This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

Two-Dimensional Quaternion Sparse Discriminant Analysis

Xiaolin Xiao, Yongyong Chen, Yue-Jiao Gong, *Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—Linear discriminant analysis has been incorporated with various representations and measurements for dimension reduction and feature extraction. In this paper, we propose two-dimensional quaternion sparse discriminant analysis (2D-QSDA) that meets the requirements of representing RGB and RGB-D images. 2D-QSDA advances in three aspects: 1) including sparse regularization, 2D-QSDA relies only on the important variables, and thus shows good generalization ability to the out-of-sample data which are unseen during the training phase; 2) benefited from quaternion representation, 2D-QSDA well preserves the high order correlation among different image channels and provides a unified approach to extract features from RGB and RGB-D images; 3) the spatial structure of the input images is retained via the matrix-based processing. We tackle the constrained trace ratio problem of 2D-QSDA by solving a corresponding constrained trace difference problem, which is then transformed into a quaternion sparse regression (QSR) model. Afterward, we reformulate the QSR model to an equivalent complex form to avoid the processing of the complicated structure of quaternions. A nested iterative algorithm is designed to learn the solution of 2D-QSDA in the complex space and then we convert this solution back to the quaternion domain. To improve the separability of 2D-QSDA, we further propose 2D-QSDA$_w$ using the weighted pairwise between-class distances. Extensive experiments on RGB and RGB-D databases demonstrate the effectiveness of 2D-QSDA and 2D-QSDA$_w$ compared with peer competitors.**

***Index Terms*—Discriminant analysis, dimension reduction, quaternion, sparse feature extraction, RGB image, RGB-D image**

I. Introduction

LINEAR discriminant analysis (LDA) [1] is a classical supervised method for dimension reduction and feature extraction. It essentially learns a discriminant subspace where the separability of different projected classes is maximized. Compared with principle component analysis (PCA) [2], LDA takes the class label of the data into consideration. It extracts the discriminant information while ignoring the components that are useless for class separability.

LDA assumes that samples are linearly separable. However, this assumption would probably fail in practical scenarios when the data are of high-dimensionality [3]. To solve this problem, the works in [4], [5] exploited different intra-class geometric measures while preserving the inter-class discrimination. They can learn effective low-dimensional subspace from the high-dimensional ambient space. kernel discriminant analysis (KDA) [3] and individualized KDA (IKDA) [6] adopted the kernel trick such that the linearly inseparable input can be cast into a high-dimensional or infinite-dimensional feature space where the input data are linearly separable. When being applied to image data, these methods necessitate the vectorization of input samples and suffer from the high computation and memory costs for constructing covariance matrices from long vectors. Another intrinsic drawback is that they ignore the spatial structure of the input images. Two-dimensional LDA (2D-LDA) [7] solves these limitations by extracting features from 2D image matrices.

Nowadays, color images have dominated practical applications [8]. Moreover, with the advance of modern cameras, RGB-D images also become popular and the complementary nature of the depth and color information creates new opportunities for computer vision [9]. However, most methods were designed for gray-scale images. When being applied to RGB/RGB-D images, they simply treat different image channels independently, failing to consider the cross-channel correlation. A practical solution is to concatenate the multiple channels into larger vectors or matrices. Nevertheless, the concatenation model captures only the pairwise correlation between image channels, and thus, still suffers performance degradation. Since a multi-channel (*e.g.*, RGB, RGB-D) image is not a simple combination of scalars but rather a vector-valued array, it is important to encode the whole structure of the array to preserve the high order cross-channel correlation. In this respect, the limitation of the concatenation model is derived from the fact that it contains only a fraction of the unfolding matrices which are needed to completely represent a vector-sensor array [10]. To address this issue, tensor representation (TR) [11] and quaternion representation (QR) [12]–[15] were utilized to represent RGB and RGB-D images. Tensor discriminant analysis (TDA) [11] was proposed by representing images as third-order tensors. The quaternion is a four-dimensional hyper-complex number system for representing multi-channel signals which exhibit complex coupling across channels. It can encode the cross-channel relationship of color images, and has been widely used in the literature [10], [16], [17], [17]. Based on QR, quaternion discriminant analysis (QDA) [18] was proposed. It converts color images into high-dimensional quaternion vectors and endures a high computation cost and loses the spatial structure of images.

This work was funded in part by The Science and Technology Development Fund, Macau SAR (File no. 189/2017/A3), and by the Research Committee at University of Macau under Grants MYRG2016-00123-FST and MYRG2018-00136-FST. (Corresponding author: Yicong Zhou.)

X. Xiao is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China and also with the Department of Computer and Information Science, University of Macau, Macau 999078, China (email: shellyxiaolin@gmail.com). Y. Chen and Y. Zhou are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (email: YongyongChen.cn@gmail.com; yicongzhou@um.edu.mo). Y.-J. Gong is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (email: gongyuejiao@gmail.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

Considering practical scenarios, the existence of outliers [19] in the training set or the out-of-sample data in the testing set which are unseen during the training phase [20] always degrades the recognition performance. To alleviate these effects, many algorithms were proposed by exploiting the l_1-norm measurement [21] either on the objective functions (robust algorithms) or as penalty terms (sparse algorithms). Representatives are 2D-LDA based on l_1-norm (2D-LDA-L1) [19] and sparse tensor discriminant analysis (STDA) [22]. However, due to the limited representation capacity of modeling the high order cross-channel relationship, those methods are inadequate in processing RGB and RGB-D images.

Mathematically, the discriminant-based dimension reduction methods end with solving a *Trace Ratio* problem in which two partially coupled objectives are simultaneously optimized. They take the form of maximizing the separability of different projected classes while minimizing the distances of within-class samples. Since the trace ratio problem does not have a closed-form solution, it is commonly transformed into a *Ratio Trace* problem [1], [7]. However, the solution of the ratio trace formulation may deviate from the original objective, and thus, is inexact [23]. Pioneer works proved that an iterative *Trace Difference* formulation can be exploited to solve the ratio trace problem [23]. Yet, how to efficiently solve it with additional constraints is an ongoing work.

Based on these observations, this paper presents two-dimensional quaternion sparse discriminant analysis (2D-QSDA) to extract sparse discriminant features directly from 2D image matrices and proposes an elegant procedure to solve 2D-QSDA. 2D-QSDA naturally takes advantage of QR and 2D-LDA such that it well preserves the high order cross-channel correlation and the spatial structure of images and is computationally efficient. The key ingredient of 2D-QSDA is the sparsity constraints imposed on the projection vectors, which is a trade-off between the original trace ratio function and the sparsity level of project basis. As a result, it improves the generalization ability of 2D-QSDA and makes it robust to the unseen data. In contrast to QDA that can be solved via quaternion eigen-decomposition (QED), 2D-QSDA is formulated as a constrained trace ratio problem and no off-the-shelf tools can be directly applied to solve it. In this work, we first rewrite 2D-QSDA to a constrained trace difference problem, then convert it to a quaternion sparse regression model, and design a nested iterative scheme to find the solution.

Besides, classical discriminant analysis methods find a subspace where the between-class distance of projected classes is maximized while the within-class distance is minimized. Essentially, they impose equal weights to all class pairs [24]. However, this brings problems since the final between-class separability is dominated by the class pairs with large between-class distances, whereas those class pairs with small between-class distances are more difficult to be correctly classified and should be properly treated. Considering this problem, we propose 2D-QSDA$_w$ using the weighted pairwise between-class distances, such that the class pairs with small between-class distances are assigned with relatively large weights to well separate these challenging class pairs. Our contributions are listed as follows.

- We propose a novel quaternion sparse regression (QSR) model to solve the constrained trace difference problem of 2D-QSDA. Including sparse regularization, 2D-QSDA can correctly identify the important variables and ignore the less important ones. Therefore, it is generalizable to classify the data that are unseen during the training phase.
- Without sparsity constraints, the QSR model reverts to a quaternion ridge regression (QRR) model. We mathematically prove that the solution of this QRR model is equivalent to that of two-dimensional QDA. This verifies the validity of integrating sparse regularization into the QRR model to construct the QSR model of 2D-QSDA.
- To solve 2D-QSDA, we reformulate the QSR model to an equivalent complex form to avoid the complicated operations of quaternion derivations. We then design a nested iterative algorithm for optimization, in which a novel sub-algorithm is devised for sparse regularization via the complex-valued alternating direction method of multipliers (complex ADMM). Moreover, a fast complex ADMM algorithm is presented by incorporating a continuation scheme, which is crucial to convergence.
- To improve the separability of 2D-QSDA, we introduce 2D-QSDA$_w$ using a weighting scheme so that the class pairs with small between-class distances can be well separated.
- Taking advantage of the four-dimensional structure of the quaternions, 2D-QSDA and 2D-QSDA$_w$ can efficiently extract features from RGB and RGB-D images. The effectiveness and the generalization ability of 2D-QSDA and 2D-QSDA$_w$ are verified by the applications of color and 3D face recognition.

Please note that the proposed 2D-QSDA has a preliminary conference version [25]. In this journal paper, we have made significant improvements in algorithm design, theoretical analysis, and experimental verification. These will be elaborated in the main body of this paper. To improve the separability of 2D-QSDA, we further propose 2D-QSDA$_w$ using the weighted pairwise between-class distances. Although 2D-QSDA and 2D-QSDA$_w$ follow the same basic optimization strategy with our previous work [26], they have completely different objective functions. Thus, different methods should be developed to formulate their objective functions into regression models. Accordingly, the equivalence between the optimization models and the corresponding regression models should be carefully established. Detailed comparisons with [26] will be provided in Sections III-A and VI-A.

In the rest of this paper, Section II presents the background knowledge. Section III proposes 2D-QSDA and its solution. The model of 2D-QSDA$_w$ is introduced in Section IV. The effectiveness of 2D-QSDA and 2D-QSDA$_w$ is examined in Section V. Then we compare 2D-QSDA with the state-of-the-art quaternion-based models in Section VI. Finally, conclusions are drawn in Section VII.

II. Preliminaries

In this section, we briefly review the quaternion background and several discriminant analysis methods. To clarify the statements, some frequently-used notations are listed in TABLE I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

TABLE I: Summary of notations.

Notation	Description
a, $\mathbf{a}$, $\mathbf{A}$	scalars, vectors, and matrices in real space ($\mathbb{R}$) or complex space ($\mathbb{C}$)
$\dot{a}$, $\dot{\mathbf{a}}$, $\dot{\mathbf{A}}$	scalars, vectors, and matrices in quaternion space ($\mathbb{H}$)
$\bar{(\cdot)}$, $(\cdot)^T$,	conjugate, transpose
$(\cdot)^*$	transpose conjugate
$(\cdot)^{-1}$	inverse of a matrix
$Tr(\cdot)$	trace of a matrix
$Re(\cdot)$	real part of a variable

A. Quaternion Fundamental

The quaternions are a hyper-complex number system that extends the complex number system [27]. A quaternion number ($\dot{q} \in \mathbb{H}$) is composed of one real part and three imaginary parts, and is generally represented as

$$\dot{q} = q_0 + q_1 i + q_2 j + q_3 k, \tag{1}$$

with real coefficients q_0, q_1, q_2, q_3 and an ordered basis $\{1, i, j, k\}$. The addition of quaternions follows that in real space, and the multiplication of quaternions is defined by

$$i^2 = j^2 = k^2 = ijk = -1. \tag{2}$$

$$ij = -ji = k, \quad jk = -kj = i, \quad ki = -ik = j. \tag{3}$$

The above rules make the multiplication of two quaternion numbers non-commutative and this complicates the processing of quaternions. Besides, the conjugate and norm of a quaternion number are defined as $\bar{\dot{q}} = q_0 - q_1 i - q_2 j - q_3 k$ and $|\dot{q}| = \sqrt{\dot{q}\bar{\dot{q}}} = \sqrt{\bar{\dot{q}}\dot{q}} = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}$, respectively. Note that the basis operators for complex vectors and matrices hold for quaternion vectors and matrices, *e.g.*, the conjugate, transpose, and conjugate transpose. To formulate objective functions in quaternion domain, the norms of quaternion vectors and matrices are used as measurements. The l_1-norm of $\dot{\mathbf{q}} = (\dot{q}_s) \in \mathbb{H}^m$ is defined as $\|\dot{\mathbf{q}}\|_1 = \sum_{s=1}^{m} |\dot{q}_s|$, where $s = 1, \cdots, m$ is a position index, and the F-norm of $\dot{\mathbf{Q}} = (\dot{q}_{s,t}) \in \mathbb{H}^{m \times n}$ is defined by $\|\dot{\mathbf{Q}}\|_F = (\sum_{s=1}^{m} \sum_{t=1}^{n} |\dot{q}_{s,t}|^2)^{\frac{1}{2}} = [Tr(\dot{\mathbf{Q}}^* \dot{\mathbf{Q}})]^{\frac{1}{2}}$, where $s = 1, \cdots, m$ and $t = 1, \cdots, n$ are the row and column indices respectively.

