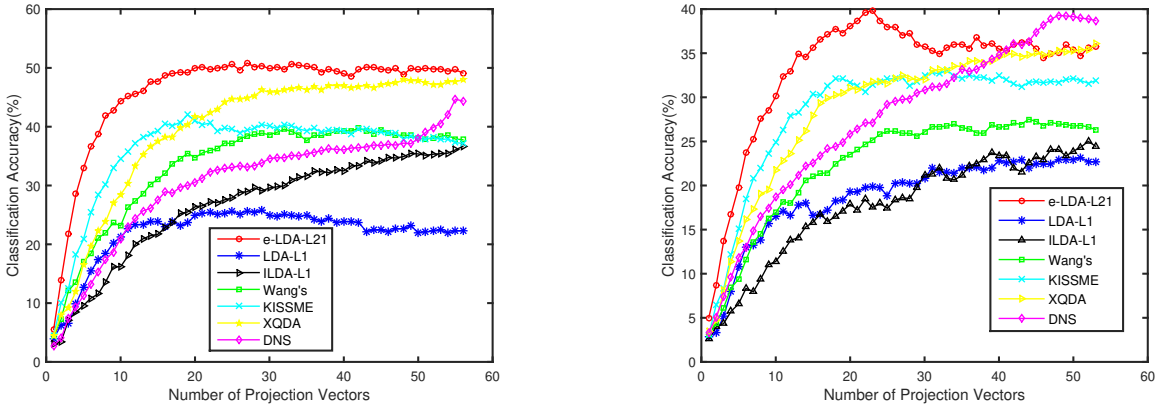
TABLE VI
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE LFWCROP AND SUFR-W DATASETS

Methods	LFWcrop		SUFR-W	
	accuracy	std	accuracy	std
e-LDA-L21	**54.55**	**3.33**	**40.11**	**2.60**
ILDA-L1	40.55	2.76	27.78	1.56
LDA-L1	28.58	2.79	24.67	2.02
Wang's	43.90	2.51	31.98	1.80
KISSME	44.08	2.06	31.89	2.38
XQDA	49.59	3.12	36.78	1.81
DNS	45.05	3.09	39.81	2.06



(a) (b) (c)

Figure 8. Average classification accuracy versus the projection vectors number of different methods on three databases. (a) AR, (b)Exp2-1 on the Extended Yale B, (c)Exp2-1 on the CMU PIE.



(a) (b)

Figure 9. Average classification accuracy versus the projection vectors number of different methods on two databases. (a) LFWcrop, (b)SUFR-W.

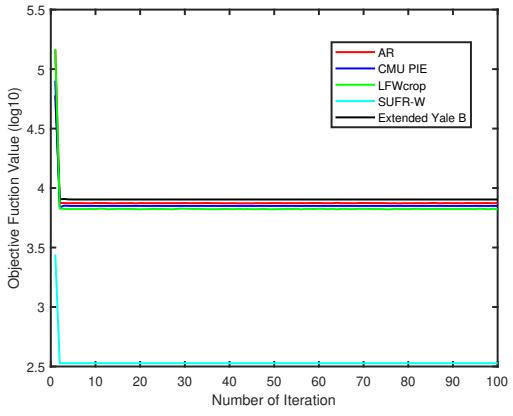
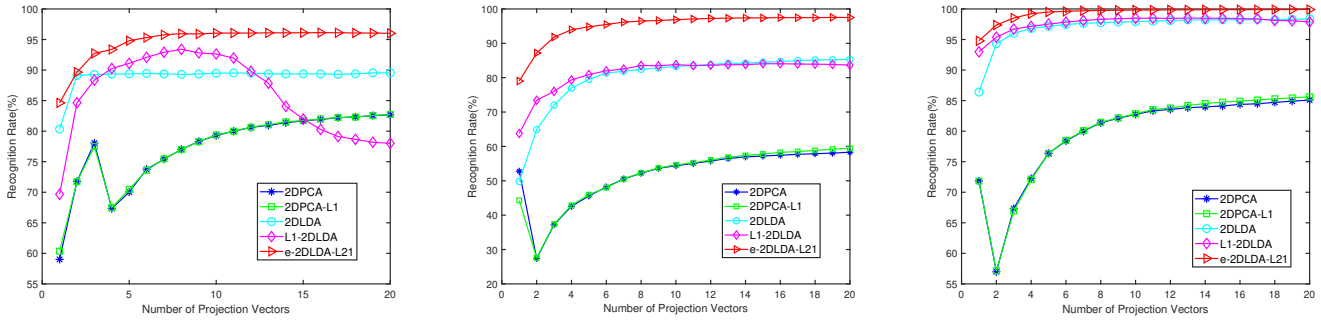


Figure 10. Convergence curves of *e*-LDA-L21 on the AR, CMU PIE, Extended Yale B, LFWcrop and SUFR-W databases.

when the covariance matrices of different classification samples are different, quadratic discrimination analysis should be used. The performance of the DNS algorithm is not stable, which may be due to the images of the same person are collapsed into a single point in the null space, thus leading to overfitting.

- Figures 8 and 9 illustrate that our method *e*-LDA-L21 is superior to the other four methods and can obtain the best recognition accuracy among the four methods with the same number of projection vectors. The reason may be that *e*-LDA-L21 calculates the dissimilarity between data in the Euler space which can suppress outliers and improve the separability of data to obtain a large margin for different classes. Moreover, as the aforementioned analysis in section III, $\ell_{2,1}$-norm enlarges the role of small between-class distance in the criterion function. This also helps encode the discriminant information.
- Table VI shows that under uncontrolled scenarios (with non precise face crops), the performance of our method degrades remarkably, compared with the performance in other datasets. The reason may be that subspace-based learning methods are not robust to some complicated conditions including misalignment, illumination variations, in-plane as well as out-of-plane rotations. But our method is still remarkably superior to the other LAD based robust subspace methods LDA-L1, ILDA-L1 and Wang's.
- Figure 10 indicates that our method monotonically decreases the value of the objective function and has a good



(a) (b) (c)

Figure 11. Average classification accuracy versus the projection vectors number of different methods on three databases. (a) AR, (b)Exp1-2 on the Extended Yale B, (c)Exp1-2 on the CMU PIE.

convergence. Table III illustrates that our method is faster than the other LDA-based methods, but slightly slower than the other three comparison algorithms. The reason may be that all LDA-based methods including ours need to iteratively solve the projection vectors, but our method converges very quickly.

C. Experiments for 2-D methods

In this section, we do the same experiments as those for the 1-D methods in the Extended Yale B, AR, and CMU PIE databases to validate *e*-2DLDA-L21, and compare with some two-dimensional methods. Tables VIII, IX and VII list the average recognition accuracy of each method and corresponding standard deviation (Std) on the Extended Yale B, CMU PIE and AR databases, respectively. Figure 11 plots the average classification accuracy of each approach versus the number projection vectors, and Figure 12 shows the convergence curve of *e*-2DLDA-L21 in these databases.

TABLE VII
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(STD) ON THE AR DATABASE.

Methods	accuracy	std
e-2DLDA-L21	**96.08**	**0.60**
2DPCA	82.71	0.94
2DPCA-L1	82.72	0.95
2DLDA	89.93	0.68
L1-2DLDA	94.22	0.79

As can be seen in the aforementioned experimental results for tow-dimensional methods, we have that

- Discriminant methods are consistently superior to 2DPCA and 2DPCA-L1. This is probably because that 2DPCA and 2DPCA-L1 are unsupervised and do not well encode the discriminant information. L1-2DLDA is overall superior to traditional 2DLDA. The reason is that ℓ_1-norm is robust to outliers, compared with squared ℓ_2-norm. However, 2DLDA is superior to L1-2DLDA in some experiments. This is probably because that L1-2DLDA does not relate to scatter matrices that well characterize geometric structure of data, while 2DLDA does.