One of the effective approaches to process quaternion matrices is to convert them into pairs of complex matrices [27]. Let $\dot{\mathbf{Q}} = \mathbf{Q}_a + \mathbf{Q}_b j \in \mathbb{H}^{m \times n}$ be the Cayley-Dickson construction of $\dot{\mathbf{Q}}$, where $\dot{\mathbf{Q}} = \mathbf{Q}_0 + \mathbf{Q}_1 i + \mathbf{Q}_2 j + \mathbf{Q}_3 k$, $\mathbf{Q}_a = \mathbf{Q}_0 + \mathbf{Q}_1 i$, and $\mathbf{Q}_b = \mathbf{Q}_2 + \mathbf{Q}_3 i$. The **complex adjoint form** [27] uniquely determines $\dot{\mathbf{Q}}$ using ($\mathbf{Q}_a$, $\mathbf{Q}_b$) as

$$\chi_{\dot{\mathbf{Q}}} = \begin{bmatrix} \mathbf{Q}_a & \mathbf{Q}_b \\ -\overline{\mathbf{Q}}_b & \overline{\mathbf{Q}}_a \end{bmatrix}, \tag{4}$$

where $\chi_{\dot{\mathbf{Q}}} \in \mathbb{C}^{2m \times 2n}$, and $\dot{\mathbf{Q}}$ and $\chi_{\dot{\mathbf{Q}}}$ are isomorphic [27]. This transformation has been widely used for quaternion matrix analysis, *e.g.*, QED [27].

B. LDA and Its Variants

1) LDA and 2D-LDA: LDA [1] and 2D-LDA [7] seek optimal projection bases, denoted by the columns of $\mathbf{V}$, to project input samples into low-dimensional subspace. In this subspace, the ratio of between-class scatter and within-class scatter is maximized. Let P_b and P_w represent the between-class and within-class scatters, and $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between-class and within-class covariance matrices of the input samples. Projecting samples into the low-dimensional subspace, the scatters of the projected samples can be evaluated by the traces of the corresponding matrices, *i.e.*, $P_b = Tr(\mathbf{V}^T \mathbf{S}_b \mathbf{V})$ and $P_w = Tr(\mathbf{V}^T \mathbf{S}_w \mathbf{V})$. The goals of LDA and 2D-LDA are to maximize the ratio $\frac{P_b}{P_w}$. There is no closed-form solution of the optimal $\mathbf{V}$. Instead, the trace ratio problem is simplified to a more tractable ratio trace problem [23], which can be efficiently solved via generalized eigen-decomposition.

2) QDA: QDA [18] incorporates the quaternion representation into discriminant analysis to preserve the high order cross-channel correlation of color images. Suppose there are images from c classes and the ith class has h_i samples, $\dot{\mathbf{x}}_j^i$ represents the jth vectorized quaternion sample from the ith class, and the mean quaternion sample of the ith class is denoted by $\bar{\dot{\mathbf{x}}}^i = \frac{1}{h_i} \sum_{j=1}^{h_i} \dot{\mathbf{x}}_j^i$, then $\dot{\mathbf{S}}_b = \sum_{i=1}^{c} h_i (\bar{\dot{\mathbf{x}}}^i - \bar{\dot{\mathbf{x}}})(\bar{\dot{\mathbf{x}}}^i - \bar{\dot{\mathbf{x}}})^*$ represents the between-class variance of the input samples. Let the columns of $\dot{\mathbf{V}}$ be the quaternion projection basis of QDA. QDA seeks an optimal basis that maximizes the between-class scatter (P_b) in the projected subspace

$$\max_{\dot{\mathbf{V}}} \; P_b = \max_{\dot{\mathbf{V}}} \; (Tr(\dot{\mathbf{V}}^* \dot{\mathbf{S}}_b \dot{\mathbf{V}})). \tag{5}$$

The solution of Eq. (5) equals to the leading eigenvectors of $\dot{\mathbf{S}}_b$. Note that QDA optimizes only the trace function instead of the trace ratio function. Besides, according to the properties of the quaternion functions ([28], TABLE 1), QDA holds either left or right linearity. Technically, it should be named quaternion left/right linear discriminant analysis. In both [18] and our work, we bypass the "left/right linear" for simplicity.

III. 2D-QSDA

QDA uses quaternion vectors to represent color images and shows performance enhancement over the configurations of independently processing different image channels or concatenating these image channels [18]. However, it fails to preserve the spatial structure of color images, endures high computation cost of processing high-dimensional vectors, and ignores the within-class scatter of the projected samples. Besides, QDA is easily influenced by noise and are not robust to classify the out-of-sample data which are unseen during the training phase (*e.g.*, occluded testing images). We propose 2D-QSDA to solve the above problems.

A. Model of 2D-QSDA

To preserve the spatial structure of images, we directly cope with 2D quaternion matrices. Specifically, let $\dot{\mathbf{X}}_j^i$ $(i = 1, \cdots, c)$ be the jth quaternion image matrix with class label i and h_i represent the number of samples in the ith class. We use $\bar{\dot{\mathbf{X}}}^i = \frac{1}{h_i} \sum_{j=1}^{h_i} \dot{\mathbf{X}}_j^i$ and $\bar{\dot{\mathbf{X}}} = \frac{1}{c} \sum_{i=1}^{c} \bar{\dot{\mathbf{X}}}^i$ to represent the mean sample

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

of the ith class and that of all images respectively. Let $\dot{\mathbf{S}}_b = \sum_{i=1}^{c} h_i(\dot{\bar{\mathbf{X}}}^i - \dot{\bar{\mathbf{X}}})(\dot{\bar{\mathbf{X}}}^i - \dot{\bar{\mathbf{X}}})^*$ and $\dot{\mathbf{S}}_w = \sum_{i=1}^{c}\sum_{j=1}^{h_i}(\dot{\mathbf{X}}_j^i - \dot{\bar{\mathbf{X}}}^i)(\dot{\mathbf{X}}_j^i - \dot{\bar{\mathbf{X}}}^i)^*$ represent the between-class and within-class covariance matrices, and $\dot{\mathbf{V}}_s = [\dot{\mathbf{v}}_{s1}, \cdots, \dot{\mathbf{v}}_{sk}]$ be the basis of 2D-QSDA. The projection scatters are calculated as $P_b = Tr(\dot{\mathbf{V}}_s^* \dot{\mathbf{S}}_b \dot{\mathbf{V}}_s)$ and $P_w = Tr(\dot{\mathbf{V}}_s^* \dot{\mathbf{S}}_w \dot{\mathbf{V}}_s)$. Incorporating the within-class scatter and the sparsity constraints into the objective of 2D-QSDA, we formulate it as a constrained trace ratio problem

$$\begin{aligned} \max_{\dot{\mathbf{V}}_s} \quad & \frac{Tr(\dot{\mathbf{V}}_s^* \dot{\mathbf{S}}_b \dot{\mathbf{V}}_s)}{Tr(\dot{\mathbf{V}}_s^* \dot{\mathbf{S}}_w \dot{\mathbf{V}}_s)} \\ \text{subject to} \quad & card(\dot{\mathbf{v}}_{sj}) \le \omega, \quad \text{for} \quad j = 1, \cdots, k, \end{aligned} \tag{6}$$

where $card(\cdot)$ denotes the cardinality (*i.e.*, the number of non-zero elements) of the basis, which can be measured via the l_0-norm. For different values of the tuning parameter ω, Eq. (6) yields adjustable levels of sparsity on the basis of 2D-QSDA.

There is no closed-form solution for the constrained trace ratio problem in Eq. (6). We therefore transform it into a constrained ratio difference problem. Considering the fact that the sparsity constraints are used to control the cardinality of the basis and not to alter the objective function, we propose a novel quaternion sparse regression (QSR) model that is equivalent to the constrained ratio difference form to find a numerical solution of 2D-QSDA. The QSR model of 2D-QSDA is presented as follows.

Theorem 1. *Let $\dot{\mathbf{\Omega}} = \dot{\mathbf{S}}_b - \mu \dot{\mathbf{S}}_w$ and its quaternion eigen-decomposition be $\dot{\mathbf{\Omega}} = \dot{\mathbf{R}} \mathbf{\Lambda} \dot{\mathbf{R}}^*$, and $\dot{\mathbf{V}}_s = [\dot{\mathbf{v}}_{s1}, \cdots, \dot{\mathbf{v}}_{sk}]$ be the solution of Eq. (6). Let $\dot{\mathbf{\Sigma}} = \dot{\mathbf{R}} \sqrt{|\mathbf{\Lambda}|} \dot{\mathbf{R}}^*$. For any $\lambda_2 \ge 0$ and $\lambda_{1,j} \ge 0$, $j = 1, \cdots, k$, if $\dot{\mathbf{A}} = [\dot{\mathbf{a}}_1, \cdots, \dot{\mathbf{a}}_k] \in \mathbb{H}^{m \times k}$ and $\dot{\mathbf{B}} = [\dot{\mathbf{b}}_1, \cdots, \dot{\mathbf{b}}_k] \in \mathbb{H}^{m \times k}$ satisfy*

$$\begin{aligned} \min_{\dot{\mathbf{A}}, \dot{\mathbf{B}}} \quad & (\|\dot{\mathbf{R}}^{-*} \dot{\mathbf{\Sigma}} - \dot{\mathbf{A}} \dot{\mathbf{B}}^* \dot{\mathbf{\Sigma}}\|_F^2 + \lambda_2 \|\dot{\mathbf{B}}\|_F^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\dot{\mathbf{b}}_j\|_1) \\ \textit{subject to} \quad & \dot{\mathbf{A}}^* \dot{\mathbf{A}} = \mathbf{I}_k, \end{aligned} \tag{7}$$

where $\dot{\mathbf{R}}^{-*} = (\dot{\mathbf{R}}^{-1})^*$, then $\dot{\mathbf{v}}_{sj} = \frac{\dot{\mathbf{b}}_j}{\|\dot{\mathbf{b}}_j\|_2}$ for appropriate $\lambda_{1,j}$, $j = 1, \cdots, k$.

By constructing $\dot{\mathbf{\Omega}} = \dot{\mathbf{S}}_b - \mu \dot{\mathbf{S}}_w$, we transform the trace ratio problem into a trace difference form. The sparsity constraints in Eq. (6) and the sparse regularization terms in Eq. (7) are used to control the cardinality of $\dot{\mathbf{v}}_{sj}$ and $\dot{\mathbf{b}}_j$ respectively at the expense of slightly decreasing the objective functions. We prove that $\dot{\mathbf{b}}_j$ is proportional to $\dot{\mathbf{v}}_{sj}$ without sparse regularization (see APPENDIX). Essentially, Theorem 1 makes a compromise between the class separability and the sparsity of the projection basis.

It is noteworthy that: 1) without sparsity constraints, the trace ratio problem can be solved by iteratively updating the value of μ since it is monotonously increasing [23]. However, due to regularization, the monotonicity is destroyed. Instead, we tune the value of μ to approximate the optimal value; 2) in Eq. (7), the sparsity of the basis of 2D-QSDA is controlled via the values of parameter $\lambda_{1,j}$ and the l_1-norm measurement since it is the tightest convex relaxation of the l_0-norm [29]. The detailed settings of model parameters will be introduced in Section V-B.

Remark: It is infeasible to optimize Eq. (6) using existing methods, and the strategy in our previous work [26] is not applicable as it is designed for a single objective based on the fact that maximizing the scatter of all projected samples equals to minimizing the sum of reconstruction errors [30]. We therefore propose Theorem 1 to optimize Eq. (6) that takes the form of maximizing a constrained trace ratio problem, in which two partially coupled objectives (*i.e.*, between-class and within-class scatters) are simultaneously optimized. In this respect, our QSR model Eq. (7) is completely different from that in [26] as the former is formulated by encodes two partially coupled measures, while the latter directly copes with image samples.

B. Solution of 2D-QSDA

Due to the complicated structure of quaternions, it is difficult to directly solve the problem in Eq. (7). Existing works convert the quaternion-valued problems to either the real space [16] or the complex space [26], [27] for efficient optimization. In this work, the complex space is adopted since the quaternion matrices and their complex adjoint forms are isomorphic [27]. In the following, we convert Eq. (7) into a complex form, extract the complex-valued solution, and then recover the quaternion-valued solution from the complex-valued one.

The two F-norm terms in Eq. (7) can be transformed into the complex space by adopting the complex adjoint forms of the quaternion matrices as (see APPENDIX)

$$\begin{aligned} & 2(\|\dot{\mathbf{R}}^{-*} \dot{\mathbf{\Sigma}} - \dot{\mathbf{A}} \dot{\mathbf{B}}^* \dot{\mathbf{\Sigma}}\|_F^2 + \lambda_2 \|\dot{\mathbf{B}}\|_F^2) \\ =& \|\chi_{\dot{\mathbf{R}}^{-*}} \chi_{\dot{\mathbf{\Sigma}}} - \chi_{\dot{\mathbf{A}}} \chi_{\dot{\mathbf{B}}^*} \chi_{\dot{\mathbf{\Sigma}}}\|_F^2 + \lambda_2 \|\chi_{\dot{\mathbf{B}}}\|_F^2. \end{aligned} \tag{8}$$

Let $\mathbf{R} = \chi_{\dot{\mathbf{R}}}$, $\mathbf{\Sigma} = \chi_{\dot{\mathbf{\Sigma}}}$, $\mathbf{A} = \chi_{\dot{\mathbf{A}}}$, $\mathbf{B} = \chi_{\dot{\mathbf{B}}}$, and $\mathbf{\Omega} = \mathbf{\Sigma}\mathbf{\Sigma}^*$, where the columns of $\mathbf{A}$ and $\mathbf{B}$ are $[\mathbf{a}_1, \cdots, \mathbf{a}_{2k}]$ and $[\mathbf{b}_1, \cdots, \mathbf{b}_{2k}]$. Eq. (8) can be rewritten as

$$Tr(\mathbf{R}^{-*} \mathbf{\Omega} \mathbf{R}^{-1}) - 2Re[Tr(\mathbf{A}^* \mathbf{R}^{-*} \mathbf{\Omega} \mathbf{B})] + Tr[\mathbf{B}^* (\mathbf{\Omega} + \lambda_2 \mathbf{I}) \mathbf{B}]. \tag{9}$$

The k l_1-norm terms in Eq. (7) can be reformulated using the operator $\xi(\cdot)$ given in Definition 1.