TABLE VIII
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE EXTENDED YALE B DATABASES

Methods	Black and white dots		Image outliers	
	$Exp1-1$	$Exp1-2$	$Exp2-1$	$Exp2-2$
e-2DLDA-L21	**98.03±0.29**	**97.69±0.30**	**97.77±0.30**	**94.34±0.59**
2DPCA	55.64±1.44	58.33±0.69	51.58±0.80	51.48±0.85
2DPCA-L1	53.77±1.07	59.44±0.60	51.61±0.72	51.43±0.88
2DLDA	81.81±0.56	85.41±0.55	85.16±1.18	84.71±1.11
L1-2DLDA	82.25±0.52	84.70±0.46	83.54±1.08	81.42±1.23

TABLE IX
THE TOP AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE CMU PIE DATABASES

Methods	Black and white dots		Image outliers	
	$Exp1-1$	$Exp1-2$	$Exp2-1$	$Exp2-2$
e-2DLDA-L21	**99.94±0.15**	**99.71±0.11**	**99.99±0.02**	**97.30±0.21**
2DPCA	77.58±0.45	85.09±0.72	77.54±1.49	76.50±1.58
2DPCA-L1	79.09±0.50	85.61±0.57	77.38±1.49	76.50±1.59
2DLDA	97.49±0.72	98.59±0.30	99.11±0.57	95.78±0.67
L1-2DLDA	99.20±0.42	98.42±0.30	99.24±0.47	95.24±0.68

- Our approach *e*-2DLDA-L21 is superior to the other two-dimensional methods. This is probably because that our method well reveals nonlinear features and discriminant features by Euler transform and $\ell_{2,1}$-norm. Another reason is that our method relates to scatter matrices and well exploits geometric structure of data. Figure 12 shows that our proposed algorithm monotonically decreases the value of the objective function and has a good convergence.
- Compared to the experiments including 1D methods and two-dimensional methods, we have that two-dimensional methods are superior to the corresponding 1D discriminant methods in Extended Yale B. The reason may be that two-dimensional methods encodes the spatial geometric structure embedded in pixels of image. However, in the

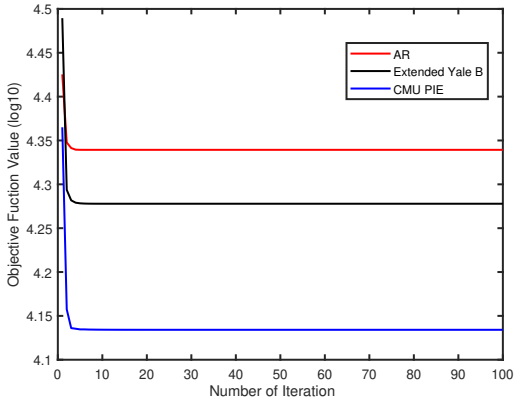


Figure 12. Convergence curves of *e*-2DLDA-L21 on three databases.

TABLE X
THE AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE EXTENDED YALE B DATASET

Methods	Black and white dots		Image outliers	
	$Exp1-1$	$Exp1-2$	$Exp2-1$	$Exp2-2$
e-LDA-L21	**90.10±0.69**	**87.90±0.77**	**90.40±0.62**	**87.49±1.12**
LDA-L21	87.94±0.62	86.58±0.67	87.73±0.72	84.40±1.12
e-LDA-L1	84.58±0.89	84.23±0.97	82.53±0.91	80.36±0.87
LDA-L1	69.99±2.47	71.12±2.04	67.53±6.53	68.12±6.41

other datasets, two-dimensional methods are inferior to the corresponding 1D discriminant methods. The reason may be that two-dimensional methods would be confronted with the heteroscedastic problem and the problem would be more serious for two-dimensional methods than the one for 1D-dimensional methods [52].

D. Discussion

1) : $\ell_{2,1}$-norm and Euler transform

To discuss the effects of Euler transform and $\ell_{2,1}$-norm for discriminant analysis respectively, we compared *e*-LDA-L1, *e*-LDA-L21, LDA-L1 and LDA-L21 on the Extended Yale B, LFWcrop and SUFR-W databases. In the experiments, we selected the aforementioned training images and corresponding testing images on these datasets. Tables X and XI list the average recognition accuracy of each approach and corresponding standard deviation (Std), respectively. Figure 13 plots the average classification accuracy of each approach versus the number projection vectors on the Extended Yale B, LFWcrop and SUFR-W databases, respectively.

As can be seen in the aforementioned experiments, we have

- Discriminant methods with Euler transformation (*e*-LDA-L21 and *e*-LDA-L1) are superior to the corresponding methods without Euler transformation (LDA-L21 and LDA-L1). The reason may be that *e*-LDA-L21 and *e*-LDA-L1 calculate the dissimilarity and similarity between data in the Euler space which can suppress outliers and reveal nonlinear features. Moreover, Euler space can help improve the separability of data and obtain a large margin for different classes.

TABLE XI
THE AVERAGE CLASSIFICATION ACCURACY(%) AND CORRESPONDING STANDARD DEVIATION(%) ON THE LFWCROP AND SUFR-W DATASETS

Methods	LFWcrop		SUFR-W	
	accuracy	std	accuracy	std
e-LDA-L21	**54.55**	**3.33**	**40.11**	**2.60**
LDA-L21	45.18	1.77	33.44	1.81
e-LDA-L1	49.08	1.48	39.85	2.53
LDA-L1	28.58	2.79	24.67	2.02

- Methods, which employ $\ell_{2,1}$-norm as distance metric, are superior to ℓ_1-norm based methods. The reason is that $\ell_{2,1}$-norm based methods reveal within-class and between-class scatters, while ℓ_1-norm based methods do not. Moreover, $\ell_{2,1}$-norm enlarges the role of small between-class distance. This helps get a large margin which is important for classification. However, it is unclear whether ℓ_1-norm plays the same role.
- Table XI shows that LDA-L21 is inferior to *e*-LDA-L1. The reason may be that LDA-L21 does not well encode nonlinear discriminant features. It also illustrates that Euler transform well revels nonlinear features.
- Figure 13 illustrates that *e*-LDA-L21 is superior to the other three methods and obtains the best recognition accuracy among the four methods under the same number of projection vectors.

2) : The sensitivity analysis of the parameter α.

In order to well illustrate the influence of parameter α in our model, we added some experiments on the Extended Yale B, AR and SUFR-W databases. In the experiments, we selected the aforementioned training images and corresponding testing images on these datasets. Figure 14 plots the curves of classification accuracy versus parameter α on three databases, respectively, and also marks the maximum value for each group experiments.

As can be seen in the aforementioned experiments, we have

- From Figure 14 (a), we can see that α has a large influence for the classification accuracy of our method under the different group experiments on the Extend Yale B database. It is unable to select the same α in this dataset. Thus, we select different α for different group experiments on the Extended Yale B dataset.
- Figure 14 (b) illustrates that α has little influence for the classification accuracy of our method on the AR and SUFR-W databases. When α is in [0.9 1.1] interval, our method overall has good performance. Thus, we set α as 1.0 in the experiments on the AR and SUFR-W datasets.

VI. CONCLUSION

We present a robust supervised approach, namely Euler LDA-L21(*e*-LDA-L21), for dimensionality reduction. *e*-LDA-L21 maps image onto Euler space and employs $\ell_{2,1}$-norm as distance metric to measure within-class and between-class scatters in the criterion function. It is similar to kernel LDA, but they are essentially different. The the main difference between them is that, *e*-LDA-L21 maps the original image