Definition 1. Let $\dot{\mathbf{q}} = \mathbf{q}_a + \mathbf{q}_b j \in \mathbb{H}^m$ and $\mathbf{q}$ be the first column of $\chi_{\dot{\mathbf{q}}}$, *i.e.*, $\mathbf{q} = \chi_{\dot{\mathbf{q}}}(:, 1) = [\mathbf{q}_a; -\overline{\mathbf{q}_b}] \in \mathbb{C}^{2m}$. $\xi(\mathbf{q})$ is defined as

$$\xi(\mathbf{q}) = [\mathbf{q}_a^T; \mathbf{q}_b^T] \in \mathbb{C}^{2 \times m}.$$

The l_1-norm of $\dot{\mathbf{q}}$ equals to the $l_{2,1}$-norm of the matrix $\xi(\mathbf{q})$:

$$\|\dot{\mathbf{q}}\|_1 = \|\xi(\mathbf{q})\|_{2,1},$$

where $\|\mathbf{M}\|_{2,1} = \sum_{j=1}^{m} \|\mathbf{M}(:, j)\|_2$.

According to Eq. (4), the complex adjoint form has a redundant structure. Hence, to recover a matrix in the complex adjoint form, we need to calculate only the first half columns and then infer the other half columns from the previous ones. Then 2D-QSDA can be reformulated into a complex form

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}} \quad & \{Tr(\mathbf{R}^{-*} \mathbf{\Omega} \mathbf{R}^{-1}) - 2Re[Tr(\mathbf{A}^* \mathbf{R}^{-*} \mathbf{\Omega} \mathbf{B})] + \\ & Tr[\mathbf{B}^* (\mathbf{\Omega} + \lambda_2 \mathbf{I}) \mathbf{B}] + 2 \sum_{j=1}^{k} \lambda_{1,j} \|\xi(\mathbf{b}_j)\|_{2,1}\} \\ \text{subject to} \quad & \mathbf{A}^* \mathbf{A} = \mathbf{I}_{2k}. \end{aligned} \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

There is no closed-form solution for Eq. (10) since variables $\mathbf{A}$ and $\mathbf{B}$ are coupled and it is intractable to simultaneously update $\mathbf{A}$ and $\mathbf{B}$. We develop an alternating minimization algorithm to iteratively learn their optimums. The iterative scheme is described as follows.

1. Update $\mathbf{A}$ for fixed $\mathbf{B}$. Given $\mathbf{B}$, the minimization of Eq. (10) is equivalent to

$$\begin{aligned} \max_{\mathbf{A}} \quad & Re[Tr(\mathbf{A}^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{B})] \\ \text{subject to} \quad & \mathbf{A}^*\mathbf{A} = \mathbf{I}_{2k}, \end{aligned} \tag{11}$$

which reduces to the orthogonal Procrustes problem in the complex domain [31]. Let $\mathbf{C} = \mathbf{R}^{-*}\mathbf{\Omega}\mathbf{B}$ and its singular value decomposition be $\mathbf{U}_c\mathbf{D}_c\mathbf{V}_c$. Then $\hat{\mathbf{A}} = \mathbf{U}_c\mathbf{V}_c$.

2. Update $\mathbf{B}$ for fixed $\mathbf{A}$. Given $\mathbf{A}$, Eq. (10) equals to

$$\min_{\mathbf{B}}\{\sum_{j=1}^{k}[\mathbf{b}_j^*(\mathbf{\Omega}+\lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j) + \lambda_{1,j}\|\xi(\mathbf{b}_j)\|_{2,1}]\}. \tag{12}$$

Thus, $\mathbf{B}$ can be optimized via k independent group Lasso problems. Specifically, $\mathbf{b}_j$ is solved by optimizing

$$\min_{\mathbf{b}_j} \; [\mathbf{b}_j^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j) + \lambda_{1,j}\|\xi(\mathbf{b}_j)\|_{2,1}]. \tag{13}$$

Since Eq. (13) does not have a closed-form solution, we rewrite it into the following constrained optimization problem by applying variable-splitting [32] to $\mathbf{b}_j$ and introducing an auxiliary variable $\mathbf{Z}$

$$\begin{aligned} \min_{\mathbf{b}_j} \quad & [\mathbf{b}_j^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j) + \lambda_{1,j}\|\mathbf{Z}\|_{2,1}] \\ \text{subject to} \quad & \mathbf{Z} = \xi(\mathbf{b}_j). \end{aligned} \tag{14}$$

We devise a novel algorithm under the framework of complex ADMM [33] to solve Eq. (14). Let $\xi^{-1}(\cdot)$ be the inverse operator of $\xi(\cdot)$. According to Definition 1, $\xi^{-1}(\cdot)$ converts a matrix of size $2\times m$ into a vector of size $2m\times 1$. The augmented Lagrangian function of Eq. (14) is

$$\begin{aligned} L(\mathbf{b}_j, \mathbf{Z}, \mathbf{y}) =& \mathbf{b}_j^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j) + \lambda_{1,j}\|\mathbf{Z}\|_{2,1} \\ & + Re(\mathbf{y}^*[\mathbf{b}_j - \xi^{-1}(\mathbf{Z})]) + \frac{\rho}{2}\|\mathbf{b}_j - \xi^{-1}(\mathbf{Z})\|_2^2, \end{aligned} \tag{15}$$

where $\mathbf{y}$ is the Lagrangian multiplier and $\rho > 0$ is the penalty parameter. To solve $L(\mathbf{b}_j, \mathbf{Z}, \mathbf{y})$, we iteratively update $\mathbf{b}_j$, $\mathbf{Z}$, and $\mathbf{y}$ while the other two variables are fixed. More specifically, given the τth update, the $(\tau+1)$th iteration to optimize $L(\mathbf{b}_j, \mathbf{Z}, \mathbf{y})$ is presented as follows.

- Update $\mathbf{b}_j^{\tau+1}$ by minimizing L w.r.t $\mathbf{b}_j$, which reduces to

$$\begin{aligned} \min_{\mathbf{b}_j} \; & \{\mathbf{b}_j^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j) \\ & + Re(\mathbf{y}^*[\mathbf{b}_j - \xi^{-1}(\mathbf{Z})]) + \frac{\rho}{2}\|\mathbf{b}_j - \xi^{-1}(\mathbf{Z})\|_2^2\}. \end{aligned} \tag{16}$$

  The solution of Eq. (16) is determined by setting the derivation of L w.r.t $\mathbf{b}_j$ to zero. Thus, $\mathbf{b}_j$ can be written explicitly as

$$\mathbf{b}_j^{\tau+1} = [\mathbf{\Omega} + (\lambda_2 + \rho)\mathbf{I}]^{-1}[\mathbf{\Omega}\mathbf{R}^{-1}\mathbf{a}_j + \rho\xi^{-1}(\mathbf{Z}^\tau) - \mathbf{y}^\tau]. \tag{17}$$

- Update $\mathbf{Z}^{\tau+1}$ by minimizing L w.r.t $\mathbf{Z}$. The optimization of L equals to

$$\begin{aligned} & \min_{\mathbf{Z}} \; \{\frac{\rho}{2}\|\mathbf{b}_j^{\tau+1} - \xi^{-1}(\mathbf{Z})\|_2^2 \\ & \qquad + Re(\mathbf{y}^*[\mathbf{b}_j^{\tau+1} - \xi^{-1}(\mathbf{Z})]) + \lambda_{1,j}\|\mathbf{Z}\|_{2,1}\} \\ =& \min_{\mathbf{Z}} \; \{\frac{1}{2}\|\xi^{-1}(\mathbf{Z}) - (\mathbf{b}_j^{\tau+1} + \frac{\mathbf{y}^\tau}{\rho})\|_2^2 + \frac{\lambda_{1,j}}{\rho}\|\mathbf{Z}\|_{2,1}\} \\ =& \min_{\mathbf{Z}} \; \{\frac{1}{2}\|\mathbf{Z} - \xi(\mathbf{b}_j^{\tau+1} + \frac{\mathbf{y}^\tau}{\rho})\|_F^2 + \frac{\lambda_{1,j}}{\rho}\|\mathbf{Z}\|_{2,1}\}. \end{aligned} \tag{18}$$

  Eq. (18) can be solved using Lemma 1, which is derived according to the optimization of the group Lasso problem [32], [33].

  Lemma 1. If a problem considering $\mathbf{Z} \in \mathbb{C}$ is to find

$$\min_{\mathbf{Z}} \; \{\frac{1}{2}\|\mathbf{Z} - \mathbf{T}\|_F^2 + \sigma\|\mathbf{Z}\|_{2,1}\}.$$

  The optimal $\mathbf{Z}$ satifies

$$\hat{\mathbf{Z}}(:,i) = \begin{cases} \frac{\|\mathbf{T}(:,i)\|_2 - \sigma}{\|\mathbf{T}(:,i)\|_2}\mathbf{T}(:,i), & \|\mathbf{T}(:,i)\|_2 > \sigma \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

- Update $\mathbf{y}^{\tau+1}$ as

$$\mathbf{y}^{\tau+1} = \mathbf{y}^\tau + \rho[\mathbf{b}_j^{\tau+1} - \xi^{-1}(\mathbf{Z}^{\tau+1})]. \tag{19}$$

The above procedures to optimize $\mathbf{b}_j$ are summarized in Algorithm 1.

Algorithm 1: Complex ADMM for computing $\mathbf{b}_j$

Input : $\mathbf{\Omega}$, $\mathbf{a}_j$, λ_2, and $\lambda_{1,j}$.
Output: Optimal $\mathbf{b}_j$.

```
Convert Eq. (13) into a constrained problem Eq. (14), and
construct the augmented Lagrangian function as Eq. (15).
Initialize b_j^0 = 0, Z^0 = 0, y^0 = 0, and ρ.
repeat
    Update b_j^{τ+1} using Eq. (17).
    Update Z^{τ+1} using Eq. (18).
    Update y^{τ+1} using Eq. (19).
until convergence;
Output b_j^{τ+1}.
```

Once $\mathbf{b}_j$ $(j = 1, \cdots, k)$ is optimized, we can obtain the optimal $\mathbf{b}_{k+j}$ from $\mathbf{b}_j$ according to the structure of the complex adjoint form. This way, the current optimal $\mathbf{B}$ is obtained. The alternating minimization algorithm for Eq. (10) continues until the stopping criteria is satisfied. Afterward, we convert this complex-valued solution into a quaternion-valued one using $\gamma(\cdot)$ in Definition 2. Note that this operator is also the one that is used to recover the eigen/singular vectors of a quaternion matrix from those of its complex adjoint matrix [34].

Definition 2. Let $\mathbf{c} = [c_1, \cdots, c_m, c_{m+1}, \cdots, c_{2m}]^T$, $\mathbf{c} \in \mathbb{C}^{2m}$. Define an operator $\gamma(\cdot)$ as

$$\gamma(\mathbf{c}) = [c_1, c_2, \cdots, c_m]^T + [c_{m+1}, c_{m+2}, \cdots, c_{2m}]^T j,$$

where $\gamma(\mathbf{c}) \in \mathbb{H}^m$ is in the Cayley-Dickson form.

Subsequently, the solution of Eq. (7), $\dot{\mathbf{V}}_s = [\dot{\mathbf{v}}_{s1}, \cdots, \dot{\mathbf{v}}_{sk}]$, can be recovered from the optimal columns of $\mathbf{B}$ as $\dot{\mathbf{v}}_{sj} = \gamma(\frac{\mathbf{b}_j}{\|\mathbf{b}_j\|_2})$, $j = 1, \cdots, k$. Finally, Algorithm 2 summarizes the detail procedures of 2D-QSDA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

Algorithm 2: 2D-QSDA

Input : Training set $\{\dot{\mathbf{X}}_i\}_{i=1}^{h}$, the dimension k, and parameters λ_2, $\lambda_{1,j}$, $j = 1, \cdots, k$.
Output: Optimal basis $[\dot{\mathbf{v}}_{s1}, \cdots, \dot{\mathbf{v}}_{sk}]$.

Reformulate 2D-QSDA into its complex form, namely, to find $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_{2k}]$ and $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_{2k}]$.
Initialize $\mathbf{B}$.
repeat
  (1) Update $\mathbf{A}$ for fixed $\mathbf{B}$ by solving the orthogonal Procrustes problem in the complex space (Eq. (11)).
  (2) Update $\mathbf{B}$ for fixed $\mathbf{A}$.
  for $j = 1, \cdots, k$, **do**
    Compute $\mathbf{b}_j$ using Algorithm 1.
  end
  for $j = k+1, \cdots, 2k$, **do**
    Derive $\mathbf{b}_j$ from $\mathbf{b}_{j-k}$.
  end
until *convergence*;
for $j = 1, \cdots, k$, **do**
  Recover $\dot{\mathbf{v}}_{sj}$ as $\gamma(\frac{\mathbf{b}_j}{\|\mathbf{b}_j\|_2})$.
end
Output $[\dot{\mathbf{v}}_{s1}, \cdots, \dot{\mathbf{v}}_{sk}]$.

C. Discussion

In this section, we examine the following issues to provide a comprehensive understanding of 2D-QSDA.

1) Optimal Conditions and Stopping Criterion: Essentially, 2D-QSDA involves a nested iterative scheme. The outer iteration can start with any $\mathbf{B}^*\mathbf{B} = \mathbf{I}_{2k}$. To improve the convergence speed, we set $\mathbf{B}$ to the complex adjoint form of the leading eigenvectors of $\dot{\mathbf{\Omega}}$. Let $\|\mathbf{b}_j\|_2$ be the residual of the jth column of $\mathbf{B}$. The stopping criterion of the outer iteration is that all column residuals are small enough, *i.e.*, $\|\mathbf{b}_j\|_2 < \varepsilon^{outer}$, where ε^{outer} is the tolerance and is fixed to 10^{-3} in our experiments.

As to the inner iteration, it is designed under the complex ADMM framework to compute $\mathbf{b}_j$ with the group Lasso penalty. Let the primal residual $r_{pri}^{\tau+1} = \mathbf{b}_j^{\tau+1} - \xi^{-1}(\mathbf{Z}^{\tau+1})$ and the dual variable residual $r_{dual}^{\tau+1} = \rho(\mathbf{Z}^{\tau+1} - \mathbf{Z}^{\tau})$. The optimal condition for the ADMM problem is that the residuals approach zero as the iteration proceeds [33]. In practice, this is achieved by setting the tolerances of the residuals to be small numbers. We empirically set the stopping criterion to $\|r_{pri}^{\tau+1}\|_2 < \varepsilon^{pri}$ and $\|r_{dual}^{\tau+1}\|_F < \varepsilon^{dual}$, where $\varepsilon^{pri} = \varepsilon^{dual} = 10^{-3}$.

2) Fast Complex ADMM with A Continuation Scheme: In ADMM, the choice of the penalty parameter ρ requires some precise tuning since it is crucial to the convergence behavior of the algorithm and also has a significant impact on the stability of the performance [35].

In this work, we modify the original complex ADMM algorithm by employing a continuation scheme [35], in which ρ is adapted according to the primal and dual variable residuals. The continuation scheme is defined as

$$\rho^{\tau+1} = \begin{cases} \nu^{incr}\rho^{\tau}, & \text{if } \|r_{pri}^{\tau}\|_2 > \mu\|r_{dual}^{\tau}\|_F \\ \rho^{\tau}/\nu^{decr}, & \text{if } \|r_{dual}^{\tau}\|_2 > \mu\|r_{pri}^{\tau}\|_F \\ \rho^{\tau}, & \text{otherwise} \end{cases} \tag{20}$$

where μ, ν^{incr}, $\nu^{decr} > 1$ are pre-defined parameters. The idea behind this continuation scheme is to keep the relative magnitudes of the primal and dual residuals within a factor μ such that the residuals converge to zero simultaneously. We set $\mu = 10$ and $\nu^{incr} = \nu^{decr} = 2$ as recommended in the literature.

The strengths of this continuation scheme lie in two folds: 1) the convergence behavior of the complex ADMM algorithm is more robust compared with a precise tuned ρ, and 2) the computation cost of the complex ADMM algorithm is greatly reduced since less iterations are needed [35]. These two advantages will be experimentally demonstrated in the following Section III-C3. With the continuation scheme, Algorithm 1 is now extended to Algorithm 3.

Algorithm 3: Fast complex ADMM with a continuation scheme

Input : $\mathbf{\Omega}$, $\mathbf{a}_j$, λ_2, $\lambda_{1,j}$, ε^{pri}, and ε^{dual}.
Output: Optimal $\mathbf{b}_j$.

Convert Eq. (13) to a constrained problem Eq. (14), and construct the augmented Lagrangian function as Eq. (15).
Initialize $\mathbf{b}_j^0 = \mathbf{0}$, $\mathbf{Z}^0 = \mathbf{0}$, $\mathbf{y}^0 = \mathbf{0}$, and $\rho^0 = 10^{-3}$.
repeat
  Update $\mathbf{b}_j^{\tau+1}$ using Eq. (17).
  Update $\mathbf{Z}^{\tau+1}$ using Eq. (18).
  Compute $r_{pri}^{\tau+1}$ and $r_{dual}^{\tau+1}$.
  Update $\mathbf{y}^{\tau+1}$ using Eq. (19).
  Adjust $\rho^{\tau+1}$ using Eq. (20).
until $\|r_{pri}^{\tau+1}\|_2 \leq \varepsilon^{pri}$ *and* $\|r_{dual}^{\tau+1}\|_F \leq \varepsilon^{dual}$;
Output $\mathbf{b}_j^{\tau+1}$.

3) Convergence Analysis: According to Eq. (10), 2D-QSDA converges to an optimum as long as the k independent group Lasso problems Eq. (12) converge. As shown in Fig. 1 (a), the outer iteration converges within 10 iterations.

For the inner iteration, the theoretical convergence of the complex ADMM algorithm for separable convex optimization was established in [33]. However, Eq. (12) is non-convex and the theoretical proof on its convergence is still an ongoing work. Following [33], we give the empirical convergence analysis. In this example, two color images (32*32 pixels) are chosen as the representations of two classes. Only one discriminant projection vector (denote its corresponding complex form as $\mathbf{b}_1$) can be learned from the training images, and we set the non-zero elements in this basis vector to four. As shown in Figs. 1 (b) and (c), if ρ is set to 0.001, the primal residual decreases slowly and Algorithm 1 cannot converge; meanwhile, Algorithm 1 achieves acceptable residuals within hundreds of iterations when $\rho = 0.1$. This verifies the importance of choosing an appropriate ρ.

Adopting Algorithm 3, we plot the residuals of the primal and dual variables with different initial values of ρ in Figs. 1 (d) and (e). Compared to the results from Algorithm 1, two observations can be made: 1) the convergence behavior of Algorithm 3 is relatively independent from the initial value of ρ; 2) Algorithm 3 converges within dozens of iterations, and thus markedly improves the computation efficiency.

4) Computation Complexity: We examine the computation cost of 2D-QSDA as: 1) the construction of $\dot{\mathbf{S}}_b$, $\dot{\mathbf{S}}_w$, and $\dot{\mathbf{\Omega}}$ costs operations of order $\mathcal{O}(hnm^2)$, and to obtain $\dot{\mathbf{R}}$ and $\dot{\mathbf{\Sigma}}$, the decomposition needs $\mathcal{O}(m^3)$ operations; 2) reformulating the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

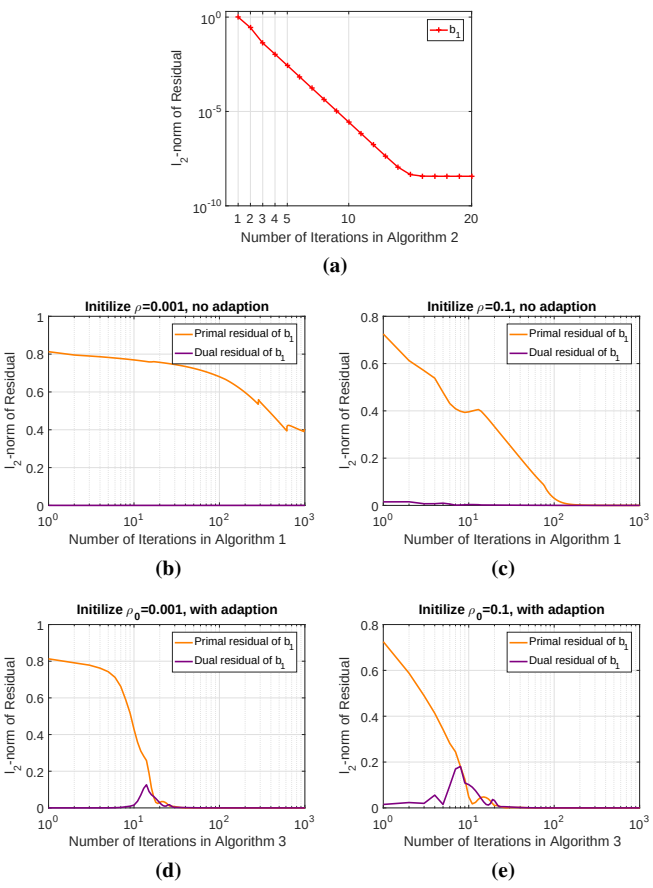

Fig. 1: Empirical convergence of 2D-QSDA: (a) outer iteration; (b) inner iteration with fixed $\rho = 0.001$; (c) inner iteration with fixed $\rho = 0.1$; (d) inner iteration with adaptive ρ and $\rho_0 = 0.001$; (e) inner iteration with adaptive ρ and $\rho_0 = 0.1$.

optimization problem into a complex form is in linear order; 3) in each iteration of Algorithm 3:

- for $\mathbf{A}$-update, the orthogonal Procrustes problem needs to be solved with $\mathcal{O}(m^3)$ operations.
- for $\mathbf{B}$-update, k basis vectors $(\mathbf{b}_j, j = 1, \cdots, k)$ are optimized individually under the complex ADMM framework, which is composed of four iterative steps as
  - i) updating $\mathbf{b}_j$. The most expensive cost is the matrix inverse operation with order $\mathcal{O}(m^3)$.
  - ii) updating $\mathbf{Z}$. The $\mathbf{Z}$-update is composed of calculating the column-wise l_2-norm and applying soft-thresholding at the cost of $\mathcal{O}(m)$.
  - iii) updating $\mathbf{y}$. The computation cost is $\mathcal{O}(m)$.
  - iv) adjusting ρ. The complexity is $\mathcal{O}(m)$.

Suppose the number of iterations of Algorithm 3 is T_1, it can be carried out at the cost of $\mathcal{O}(T_1 m^3)$. Then, the cost for $\mathbf{B}$-update will be $\mathcal{O}(kT_1 m^3)$. The cost of $\mathbf{A}$-update is negligible compared to that of $\mathbf{B}$-update. Let the number of iterations in Algorithm 2 be T_2. The total computation cost of 2D-QSDA is $\mathcal{O}(hnm^2 + kT_1T_2m^3)$.

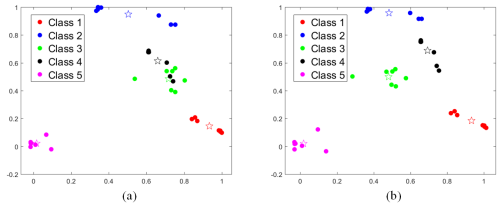

Fig. 2: Comparison of the separability of (a) 2D-QSDA and (b) 2D-QSDA$_w$. (Stars represent the class centers.)

IV. 2D-QSDA$_w$ USING WEIGHTED PAIRWISE BETWEEN-CLASS DISTANCES

The proposed 2D-QSDA is designed to maximize the between-class scatter while minimizing the within-class scatter with sparse constraints. Following the strategy in [24], we rewrite the final between-class scatter of 2D-QSDA as the mean scatter of all class pairs, *i.e.*, $\dot{\mathbf{S}}_b = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} h_i h_j (\bar{\mathbf{X}}^i - \bar{\mathbf{X}}^j)(\bar{\mathbf{X}}^i - \bar{\mathbf{X}}^j)^*$. That is, the between-class scatters of all class pairs are equally weighted, the final between-class scatter is thus dominated by large between-class scatters. However, the underlying goal of discriminant analysis is to maximize the between-class scatter of each class pair rather than separating each class center from the total mean. From this respect, maximizing small between-class scatters of class pairs is more challenging since the class pairs with large between-class scatters have already been well-separated.

We adopt a real-world dataset to illustrate this problem. Specifically, the first five classes from the PIE database [36] are selected with seven samples per class. We reshape the samples into quaternion vectors and project them into a two-dimensional subspace. As shown in Fig. 2 (a), Classes 1, 2, and 5 are well-separated, and the distance between Classes 3 and 4 is small. This is because the final between-class scatter is dominated by the large distances of class pairs, and hence, the pairwise between-class distance of Classes 3 and 4 is not maximized. Nevertheless, it is rather difficult to correctly separate Classes 3 and 4.

To solve this limitation, we propose 2D-QSDA$_w$ to improve the separability of 2D-QSDA using the weighted pairwise between-class distances. The core idea of 2D-QSDA$_w$ is to set large weights to the small between-class distances and vice verse. We define the weighting scheme of 2D-QSDA$_w$ as

$$\dot{\mathbf{S}}_b = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} w_{i,j} h_i h_j (\bar{\mathbf{X}}^i - \bar{\mathbf{X}}^j)(\bar{\mathbf{X}}^i - \bar{\mathbf{X}}^j)^* \qquad (21)$$

where $w_{i,j} = a^{d_{i,j}}$ is the weight of the (i,j) class pair, $0 < a < 1$ is a constant, and $d_{i,j}$ measures the between-class distance of the (i,j) class pair. In the experiments, we empirically set $a = 0.5$ and adopt the squared Euclidean distance to calculate $d_{i,j}$. As can be seen from Fig. 2 (b), adopting the weighting scheme, 2D-QSDA$_w$ finds a better subspace than 2D-QSDA since it is much easier to separate Classes 3 and 4 in this subspace.