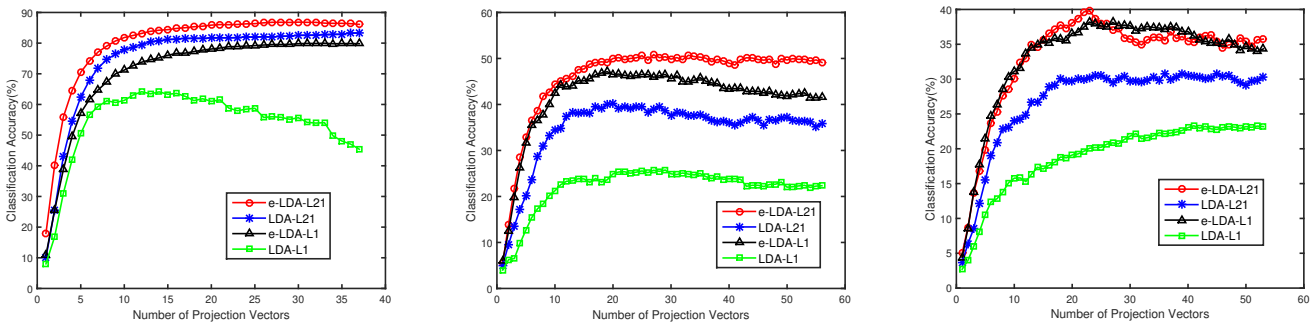


Figure 13. Average classification accuracy versus the projection vectors number of different methods on three databases. (a)Exp2-2 on the Extended Yale B, (b)LFWcrop, (c)SUFR-W.

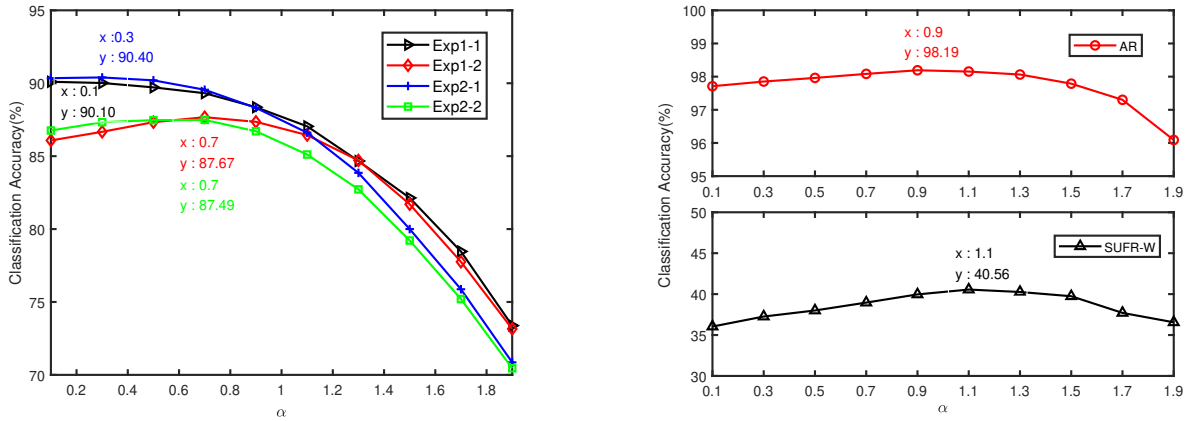


Figure 14. The curve of classification accuracy versus the parameter α. (a) Extended Yale B, (b)AR and SUFR-W.

space to an explicit Euler feature space and does not increase the dimensionality of features, while kernel LDA does not. Compared with most existing robust LDA methods, our method not only is robust to outliers but also helps obtain a large margin in the low-dimensional space. Thus, our method encodes discriminant information. Experiment results illustrate that our proposed algorithms have a good convergence and our methods are superior to the other robust methods for image classification.

For feature extraction, rotational invariance is one of the important properties [50] [51]. As the aforementioned analysis, from the norm point of view, $\ell_{2,1}$-norm and traditional squared ℓ_2-norm have no essential difference, thus, our method retains traditional LDA's rotation invariance. We will study this problem in future work.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and AE for their constructive comments and suggestions, which improved the paper substantially. We also thank Prof. X. Gao for polishing our paper. Finally, we wold like to thank Prof. S. Liao and Dr. Zhang for providing the codes of XQDA and DNS.

REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for recognition", *J. Cognitive Neurosci.*, vol. 3, no. 7, pp. 71-86, winter. 1991.

[2] P. N. Belhumeur, J. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711-720, Jul. 1997.

[3] Jian Yang, D. Zhang, A. F. Frangi and Jing-yu Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131-137, Jan. 2004.

[4] H. Lu, K. Plataniotis and A. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects", *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 8-39, Jan. 2008.

[5] J. Yang, D. Zhang, Y. Xu and J. Yang, "Two-dimensional discriminant transform for face recognition", *Pattern Recognit.*, vol. 38, no. 7, pp. 1125-1129, Jul. 2005.