After formulating the final between-class scatter of 2D-QSDA$_w$ using Eq. (21), 2D-QSDA$_w$ is optimized following the same strategy in 2D-QSDA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

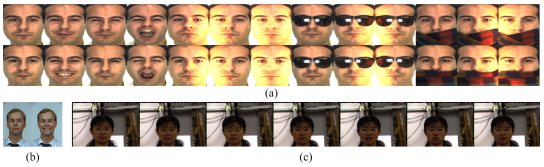

Fig. 3: Examples of color face images from: (a) AR, (b) Color FERET, and (c) CMU PIE.

V. Experiments

2D-QSDA and 2D-QSDA$_w$ are designed to extract sparse discriminant features from the RGB and RGB-D images while reducing the feature dimension. More importantly, they exhibit good generalization ability to the unseen data due to sparse constraints. Benefited from the intrinsic structure of quaternion, 2D-QSDA and 2D-QSDA$_w$ provide a unified approach to process RGB and RGB-D images and the extension from RGB images to RGB-D ones will not bring extra computation burden. In addition, since the depth channel contains complementary information to the color channels [9], 2D-QSDA and 2D-QSDA$_w$ can well extract the discriminant features by feeding the complementary data for performance enhancement. In this section, we validate the effectiveness of 2D-QSDA and 2D-QSDA$_w$ with applications of color and 3D face recognition. We introduce the databases and the experiment settings in Sections V-A and V-B, respectively. The performance of 2D-QSDA and 2D-QSDA$_w$ is compared with that of the state-of-the-arts in Sections V-C and V-D.

A. Databases

1) Color Face Databases: AR database contains 3276 color face images of 126 individuals with different expressions, illumination conditions, and occlusions. We employ a popular subset of AR [37] in which color face images from 100 individuals are cropped.

Color FERET database [38] contains 14,126 color face images of 1199 individuals. We collect a subset of FERET that contains 265 subjects with expression variations (images marked by "fa" and "fb").

CMU PIE database [36] is composed of color face images from 68 individuals. For each subject, face images with different poses, illumination, expressions, and frames from a talking sequence are recorded. We collect a subset (images captured by "C27") of CMU PIE. For each person, one neutral face image, two face images with blinking and smiling expressions, and four images from the talking sequence (frames labeled by 00, 19, 39, and 59) are selected.

2) 3D Face Databases: EURECOM Kinect database [39] provides 3D face images of 52 subjects. The face images are captured with different expressions, lighting conditions, and occlusions. In our experiments, 728 3D face images with frontal position are used.

IST-EURECOM light field face database (LFFD) [40] contains 3D face images of 100 individuals taken in two sessions with a temporal separation. Variations including emotions, actions, poses, illuminations, and occlusions are captured in each session for each subject. The face images are captured by a light field camera and then rendered into RGB and depth images. All frontal images are involved for our experiments, namely, 2800 3D face images are used in total.

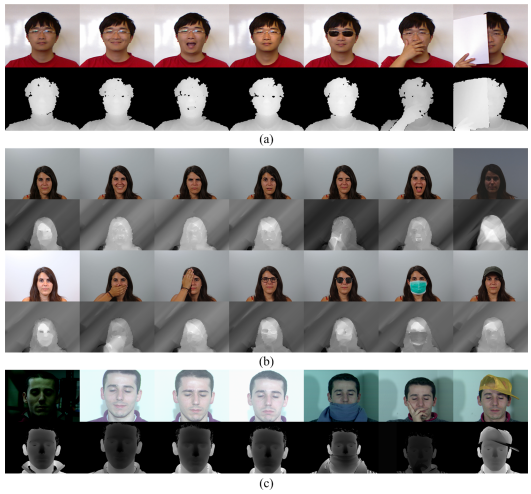

Fig. 4: Examples of 3D face images from: (a) EURECOM, (b) LFFD, and (c) UMB.

UMB database [41] is a set of 3D face images of 126 persons with a particular focus on real-world occlusions, *e.g.*, scarves, hats, hands. In our experiments, all persons that associated with occlusion images of scarfs, hands, and hats are selected, composing a subset of 882 3D images from 126 subjects.

In all experiments, the face images are aligned and cropped to 32*32 pixels based on the location of the eyes.

B. Experiment Settings

1) Parameters for 2D-QSDA and 2D-QSDA$_w$: The QSR models of 2D-QSDA and 2D-QSDA$_w$ are given in Eq. (7). To start with, μ controls the relative importance of the between-class and within-class scatters, and it is tuned among the values 10^{-3}, 10^{-2}, $\cdots$, 10^{3}.

λ_2 is used to avoid the potential colinearity problem when the number of training samples is far less than the input dimension of samples [29]. We empirically set it to 10^{-3}. Our algorithm is robust to the choice of λ_2 since we directly cope with 2D quaternion matrices, and hence the number of samples is generally larger than the number of processing dimensions.

$\lambda_{1,j}$ controls the sparsity of the jth basis vector, and this is implemented via column-wise soft-thresholding. According to Eq. (18) and Lemma 1, the threshold is set to $\sigma = \frac{\lambda_{1,j}}{\rho}$. In practice, to relieve the burden of manually tuning $\lambda_{1,j}$, we specify the cardinality ω. More specifically, we sort the values of $\|\mathbf{T}(:,i)\|_2$ and save them to $\mathbf{t}_{sort}$ in a descending order. The threshold σ is set to the $(\omega+1)$th element of $\mathbf{t}_{sort}$. This way, only ω columns of $\mathbf{T}$ are retained, and thus exactly ω non-zero entries in the basis vector are preserved when converting back into the quaternion space. For convenience, ω is fixed for all projection vectors and is chosen from $2, 4, \cdots, 32$.

2) Competing Algorithms: 15 state-of-the-art peer algorithms are used for comparison, including six unsupervised methods (PCA-based) and eight supervised ones (LDA-based).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

The competing algorithms are PCA [2], 2D-PCA [42], 2D-PCA-L1 [43], QPCA [12], MPCA [44], MSPCA [45], 2D-QSPCA [26], LDA [1], 2D-LDA [7], 2D-LDA-L1 [19], QDA [18], TDA [11], STDA [22], KDA [3], and IKDA [6].

Among them, PCA, LDA, KDA, and IKDA use vectorized samples, while 2D-PCA, 2D-PCA-L1, 2D-LDA, and 2D-LDA-L1 directly process 2D matrices. These algorithms are designed for gray-scale images, and we extend them to process RGB or RGB-D images by concatenating different image channels. MPCA and MSPCA utilize the third-order tensors to represent color or 3D face images. QPCA, QDA, 2D-QSDA, and 2D-QSDA$_w$ utilize the quaternion representation. Note that QPCA and QDA cope with vectorized samples while 2D-QSDA and 2D-QSDA$_p$ directly process quaternion matrices.

Following the literature [10], [16], the R, G, and B channels of color images are placed into the three imaginary parts of the quaternion components. When being applied to RGB-D images, we follow this convention to impose the color channels into the imaginary parts, and thus the depth channel is placed into the scalar part. In practice, we can arbitrarily place the R, G, B, and D channels into the four quaternion components since the advantage of quaternion representation is to holistically explore the correlation among multiple channels rather than the information from a particular channel [46].

3) Setups for Peer Algorithms: We first specify the projection dimension (k) of the competing algorithms. For PCA and QPCA, k is individually selected from $10, 20, 30, \cdots, ind$, where $ind = \min(h, m * n)$, h is the total number of training samples, and $m * n$ is the size of the image matrices ($m = n = 32$ in our experiments); for LDA and QDA, k is selected from $10, 20, 30, \cdots, c-1$, where c is the total number of classes; for MPCA, MSPCA, TDA, and STDA, their row (k_r) and column (k_c) dimensions are denoted by $k_r = k_c$, and they are selected from $1, 2, 3, \cdots, 32$, while the third dimension is chosen from $1, 2, 3$ for color face images and $1, 2, 3, 4$ for 3D face images; for 2D-PCA, 2D-PCA-L1, 2D-LDA, and 2D-LDA-L1, k is set to $2, 4, \cdots, 32$; for the kernel methods, *i.e.*, KDA and IKDA, we test all recommended parameters and record the best performance. Considering the sparse algorithms, MSPCA and STDA, the cardinality ω is set to $2, 4, \cdots, 32$ and is fixed for all basis vectors, as with 2D-QSDA and 2D-QSDA$_w$.

For all experiments in this work, the classification is based on the nearest neighbor classifier with l_1-norm distance. We report the best recognition rates and the corresponding dimension of features for all competing algorithms.

C. Color Face Recognition

Color information is an important cue for face recognition, and the high order cross-channel correlation should be considered to preserve the discriminant details of each specific class [8]. Treating different color channels in a holistic way, 2D-QSDA and 2D-QSDA$_w$ show significant improvements over the competing algorithms. Besides, 2D-QSDA$_w$ consistently outperforms 2D-QSDA, demonstrating the effectiveness of the weighting scheme.

1) Performance on Clean Face Images: 2D-QSDA and 2D-QSDA$_w$ are compared to the state-of-the-arts on three color face databases. The results are detailed in TABLE II and are summarized as:

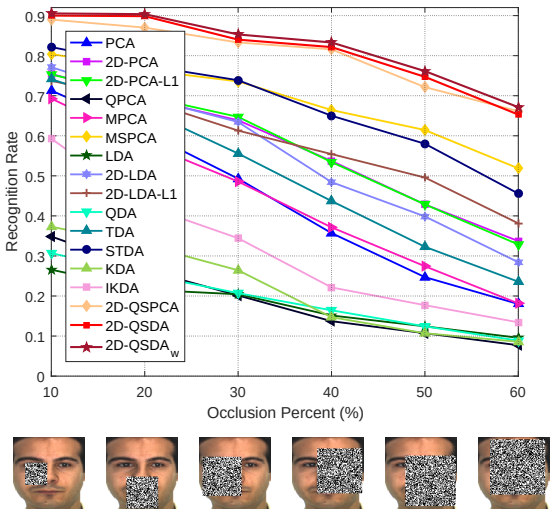

Fig. 5: Recognition rates on color face recognition with varying portions of occlusions.

- For the AR database, we test the performance of different algorithms with variations over time. The clean color face images in session one and session two are used for training and testing, respectively. 2D-QSDA$_w$ obtains the highest recognition rate, followed by 2D-QSDA.
- On color FERET, the recognition performance over varying expressions is examined. Note that the training set contains only one sample for each class, and hence the competing algorithms may suffer from the small sample size problem [29]. 2D-QSDA and 2D-QSDA$_w$ obtain consistently good performance because they essentially work in the column direction of color images and thus the number of samples is sufficient compared with the number of feature dimensions. In addition, in Eq. (7), setting $\lambda_2 > 0$ further improves the robustness of 2D-QSDA and 2D-QSDA$_w$ to the small sample size problem.
- On CMU PIE, similar to FERET, we impose a challenge on face recognition with single training image per person. In this test, the neutral color face images are used for training and the rest images with expression and facial action changes are used for testing. 2D-QSDA and 2D-QSDA$_w$ are superior to the other methods.

2) Performance on Partially Occluded Face Images: To evaluate the generalization ability of different algorithms in dealing with unseen data, we examine their robustness to both real-world and synthetic occlusions that are not involved in the training phase on the AR database.

- For real occlusions, the clean and natural-occluded face images are used to construct the training and testing sets respectively. According to TABLE III, 2D-QSDA$_w$ and 2D-QSPCA obtain the highest recognition rate and 2D-QSDA comes second in performance. This verifies the good generalization ability of the sparse feature extraction algorithms when the testing sets are contaminated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

TABLE II: Experiment results on clean color face images.

		PCA	2D-PCA	2D-PCA-L1	QPCA	MPCA	MSPCA	2D-QSPCA	LDA	2D-LDA	2D-LDA-L1	QDA	TDA	STDA	KDA	IKDA	2D-QSDA	2D-QSDA$_w$
AR	*acc*	0.7486	0.7957	0.7971	0.8142	0.7614	0.8357	0.8914	0.8057	0.8257	0.7571	0.4571	0.8271	0.8443	0.7971	0.8642	***0.8971***	**0.9043**
	dim	300	96*20	96*24	150*4	18*11*3	20*20*3	32*20*4	80	96*32	76*32	160*4	28*28*3	28*28*2	-	-	32*24*4	32*20*4
FERET	*acc*	0.7396	0.7434	0.7547	0.7094	0.7344	0.8438	0.8340	0.7358	0.7461	0.8038	0.7358	0.7472	0.8604	0.7547	0.7472	***0.8679***	**0.8830**
	dim	260	96*4	96*4	190*4	9*6*3	5*5*2	32*12*4	260	96*32	96*12	190*4	24*24*2	28*28*2	-	-	32*12*4	32*12*4
PIE	*acc*	0.5490	0.5735	0.5466	0.5123	0.5613	0.6520	0.7034	0.5392	0.5490	0.6348	0.5123	0.5784	0.6863	0.5098	0.5515	***0.7108***	**0.7230**
	dim	40	96*4	96*4	60*4	9*4*1	5*4*1	32*4*4	40	96*28	96*8	60*4	8*8*2	12*12*2	-	-	32*2*4	32*6*4

1. Bold number indicates the best performance, and italic bold number denotes the second best performance.
2. KDA and IKDA are kernel-based methods, and hence the best projection dimension cannot be reported.

TABLE III: Experiment results on natural-occluded color face images.

		PCA	2D-PCA	2D-PCA-L1	QPCA	MPCA	MSPCA	2D-QSPCA	LDA	2D-LDA	2D-LDA-L1	QDA	TDA	STDA	KDA	IKDA	2D-QSDA	2D-QSDA$_w$
AR	*acc*	0.3983	0.7575	0.7592	0.4575	0.5283	0.6667	**0.8492**	0.6183	0.6233	0.2733	0.4250	0.6800	0.7275	0.5291	0.5342	***0.8208***	**0.8492**
	dim	300	96*28	96*20	190*4	18*9*3	28*28*2	32*20*4	100	96*32	96*24	150*4	20*20*3	28*28*2	-	-	32*20*4	32*20*4

TABLE IV: Experiment results on clean 3D face images.

		PCA	2D-PCA	2D-PCA-L1	QPCA	MPCA	MSPCA	2D-QSPCA	LDA	2D-LDA	2D-LDA-L1	QDA	TDA	STDA	KDA	IKDA	2D-QSDA	2D-QSDA$_w$
EU	*acc*	0.7740	0.7212	0.6971	0.5817	0.7740	0.8173	0.7788	0.7885	0.7692	0.7404	0.5673	0.7644	0.7885	0.7692	0.8029	***0.8269***	**0.8365**
	dim	200	128*20	128*12	150*4	11*7*4	8*7*3	32*8*4	40	128*28	128*20	80*4	28*28*3	24*24*3	-	-	32*24*4	32*20*4
LFFD	*acc*	0.5488	0.4788	0.4825	0.2525	0.5100	0.5388	0.5637	0.5937	0.6012	0.4188	0.2325	0.6013	0.5025	0.5713	**0.6275**	0.5963	***0.6088***
	dim	300	128*32	128*28	180*4	9*6*4	8*7*2	32*8*4	120	128*32	128*24	110*4	32*32*4	24*24*3	-	-	32*20*4	32*16*4
UMB	*acc*	0.8122	0.8122	0.8175	0.7831	0.8042	***0.8413***	0.7857	0.8122	0.7989	0.7857	0.7831	0.8214	0.8386	0.7249	0.7143	0.8307	**0.8466**
	dim	120	128*4	128*4	20*4	9*6*3	8*7*2	32*12*4	100	128*24	128*24	20*4	12*12*3	8*8*3	-	-	32*4*4	32*6*4

TABLE V: Experiment results on natural-occluded 3D face images.

		PCA	2D-PCA	2D-PCA-L1	QPCA	MPCA	MSPCA	2D-QSPCA	LDA	2D-LDA	2D-LDA-L1	QDA	TDA	STDA	KDA	IKDA	2D-QSDA	2D-QSDA$_w$
EU	*acc*	0.6635	0.6250	0.6218	0.1731	0.6763	0.7885	0.7532	0.4230	0.4071	0.7789	0.2308	0.7372	0.8077	0.6378	0.6699	***0.8718***	**0.8814**
	dim	130	128*24	128*32	70*4	13*11*4	6*4*2	32*4*4	40	128*32	128*24	50*4	28*28*3	12*12*3	-	-	32*16*4	32*12*4
LFFD	*acc*	0.6825	0.7483	0.7475	0.0750	0.6700	0.7558	0.7842	0.59	0.6325	0.6858	0.0775	0.7333	0.7917	0.6475	0.7008	***0.8250***	**0.8333**
	dim	280	128*16	128*16	150*4	9*7*3	7*4*2	32*8*4	80	128*32	128*20	50*4	8*8*4	24*24*3	-	-	32*28*4	32*24*4
UMB	*acc*	0.4127	0.5582	0.5556	0.2989	0.4497	0.5979	0.5675	0.4232	0.4523	0.3915	0.2857	0.4894	0.5106	0.3439	0.4365	***0.6058***	**0.6164**
	dim	210	128*12	128*8	50*4	11*7*3	4*4*1	32*8*4	100	128*32	128*24	20*4	28*28*3	8*8*3	-	-	32*14*4	32*18*4

- For synthetic occlusions, the clean color face images from session one are used for training. We randomly add white-and-black blocks on the clean face images from session two to form the testing sets. The blocks are adjusted into different sizes ranging from 10% to 60% of the size of images. For each testing image, the random block is imposed on a random position and then the whole testing set is fixed to avoid the interference of randomness. As shown in Fig. 5, 2D-QSDA$_w$ and 2D-QSDA reach the highest and the second highest recognition rates under all experimental settings with 5%-10% improvements over the best competing algorithms.

D. 3D Face Recognition

3D (RGB-D) face images contain more robust features of a subject and thus offer more comprehensive representations. Incorporating the depth cue into traditional color face recognition has led to improvements with comparison to the usage of color face images alone [14]. 2D-QSDA and 2D-QSDA$_w$ exploit the quaternion representation, which intrinsically provide a way to encode the depth cue into the real dimension. Therefore, 3D face recognition can be fulfilled without extra computation cost. In this section, we evaluate the performance of different algorithms under the circumstances of 3D face recognition. In general, 2D-QSDA and 2D-QSDA$_w$ outperform or are comparable with the state-of-the-arts on clean 3D face images and are more reliable and generalizable to recognize 3D occluded face images. Besides, 2D-QSDA$_w$ obtains consistently improvements over 2D-QSDA.

1) Performance on Clean Face Images: The clean 3D face images in EURECOM, LFFD, and UMB databases are employed for experiments and the final results are presented in TABLE IV. The detailed comparison is analyzed as follows.

- For EURECOM and LFFD, clean 3D face images from two sessions are used for training and testing, respectively. 2D-QSDA$_w$ and 2D-QSDA are the top two methods on EURECOM; 2D-QSDA$_w$ is the second-best-performing method on LFFD, following IKDA.
- On the UMB database, a single training image per person is used to train optimal bases and the remaining face images with different expressions and intensive lightness changes compose the testing set. 2D-QSDA$_w$ is the best-performing method, followed by MSPCA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

TABLE VI: Ablation study on different modules of 2D-QSDA.

	Clean AR	Occluded AR	AR Occ =10%	AR Occ =20%	AR Occ =30%	AR Occ =40%	AR Occ =50%	AR Occ =60%
1D-SLDA	0.7514	0.5042	0.7114	0.6729	0.6600	0.5429	0.4886	0.3986
2D-SLDA	0.8414	0.6408	0.8029	0.7457	0.6986	0.6057	0.5271	0.4386
1D-QSDA	0.7886	0.5267	0.7600	0.7057	0.7071	0.6029	0.5357	0.4686
2D-QDA	0.8200	0.3933	0.7700	0.7300	0.6914	0.5929	0.5043	0.4029
2D-QSDA	**0.8971**	**0.8208**	**0.9000**	**0.8986**	**0.8400**	**0.8214**	**0.7471**	**0.6529**

2) Performance on Partially Occluded Face Images: We also compare the performance of competing algorithms in terms of their generalization ability by investigating their robustness to occlusions. The EURECOM, LFFD, and UMB databases are employed with all clean images for training and the natural-occluded face images for testing. As reported in TABLE V, 2D-QSDA$_w$ and 2D-QSDA consistently achieve the best and second-best performance and they outperform the best peer algorithms by the margins of 4%-10%. This validates the good generalization ability of 2D-QSDA and 2D-QSDA$_w$ in extracting features from RGB-D images.

To summarize: 1) 2D-QSDA and 2D-QSDA$_w$ are comparable with 2D-QSPCA on the RGB databases, and they show advantages over their competitors in most cases; 2) on RGB-D databases, 2D-QSDA$_w$ and 2D-QSDA obtain consistently the best and second-best performance, and they outperform 2D-QSPCA by the margins of 5%-17%. This observation coincides with the fact that 2D-QSDA and 2D-QSDA$_w$ can naturally take advantage of the discriminant information associated with the complement of depth and color channels; 3) 2D-QSDA$_w$ always outperforms 2D-QSDA since it pays more attention to the challenging class pairs using a weighting scheme.

E. Model Analysis

In this section, we present an in-depth study on the modules of 2D-QSDA to advance the understanding of the mechanisms behind 2D-QSDA. Similar observations can be made for 2D-QSDA$_w$ since it shares the same optimization strategy with 2D-QSDA.

1) Ablation Study: Firstly, we examine the ablation phenomena associated with the modules of 2D-QSDA, *i.e.*, quaternion representation, matrix-based processing, and sparse regularization. The results on the AR dataset are shown in TABLE VI. In general, the three modules of 2D-QSDA jointly work to improve the performance of 2D-QSDA. Specifically, 1) comparing the performance of 1D-SLDA and 2D-SLDA to that of 1D-QSDA and 2D-QSDA, we find that the quaternion representation helps to improve the performance in both clean and occluded images, and shows higher robustness to partial occlusion; 2) comparing the recognition rates of 1D-SLDA and 1D-QSDA with those of 2D-SLDA and 2D-QSDA, matrix-based processing shows more performance enhancements when dealing with clean images; 3) we can also find that the sparse regularization is highly beneficial to improve the robustness to partial occlusion according to the performance of 2D-QDA and 2D-QSDA.

2) Benefits from Sparse Regularization: We further explore the benefits gained from sparse regularization as it is an important module of 2D-QSDA. As shown in Fig. 6, the influence of the sparsity level on the classification accuracy is investigated with clean and partially-occluded images from the AR database. The red dots indicate the best recognition rates. As can be seen, 1) sparse regularization improves the recognition accuracy, and 2) a relatively higher sparsity level of the projection basis is required when being applied to partially-occluded images.

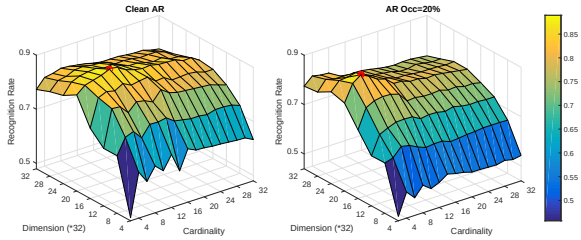

Fig. 6: Classification accuracy vs. cardinality and feature dimension.

The sparse regularization also provides a good interpretation for the basis vectors. Conceptually, 2D-QSDA works in the column direction of images, and thus the basis vectors maintain the discriminant information in the column space. With sparsity constraints, the obtained basis vectors emphasize the most important rows of images while ignoring the less important ones. For illustration, we visualize the non-zero entries in the first ten sparse basis vectors of 2D-QSDA trained on AR in Fig. 7. In this example, the cardinality of each basis vectors is fixed to eight. When projecting face images onto these sparse basis vectors, only non-black regions are retained and further considered in the subsequent processing. The non-black regions indicate that the discriminant features selected by 2D-QSDA are the informative parts of face images such as eyes, chin, nose, mouth, and cheek, which coincide with the discriminative parts reported in [47].

Fig. 7: Visualization of non-zero entries in the first ten sparse basis vectors of 2D-QSDA.

VI. Comparison with Existing Quaternion Models

To provide a comprehensive understanding of 2D-QSDA, in this section, we compare 2D-QSDA with several newly proposed quaternion-based image processing methods.

A. Comparison with 2D-QSPCA

Our previous work 2D-QSPCA [26] is relevant to this work because they follow the same basic optimization strategy. But

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

they are significantly different from the motivations, mathematical formulations, and have different applicable conditions: 1) 2D-QSPCA is designed to reduce the dimension of input samples while retaining the variation of the entire database as much as possible. Meanwhile, 2D-QSDA does not focus on preserving the common information of the whole database. Instead, it pays close attention to the divergence of different classes; 2) the objective of 2D-QSPCA is to maximize a constrained trace function. On the other hand, 2D-QSDA is constructed into a constrained trace ratio form, which is a more complicated and general model in the field of dimension reduction. Essentially, the objective of 2D-QSPCA falls into a special case of the trace ratio problem where the numerator equals the trace of an identity matrix; 3) since 2D-QSPCA extracts the common structure from the whole database, when being applied to process RGB-D images, it may lose the complementary information from the color and depth channels and thus suffer performance degradation. Meanwhile, 2D-QSDA focuses on the separability of the projected samples, and thus, can better utilize the complementary information.

To quantitatively show the properties of 2D-QSPCA and 2D-QSDA, we compare their separable ability in their corresponding projected spaces. More specifically, the class compactness and the separability of the projected samples are simultaneously considered via the Dunn index (DI) [48]:

$$DI = \frac{\min_{1\le i<j\le c}\delta(C_i, C_j)}{\max_{1\le k\le c}\Delta_k}, \tag{22}$$

where c is the class number, C_i and C_j represent the ith and jth classes, $\delta(\cdot)$ is the interclass distance metric, and Δ_k measures the compactness of the kth class. We extend the DI measure into the quaternion domain by calculating the corresponding quaternion distances. As verified in Fig. 8, 2D-QSDA has higher DI values on all databases, which is consistent with the motivation of discriminant analysis.