[6] J. Ye, J. Ravi Janardan and L. Qi, "Two-dimensional linear discriminant analysis", *In proc. Advances in Neural Information Processing Systems*, pp. 1569-1576, 2005.

[7] L. Zhang, Q. Gao and D. Zhang, "Directional independent component analysis with tensor representation", *Proc. IEEE Int. Computer Vision and Pattern Recogni.*, pp. 1-7, 2008.

[8] Q. Gao, F. Gao, H. Zhang, X. Hao, and X. Wang, "Two-dimensional maximum local variation based on image Euclidean distance for face recognition", *IEEE Trans. Image Processing*, vol. 22, no. 10, pp.3807-3817, 2013.

[9] T. Li, M. Li, Q. Gao, and D. Xie, "F-norm distance metric based robust 2DPCA and face recognition", *Neural Networks*, vol. 94, no. 10, pp. 204-211, 2017.

[10] P. Roth, P. Wohlhart, and M. Hizer, "Large Scale Metric Learinng from Equivalence Constraints", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288-2295, 2012.
[11] N. Martinel, C. Micheloni, G. L. Foresti, "Kernelized Saliency-based Person Re-Identification through Multiple Metric Learning", *IEEE Transactions on Image Processing*, vol. 24, no.12, pp. 5645-5658, 2015.
[12] Z. Li, S.Chang, F. Liang, T.S. Huang, L. Cao, and J.R. Smith,"Learning locally-adaptive decision functions for person verification", *Computer Vision and Pattern Recognition*, vol. 9 no. 4 pp. 3610-3617, 2013.
[13] S. Liao, Y. Hu, X. Zhu, "Person Re-identification by Local Maximal Occurrence Representation and Metric Learning", *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 8, no. 4, pp. 2197-2206, 2015.
[14] S. Liao , Z. Li, "Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification", *IEEE International Conference on Computer Vision* pp. 3685-3693, 2015.
[15] L. Zhang, T. Xiang, and S.Gong, "Learning a Discriminative Null Space for Person Re-identification", *IEEE Conference on Computer Vision and Pattern Recognition*, pp:1239-1248, 2016.
[16] N. Martinel, A. Das, C. Micheloni, and A. Roy-Chowdhury, "Temporal model adaptation for person re-identification", *European Conference on Computer Vision* (ECCV), pp. 858-877, 2016.
[17] X. Zhu, X. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics", *International Joint Conference on Artificial Intelligence*, pp:3552-3558,2016.
[18] L. Luo, J. Yang, J. Qian, T. Ying, and G. Lu, "Robust image regression based on the extended matrix variate power exponential distribution of dependent noise", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 9, pp. 2168-2182, 2017.
[19] Z. Zhang, F. Li, M. Zhao, L. Zhang and S. Yan, "Robust Neighborhood Preserving Projection by Nuclear/L2,1-Norm Regularization for Image Feature Extraction", *IEEE Trans. on Image Processing*, vol. 26, no. 4, pp. 1607-1622, 2017.
[20] Q. Ke and T. Kanade, "Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming", *Proc. IEEE Int. Computer Vision and Pattern Recogni.*, pp. 592-599, 2005.
[21] N. Kwak, "Principal component analysis based on L1-norm maximization", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, Jun. 2008.
[22] X. Li, Y. Pang and Y. Yuan, "L1-norm-based 2DPCA", *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1170-1175, Jan. 2010.
[23] F. Nie, H. Wang, C. H. Ding, D. Luo and H. Huang, "Robust principal component analysis with non-Greedy l1-Norm maximization", *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 1433-1438, 2011.
[24] R. Wang, F. Nie, X. Yang, F. Gao and M. Yao, "Robust 2DPCA with non-greedy L1-norm maximization for image analysis", *IEEE Trans. Cybernetics*, vol. 45, no. 5, pp. 1108-1112, May. 2015.
[25] N. Kwak, "Principal Component Analysis by Lp-Norm Maximization", *IEEE Trans. on Cybern.*, vol. 44, no. 5, pp. 594-609, Apri. 2014.
[26] F. Zhang, J. Yang, J. Qian and Y. Xu, "Nuclear norm-based 2-DPCA for extracting features from images", *IEEE Trans. Neural Networks and Learning Systems*, Vol. 26, no. 10, pp. 2247-2260, Oct. 2015.
[27] H. Wang, F. Nie and H. Huang, "Robust distance metric learning via simultaneous L1-norm minimization and maximization", *Proc. Int. Conf. Mach. Learn.*, pp. 1836-1844, 2014.