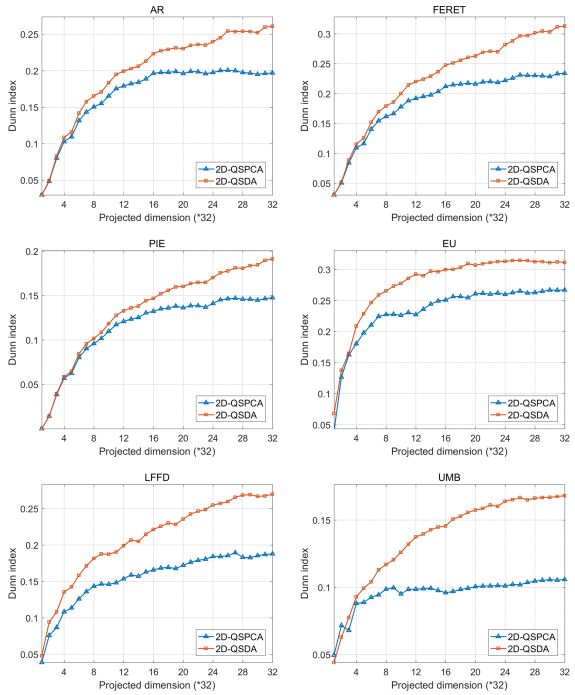

Fig. 8: Comparison of the Dunn index of 2D-QSPCA and 2D-QSDA.

B. Comparison with QMMC, QSRC, and QPCANet

Due to the powerful representation ability in capturing the high order cross-channel correlation, the quaternion algebra has been well integrated with other real domain techniques for color image processing [16], [49], [50].

A novel quaternion based maximum margin criterion (QMMC) algorithm was proposed in [49] to extract the color features and has shown advantages over the traditional QPCA and QDA criteria for the task of classification. While effective, QMMC transforms the color images into quaternion vectors, and thus, the spatial structures may be destroyed. In addition, QMMC cannot well process the unseen data since it exploits the l_2-norm as the measurement. Similar to QMMC, 2D-QSDA also focuses on discriminant features. The difference lies in the joint considering of the matrix-based operation, quaternion representation, and sparse regularization, through which the spatial and cross-channel structures of color images are well preserved and the robustness of 2D-QSDA is ensured.

Inspired by the fact that sparse representation-based classification (SRC) has achieved great success in face recognition, the work in [16] proposed quaternion-based SRC (QSRC). In contrast to 2D-QSDA that aims to extract the low-dimensional discriminant features from high-dimensional data, QSRC is specialized for classification and it will not reduce the dimension of samples. Imposing sparse coefficients, QSRC is robust to the naturally-occluded face images to a certain degree. However, its performance is greatly decreased for synthetic occlusions. This may come from the fact that the distribution of the synthetic occlusions is too far away from the clean images, and hence, it is very hard to obtain the correct representation coefficients.

Nowadays, neural network-based models have witnessed great performance improvement in many real-world applications. Taking advantage of the cascade architecture of the network in extracting high-dimensional multi-scale features, PCANet [51] devised a simple yet effective structure for face recognition. To facilitate the processing of color images, QPCANet [50] was proposed and has shown performance enhancement. By contrast, 2D-QSDA is a statistical method and is used to extract low-dimensional features.

The numerical comparison of the above-mentioned algorithms is reported in TABLE VII. Generally, QMMC and QSRC are not robust enough to the unseen data. Benefited from the cascade network structure, QPCANet usually obtains the best performance in recognizing clean face images and the face images with small occlusions; meanwhile, 2D-QSDA is more robust than QPCANet when the face images suffer from a large portion of occlusions. This is derived from the facts that: 1) the multi-layer high-dimensional features extracted from QPCANet contain more rich information for recognition, while they are prone to the over-fitting problem; 2) the sparse regularization improves the generalization ability of 2D-QSDA. Therefore, when being applied to images with partial occlusions, 2D-QSDA shows competitive advantages.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

TABLE VII: Comparison of 2D-QSDA and other quaternion-based methods.

	RGB databases										RGB-D databases					
	AR	FERET	PIE	Occluded AR	AR Occ =10%	AR Occ =20%	AR Occ =30%	AR Occ =40%	AR Occ =50%	AR Occ =60%	EU	LFFD	UMB	Occluded EU	Occluded LFFD	Occluded UMB
QMMC	0.6929	0.7434	0.5392	0.46	0.6771	0.5271	0.4143	0.2743	0.2071	0.1371	0.75	0.1725	***0.7989***	0.4615	0.2617	0.4127
QSRC	***0.94***	0.766	0.5	***0.8492***	0.5914	0.3286	0.1929	0.1	0.0543	0.0429	0.8173	**0.7488**	0.6905	0.8333	0.7308	0.4206
QPCANet	**0.96**	**0.9358**	**0.7451**	**0.9758**	**0.93**	***0.8743***	***0.8271***	***0.6386***	***0.4571***	***0.2371***	**0.9471**	0.5788	0.76	**0.9455**	**0.7407**	***0.7517***
2D-QSDA	0.8971	***0.8679***	***0.7108***	0.8208	***0.9029***	**0.8986**	**0.84**	**0.8214**	**0.7471**	**0.6529**	***0.8269***	***0.5963***	**0.8307**	***0.8718***	***0.5974***	**0.825**

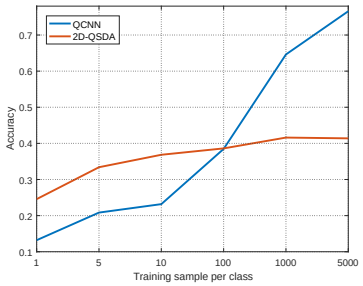

Fig. 9: Comparison of QCNN and 2D-QSDA on CIFAR10.

TABLE VIII: Comparison of QCNN and 2D-QSDA.

	AR	FERET	PIE
QCNN	0.8286	0.6264	0.5172
2D-QSDA	**0.8971**	**0.8679**	**0.7108**

C. Comparison with QCNN

Though effective in classification, QPCANet uses only simple operations to emulate the processing layers of convolutional neural networks (CNN), and thus it may lose the optimal representation ability of quaternion algebra in preserving the cross-channel relationship. Recently, a novel quaternion CNN [52] was proposed by re-designing the basic network modules in quaternion domain, and it shows promising performance in generic object recognition. We therefore compare the performance of QCNN and 2D-QSDA on CIFAR10 [53] and several color face databases. Please note that the data augmentation operation in QCNN is turned off for fair comparison.

For CIFAR10, we use the first 1, 5, 10, 100, 1000, and 5000 images per subject in the training set for model training, respectively. As shown in Fig. 9, QCNN achieves much better performance when there are sufficient training samples, while 2D-QSDA has advantages in dealing with the limited training sample problem. This is because QCNN can learn more discriminative features from massive training samples, and its cascade network architecture is able to extract high-dimensional multi-scale features that contain rich information for recognition. Meanwhile, 2D-QSDA is a statistical dimension reduction method that works in the column space of images. Thus, a limited number of training samples per subject is adequate to discover the statistics of the dataset. For color face recognition, generally speaking, 2D-QSDA achieves better performance than QCNN as recorded in TABLE VIII.

VII. Conclusion

In this paper, we developed 2D-QSDA and 2D-QSDA$_w$ to extract sparse discriminant features from RGB and RGB-D images while reducing their dimensions. The constrained trace ratio problems of 2D-QSDA and 2D-QSDA$_w$ were first transformed into constrained trace difference problems, and then converted to flexible QSR models for optimization. To solve the QSR models, we converted them into equivalent complex forms, where the quaternion-valued sparse regularization terms were transformed into the complex-valued group Lasso penalties. We then designed a nested iterative algorithm to optimize the complex-valued models. In each iteration, the group Lasso problems were solved using a novel sub-algorithm that was devised under complex ADMM. When being applied to practical applications, 2D-QSDA and 2D-QSDA$_w$ have shown the enhanced performance on clean samples and the robustness to the out-of-sample data that is unseen in the training phase. Extensive experiments on color and 3D face recognition verified the effectiveness and the generalization ability of 2D-QSDA and 2D-QSDA$_w$.

The proposed 2D-QSDA and 2D-QSDA$_w$ exploit a nested iterative optimization scheme, in which the complex ADMM algorithm is involved. Nevertheless, the optimization procedure is inefficient when the samples are of large-size since the computation of complex ADMM is proportional to the cube of the row numbers of images. Inspired by [54], our future work will develop fast quaternion optimization algorithms.

VIII. Appendix

Proof: Without sparsity constraints, Eq. (6) reduces to

$$\frac{Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_b\dot{\mathbf{V}}_s)}{Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_w\dot{\mathbf{V}}_s)} \tag{23}$$

or equivalently, a trace difference problem

$$\begin{aligned} &\max_{\dot{\mathbf{V}}_s} \quad Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_b\dot{\mathbf{V}}_s) - \mu Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_w\dot{\mathbf{V}}_s) \\ =&\max_{\dot{\mathbf{V}}_s} \quad Tr(\dot{\mathbf{V}}_s^*(\dot{\mathbf{S}}_w - \mu\dot{\mathbf{S}}_w)\dot{\mathbf{V}}_s), \end{aligned} \tag{24}$$

where $\mu = \max Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_b\dot{\mathbf{V}}_s)/Tr(\dot{\mathbf{V}}_s^*\dot{\mathbf{S}}_w\dot{\mathbf{V}}_s)$ is the optimal ratio of the between-class and within-class scatters, and the solution of Eq. (24) equals to the leading eigenvectors of $\dot{\mathbf{S}}_b - \mu\dot{\mathbf{S}}_w$. Also note that $\dot{\mathbf{\Omega}} = \dot{\mathbf{S}}_b - \mu\dot{\mathbf{S}}_w$ in Theorem 1.

To provide an efficient tool for quaternion analysis, the properties of the complex adjoint form of the quaternion matrix are reviewed in TABLE IX. Let the multiplication and addition of quaternion matrices $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ be compatible. According to [27], we have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2947775, IEEE Transactions on Image Processing

TABLE IX: Properties of the complex adjoint form.

1. $(\chi_{\dot{\mathbf{P}}})^* = \chi_{\dot{\mathbf{P}}^*}$
2. $(\chi_{\dot{\mathbf{P}}})^{-1} = \chi_{\dot{\mathbf{P}}^{-1}}$ if $\dot{\mathbf{P}}^{-1}$ exists
3. $\chi_{(\dot{\mathbf{P}}+\dot{\mathbf{Q}})} = \chi_{\dot{\mathbf{P}}} + \chi_{\dot{\mathbf{Q}}}$
4. $\chi_{\dot{\mathbf{P}}\dot{\mathbf{Q}}} = \chi_{\dot{\mathbf{P}}}\chi_{\dot{\mathbf{Q}}}$
5. $2\|\dot{\mathbf{Q}}\|_F^2 = 2Tr(\dot{\mathbf{Q}}^*\dot{\mathbf{Q}}) = \|\chi_{\dot{\mathbf{Q}}}\|_F^2 = Tr(\chi_{\dot{\mathbf{Q}}^*}\chi_{\dot{\mathbf{Q}}})$

Based on these properties, Eq. (7) can be transformed to

$$\begin{aligned}&2(\|\dot{\mathbf{R}}^{-*}\dot{\mathbf{\Sigma}} - \dot{\mathbf{A}}\dot{\mathbf{B}}^*\dot{\mathbf{\Sigma}}\|_F^2 + \lambda_2\|\dot{\mathbf{B}}\|_F^2)\\ &= \|\chi_{\dot{\mathbf{R}}^{-*}}\chi_{\dot{\mathbf{\Sigma}}} - \chi_{\dot{\mathbf{A}}}\chi_{\dot{\mathbf{B}}^*}\chi_{\dot{\mathbf{\Sigma}}}\|_F^2 + \lambda_2\|\chi_{\dot{\mathbf{B}}}\|_F^2. \end{aligned} \tag{25}$$

Let $\mathbf{R} = \chi_{\dot{\mathbf{R}}},\ \mathbf{\Sigma} = \chi_{\dot{\mathbf{\Sigma}}}, \mathbf{A} = \chi_{\dot{\mathbf{A}}},\ \mathbf{B} = \chi_{\dot{\mathbf{B}}}$, where the columns of $\mathbf{A}$ and $\mathbf{B}$ are $[\mathbf{a}_1, \cdots, \mathbf{a}_{2k}]$ and $[\mathbf{b}_1, \cdots, \mathbf{b}_{2k}]$. Eq. (25) can be rewritten as

$$Tr[(\mathbf{R}^{-1} - \mathbf{B}\mathbf{A}^*)(\mathbf{R}^{-*} - \mathbf{A}\mathbf{B}^*)\mathbf{\Sigma}\mathbf{\Sigma}^*] + \lambda_2 Tr(\mathbf{B}^*\mathbf{B}). \tag{26}$$