[28] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm maximization", *IEEE Trans. Image Processing*, vol. 22, no. 8, pp. 3018-3027, Aug. 2013.
[29] W. Zheng, Z. Lin and H. Wang, "L1-Norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 793-805, Apr. 2014.
[30] X. Chen, J. Yang and Z. Jin, "An improved linear discriminant analysis with L1-norm for robust feature extraction", *Proc. Int. Conf. Pattern Recogni.*, pp. 1585-1590, 2014.
[31] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, A non-greedy algorithm for L1-norm LDA, *IEEE Trans. Image Processing*, vol. 26, no. 2, pp. 684-695, 2017.
[32] Q. Wang, Q. Gao, D. Xie, X. Gao, and Y. Wang, "Robust DLPP With Nongreedy L1-Norm Minimization and Maximization", *IEEE Trans. on Neural Networks and Learning Systems*, vol. 29. no. 3, pp. 738-743, 2018.
[33] C. Li, Y. Shao and N. Deng, "Robust L1-norm two-dimensional linear discriminant analysis", *Neural Networks*, vol. 65, pp. 92-104, May. 2015.
[34] Q. Gao, L. Ma, Y. Liu, X. Gao, F. Nie, "Angle 2DPCA: a new formulation for 2DPCA", *IEEE Transactions on Cybernetics*, vol. 48, no. 5, pp. 1672-1678, 2018.
[35] C. Ding, D. Zhou, X. He and H. Zha, "R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization", *Proc. Int. Conf. Mach. Learn.*, pp. 281-288, 2006.
[36] Q. Wang, Q, Gao, X.Gao, and F. Nie, "Optimal mean two-dimensional principal component analysis with F-norm minimization", *Pattern recognition*, vol. 68, pp. 286-294, 2017.
[37] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You and Y. Tang, "Group sparse multiview patch alignment framework with view consistency for image classification", *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3126-3137, Jul. 2014.
[38] F. Nie, H. Huang, X. Cai and C. Ding, "Efficient and robust feature selection via joint L2,1-norms minimization", *Advances in Neural Information Processing Systems*, pp. 1813-1821, 2010.
[39] X. Shi, Y. Yang, Z. Guo and Z. Lai, "Face recognition by sparse discriminant analysis via joint L2,1-norm minimization", *Pattern Recognit.*, vol. 47, pp. 2447-2453, Jul. 2014.
[40] W. Wong, Z. Lai, Y. Xu, J. Wen and C. Ho, "Joint tensor feature analysis for visual object recognition", *IEEE Trans. Cybernetics*, vol. 45, no. 11, pp. 2425-2436, Nov. 2015.
[41] C. Hou, F. Nie, D. Yi and Y. Wu, "Feature selection via joint embedding learing and sparse regression", *Proc. 22nd Int. Joint Conf. Artif. Intell.*, pp. 1324-1329, 2011.
[42] A. J. Fitch, A. Kadyrov, W. J. Christmas, and J. Kittler, "Fast robust correlation", *IEEE Trans. Image Processing*, vol. 14, no. 8, pp. 1063-1073, 2005.
[43] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Euler principal component analysis", *International Journal of Computer Vision*, vol. 101, no. 3, pp. 498-518, 2013
[44] M. H. Nguyen and F. Torre, "Robust kernel principal component analysis", *In Advances in Neural Information Processing Systems*, pp. 1185C1192, 2009.
[45] Fukunaga K, *Introduction to statistical pattern recognition*, Academic press, 2013.
[46] A. S. Georghiades, P. N. Belhumeur and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643-660, Jun. 2001.
[47] A. M. Martinez, "The AR face database", CVC Technical Report, 24, 1998.
[48] T. Sim, S. Baker and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) database", *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, Washington, DC, pp. 46-51, 2002.
[49] J. Leibo, Q. Liao, T. Poggio, "Subtasks of unconstrained face recognition", *VISAPP*, vol. 2, pp. 113-121, 2014.
[50] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference", *Proc. Int. Conf. Biometrics*, pp.199-208, 2009.
[51] Q. Liao, J. Leibo, and T. Poggio, "Learning invariant representations and applications to face verification", *proc. Advances in Neural Information Processing Systems*, pp. 3057-3065, 2013.
[52] W. Zheng, J. Lai, and Stan Z. Li, "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?" *Pattern Recognition*, vol. 41, no. 7, pp. 2156-2172, 2008.