Let $\mathbf{\Omega} = \mathbf{\Sigma}\mathbf{\Sigma}^*$. Eq. (26) can be further reduced to

$$\begin{aligned}&Tr(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}) - 2Re[Tr(\mathbf{A}^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{B})] + Tr[\mathbf{B}^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{B}]\\ =&Tr(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}) + \sum_{j=1}^{2k}\mathbf{b}_j^*(\mathbf{\Omega} + \lambda_2\mathbf{I})\mathbf{b}_j - 2Re(\mathbf{a}_j^*\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{b}_j). \end{aligned} \tag{27}$$

Given $\mathbf{A}$, the optimal $\mathbf{b}_j$ can be solved individually as

$$\hat{\mathbf{b}}_j = (\mathbf{\Omega} + \lambda_2\mathbf{I})^{-1}\mathbf{\Omega}\mathbf{R}^{-1}\mathbf{a}_j, \tag{28}$$

or equivalently,

$$\hat{\mathbf{B}} = (\mathbf{\Omega} + \lambda_2\mathbf{I})^{-1}\mathbf{\Omega}\mathbf{R}^{-1}\mathbf{A}. \tag{29}$$

Substituting Eq. (29) to Eq. (27), we have

$$Tr\{\mathbf{A}^*[\mathbf{R}^{-*}\mathbf{\Omega}(\mathbf{\Omega} + \lambda_2\mathbf{I})^{-1}\mathbf{\Omega}\mathbf{R}^{-1}]\mathbf{A}\}. \tag{30}$$

With the orthonormal constraint, the optimal columns of $\mathbf{A}$ are the leading eigenvectors of

$$\begin{aligned}&\mathbf{R}^{-*}\mathbf{\Omega}(\mathbf{\Omega} + \lambda_2\mathbf{I})^{-1}\mathbf{\Omega}\mathbf{R}^{-1}\\ =&\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1} + \lambda_2\mathbf{I})^{-1}\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}, \end{aligned} \tag{31}$$

which equals to the leading eigenvector of $\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}$. Let the eigen-decomposition of the Hermitian matrix $\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1}$ be $\mathbf{E}\mathbf{D}\mathbf{E}^*$ and $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_{2k}]$ where the columns of $\mathbf{E}$ are sorted with their corresponding eigenvalues in a descending order.

Then, $\hat{\mathbf{A}} = \mathbf{E}$, or equivalently, $\hat{\mathbf{a}}_j = \mathbf{e}_j$ $(j = 1, \cdots, 2k)$. Substituting $\mathbf{a}_j$ back into Eq. (28) gives

$$\begin{aligned}\mathbf{b}_j =&(\mathbf{\Omega} + \lambda_2\mathbf{I})^{-1}\mathbf{\Omega}\mathbf{R}^{-1}\mathbf{a}_j\\ =&\mathbf{R}^{-1}(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1} + \lambda_2\mathbf{I})^{-1}(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1})\mathbf{a}_j\\ =&\mathbf{R}^{-1}\mathbf{E}(\mathbf{D} + \lambda_2\mathbf{I})^{-1}\mathbf{E}^*\mathbf{E}\mathbf{D}\mathbf{E}^*\mathbf{a}_j\\ =&\frac{d_{jj}}{d_{jj} + \lambda_2}\mathbf{R}^{-1}\mathbf{a}_j, \end{aligned} \tag{32}$$

where d_{jj} is the jth diagonal element of $\mathbf{D}$.

Since $\dot{\mathbf{R}}$ is a unitary matrix and $\mathbf{R} = \chi_{\dot{\mathbf{R}}}$, we have $\mathbf{R}^* = \mathbf{R}^{-1}$. From $\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1} = \mathbf{E}\mathbf{D}\mathbf{E}^*$ it follows that

$$\begin{aligned}\mathbf{\Omega} =&\mathbf{R}^*(\mathbf{R}^{-*}\mathbf{\Omega}\mathbf{R}^{-1})\mathbf{R}\\ =&\mathbf{R}^*(\mathbf{E}\mathbf{D}\mathbf{E}^*)\mathbf{R}\\ =&(\mathbf{R}^{-1}\mathbf{E})\mathbf{D}(\mathbf{R}^{-1}\mathbf{E})^*. \end{aligned} \tag{33}$$

According to Eqs. (32) and (33), we know: 1) the optimal columns of $\mathbf{B}$ are proportional to the columns of $\mathbf{R}^{-1}\mathbf{A}$; 2) the columns of $\mathbf{R}^{-1}\mathbf{A} = \mathbf{R}^{-1}\mathbf{E}$ are the leading eigenvectors of $\mathbf{\Omega}$. That is, the optimal columns of $\mathbf{B}$ are proportional to the leading eigenvectors of $\mathbf{\Omega}$.

Recall that $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_{2k}]$ and $\mathbf{\Omega}$ are the complex adjoint forms of $\dot{\mathbf{B}}$ and $\dot{\mathbf{\Omega}}$ respectively, and the eigen-decomposition of $\dot{\mathbf{\Omega}}$ can be fully recovered from that of $\mathbf{\Omega}$ using operator $\gamma(\cdot)$ in Definition 2 [34]. Thus, after the recover operation, $\dot{\mathbf{b}}_j$ is proportional to the jth eigenvector of $\dot{\mathbf{\Omega}}$, or equivalently, $\dot{\mathbf{b}}_j$ is proportional to the optimal $\dot{\mathbf{v}}_{sj}$. ■

REFERENCES

[1] P. N. Belhumeur, J. P. 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, 1997.

[2] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, no. 1-3, pp. 37–52, 1987.

[3] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Netw. for Signal Process. Workshop*. IEEE, 1999, pp. 41–48.

[4] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, 2009.

[5] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, 2011.

[6] Z. Fan, Y. Xu, M. Ni, X. Fang, and D. Zhang, "Individualized learning for improving kernel Fisher discriminant analysis," *Pattern Recog.*, vol. 58, pp. 100–109, 2016.

[7] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1569–1576.

[8] A. W. Yip and P. Sinha, "Contribution of color to face recognition," *Perception*, vol. 31, no. 8, pp. 995–1003, 2002.

[9] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, 2013.

[10] Y. Xu, L. Yu, H. Xu, H. Zhang, and T. Nguyen, "Vector sparse representation of color image using quaternion matrix analysis," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1315–1329, 2015.

[11] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, 2007.

[12] N. Le Bihan and S. J. Sangwine, "Quaternion principal component analysis of color images," in *IEEE Int. Conf. Image Process.*, vol. 1. IEEE, 2003, pp. I–809.

[13] R. Lan and Y. Zhou, "Quaternion-Michelson descriptor for color image classification," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5281–5292, 2016.

[14] B. Chen, J. Yang, B. Jeon, and X. Zhang, "Kernel quaternion principal component analysis and its application in RGB-D object recognition," *Neurocomputing*, vol. 266, pp. 293–303, 2017.

[15] Y. Chen, X. Xiao, and Y. Zhou, "Low-rank quaternion approximation for color image processing." *IEEE Trans. Image Process.*, 2019.

[16] C. Zou, K. I. Kou, and Y. Wang, "Quaternion collaborative and sparse representation with application to color face recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3287–3302, 2016.

[17] R. Lan, Y. Zhou, and Y. Y. Tang, "Quaternionic local ranking binary pattern: a local descriptor of color images," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 566–579, 2015.

[18] Y. Xu, "Quaternion-based discriminant analysis method for color face recognition," *PloS one*, vol. 7, no. 8, p. e43493, 2012.

[19] C.-N. Li, Y.-H. Shao, and N.-Y. Deng, "Robust L1-norm two-dimensional linear discriminant analysis," *Neural Netw.*, vol. 65, pp. 92–104, 2015.

[20] X. Chang, F. Nie, Y. Yang, C. Zhang, and H. Huang, "Convex sparse PCA for unsupervised feature learning," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 1, p. 3, 2016.

[21] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Intell.* ACM, 2004, p. 78.

[22] Z. Lai, Y. Xu, J. Yang, J. Tang, and D. Zhang, "Sparse tensor discriminant analysis," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3904–3915, 2013.

[23] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, 2009.

[24] S. Zheng, C. H. Ding, F. Nie, and H. Huang, "Harmonic mean linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, 2018.

[25] X. Xiao and Y. Zhou, "Quaternion sparse discriminant analysis for color face recognition," in *IEEE Int. Conf. Multimedia & Expo*. IEEE, 2018, pp. 1–6.

[26] ——, "Two-dimensional quaternion PCA and sparse PCA," *IEEE Trans. Neural Netw. Learn. Syst.*, 2018.

[27] F. Zhang, "Quaternions and matrices of quaternions," *Linear Alg. Appl.*, vol. 251, pp. 21–57, 1997.

[28] E. M. Hitzer, "Quaternion Fourier transform on quaternion fields and generalizations," *Adv. Appl. Clifford Algebr.*, vol. 17, no. 3, pp. 497–517, 2007.

[29] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[30] I. Jolliffe, "Principal component analysis," 2002.

[31] M. T. Hanna, N. P. A. Seif, and W. A. E. M. Ahmed, "Hermite-Gaussian-like eigenvectors of the discrete Fourier transform matrix based on the direct utilization of the orthogonal projection matrices on its eigenspaces," *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2815–2819, 2006.

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[33] L. Li, X. Wang, and G. Wang, "Alternating direction method of multipliers for separable convex optimization of real functions in complex variables," *Math. Probl. Eng.*, vol. 2015, 2015.

[34] N. Le Bihan and J. Mars, "Singular value decomposition of quaternion matrices: a new tool for vector-sensor signal processing," *Signal Prog.*, vol. 84, no. 7, pp. 1177–1199, 2004.

[35] C. Song, S. Yoon, and V. Pavlovic, "Fast ADMM algorithm for distributed optimization with adaptive penalty." in *Am. Assoc. Artif. Intell.*, 2016, pp. 753–759.

[36] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database of human faces," Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-02, January 2001.

[37] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.

[38] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, 2000.

[39] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 44, no. 11, pp. 1534–1548, Nov 2014.

[40] A. Sepas-Moghaddam, V. Chiesa, P. L. Correia, F. Pereira, and J.-L. Dugelay, "The IST-EURECOM light field face database," in *Int. Workshop Forensic Biom.* IEEE, 2017, pp. 1–6.

[41] A. Colombo, C. Cusano, and R. Schettini, "UMB-DB: A database of partially occluded 3D faces," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*. IEEE, 2011, pp. 2113–2119.

[42] J. Yang, D. Zhang, A. F. Frangi, and J.-y. 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, 2004.

[43] 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, 2010.

[44] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, 2008.

[45] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, "Multilinear sparse principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1942–1950, 2014.

[46] C. C. Took and D. P. Mandic, "The quaternion LMS algorithm for adaptive filtering of hypercomplex processes," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1316–1327, 2008.

[47] O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "Which parts of the face give out your identity?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2011, pp. 641–648.

[48] H. Wan, H. Wang, G. Guo, and X. Wei, "Separability-oriented subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 409–422, 2018.

[49] Z. Liu, Y. Qiu, Y. Peng, J. Pu, and X. Zhang, "Quaternion based maximum margin criterion method for color face recognition," *Neural Process. Lett.*, vol. 45, no. 3, pp. 913–923, 2017.

[50] R. Zeng, J. Wu, Z. Shao, Y. Chen, B. Chen, L. Senhadji, and H. Shu, "Color image classification via quaternion principal component analysis network," *Neurocomputing*, vol. 216, pp. 416–428, 2016.

[51] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, 2015.

[52] X. Zhu, Y. Xu, H. Xu, and C. Chen, "Quaternion convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[53] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[54] D. Song, D. A. Meyer, and M. R. Min, "Fast nonnegative matrix factorization with rank-one ADMM," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.

Xiaolin Xiao received the B.E. degree from Wuhan University, China, in 2013 and the Ph.D. degree from University of Macau, Macau, China, in 2019. Currently, she is a Postdoc Fellow with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include superpixel segmentation, saliency detection, and color image processing and understanding.

Yongyong Chen received his B.S. and M.S. degrees in the College of Mathematics and Systems Science, Shandong University of Science and Technology and visited the National Key Lab for Novel Software Technology in Nanjing University as an exchange student in 2017. He is now pursuing his Ph.D. degree in the Department of Computer and Information Science, University of Macau. His research interests include low-rank and sparse matrix/tensor decomposition models, with applications to image processing, data mining and computer vision.

Yue-Jiao Gong (M'15) received the B.S. and Ph.D. degree in Computer Science from Sun Yat-sen University, China, in 2010 and 2014, respectively. Currently, she is a Full Professor with the School of Computer Science and Engineering, South China University of Technology, China. Her research interests include evolutionary computation, swarm intelligence, and their applications to intelligent transportation and smart city scheduling. She has published over 80 papers, including more than 30 IEEE Transactions papers, in her research area.

Yicong Zhou (M07-SM'14) received his B.S. degree from Hunan University, Changsha, China, and his M.S. and Ph.D. degrees from Tufts University, Massachusetts, USA, all in electrical engineering.

He is an Associate Professor and Director of the Vision and Image Processing Laboratory in the Department of Computer and Information Science at University of Macau. His research interests include image processing and understanding, computer vision, machine learning, and multimedia security.

Dr. Zhou is a senior member of the International Society for Optical Engineering (SPIE). He was a recipient of the Third Price of Macau Natural Science Award in 2014. He is a Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He serves as an Associate Editor for *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Geoscience and Remote Sensing*, and four other journals.