Shuangli Liao received the B.Eng. degree in communication engineering from China University of Geosciences, Wuhan, China, in 2016. She is currently working toward the Ph.D. degree in telecommunications engineering in Xidian University, Xian, China. Her research interests include pattern recognition and machine learning.

Quanxue Gao received the B. Eng. degree from xi'an Highway University, Xi'an, China, in 1998, the M.S. degree from the Gansu University of Technology, Lanzhou, China,in 2001, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an China, in 2005. He was a associate research with the Biometrics Center, The Hong Kong Polytechnic University, Hong Kong from 2006 to 2007. From 2015 to 2016, he was a visiting scholar with the department of computer science, The University of Texas at Arlington, Arlington USA. He is currently a professor with the School of Telecommunications Engineering, Xidian University, and also a key member of State Key Laboratory of Integrated Services Networks. He has authored 40 technical articles in refereed journals and proceedings, including IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, CVPR, AAAI, and IJCAI. His current research interests include pattern recognition and machine learning.

Zhaohua Yang received the B.S. degree in precision instruments and machinery from Harbin Institute of Technology, Harbin, China, in 1998, the M.S. degree in control theory and engineering from Lanzhou University of Technology, Lanzhou, China, in 2001, and the Ph.D. degree in precision instruments and machinery from Harbin Institute of Technology in 2004. She was Visiting scholar from Washington University in St. Louis in 2012. She is currently an Associate Professor in the School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing, China. She has authored 50 technical articles in refereed journals and proceedings including IEEE Transactions on Industrial ElectronicsInstrument and Measurement and etc. Her current research interests include ghost imaging, Pattern recognition, and engineering applications.

Fang Chen received the B. Eng. Degree in Communication Engineering from Lanzhou Jiaotong University, China, in 2014. She is currently a M.S. degree candidate in traffic information engineering and control in Xidian University, China. Her research interests include pattern recognition, and dimensionality reduction and face recognition.

Feiping Nie received the Ph.D.degree in computer science from Tsinghua University, Beijing,China, in 2009. He was respectively a Post Doctoral Research Associate, Research Assistant Professor,and Research Professor with the University of Texas at Arlington, Arlington, TX,USA, from 2009 to 2015.He is currently a professor of Northwestern Polytechnical University, Xi'an China. He has authored 160 technical articles in refereed journals and proceedings, including IEEE Trans. on Pattern Analysis and Machine Intelligence,International Journal of Computer Vision, IEEE Trans. on Image Processing, IEEE Trans. Neural Networks and Learning Systems, IEEE Trans.Cybernetics, IEEE Trans. Knowledge and Data Engineering, ICCV, CVPR, ICML, AAAI, IJCAI, and NIPS. His current research interests include machine learning and its application fields, such as pattern recognition, data mining,computer vision,image processing, and information retrieval.

Jungong Han is a tenured Associate Professor of Data Science Institute at Lancaster University (LU), UK. In the past 15 years, he has been continuously conducting research in the fields of video analysis, computer vision and machine learning, and has published over 150 articles in leading journals and prestigious conferences, in which one of the first-authored papers has been cited for more than 1000 times. Dr. Han is the member of the editorial board of several international journals, such as Elsevier Neurocomputing, Springer Multimedia Tools and Applications and IET Computer Vision, and has been (lead) Guest Editors for IEEE T-NNLS and IEEE T-CYB.