Discriminant feature extraction by generalized difference subspace

Kazuhiro Fukui, *Member, IEEE,* Naoya Sogi, *Student Member, IEEE,* Takumi Kobayashi, *Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* and Atsuto Maki, *Member, IEEE*

Abstract—In this paper, we reveal the discriminant capacity of orthogonal data projection onto the generalized difference subspace (GDS), both theoretically and experimentally. In our previous work, we demonstrated that the GDS projection works as a quasi-orthogonalization of class subspaces, which is an effective feature extraction for subspace based classifiers. Here, we further show that GDS projection also works as a discriminant feature extraction through a similar mechanism to the Fisher discriminant analysis (FDA). A direct proof of the connection between GDS projection and FDA is difficult due to the significant difference in their formulations. To circumvent the complication, we first introduce geometrical Fisher discriminant analysis (gFDA) based on a simplified Fisher criterion. It is derived from a heuristic yet practically plausible assumption: the direction of the sample mean vector of a class is largely aligned to the first principal component vector of the class, given that the principal component analysis (PCA) is applied without data centering. gFDA works stably even under few samples, bypassing the small sample size (SSS) problem of FDA. We then prove that gFDA is equivalent to GDS projection with a small correction term. This equivalence ensures GDS projection to inherit the discriminant ability from FDA via gFDA. Furthermore, we discuss two useful extensions of these methods, 1) a nonlinear extension by kernel trick, 2) a combination with CNN features. The equivalence and the effectiveness of the extensions have been verified through extensive experiments on the extended Yale B+, CMU face database, ALOI, ETH80, MNIST, and CIFAR10, mainly focusing on image recognition under small samples.

Index Terms—Discriminant analysis, Fisher criterion, subspace representation, PCA without data centering

1 Introduction

In this paper, we reveal the discriminant ability of orthogonal projection of data onto the generalized difference subspace (GDS) [1], called GDS projection. GDS is a mathematical concept that represents the difference between multiple subspaces, and it is defined as a natural extension of the difference vector between two vectors.

GDS projection can be seen with two natures in feature extraction and they can be alternatively exploited. One is to enlarge the angles between class subspaces to make their relationship closer to the orthogonal status [1]. As a result, GDS projection works as quasi-orthogonalization, and is an effective feature extraction technique for subspace based classifiers such as the subspace method and the mutual subspace method [2], [3], [4], [5], [6]. The other nature, on which we focus in this paper, is to serve for discriminative feature extraction, through a mechanism similar to the Fisher discriminant analysis (FDA) [7], [8]; we approach its mechanism from both theoretical and empirical aspects, by exploring the close connection between GDS projection and FDA. However, a direct proof of their close connection would not be straightforward due to the significant difference in their formulations. To circumvent the complication, we introduce geometrical Fisher discriminant analysis (gFDA) that is a discriminant analysis based on a simplified Fisher criterion in terms of class representation. We indirectly prove the close connection via gFDA, where gFDA serves as an intermediate concept as it inherits the discriminant ability from FDA and the intrinsic mechanism from GDS projection, respectively.

The simplification starts by considering that the lengths of data are normalized to a unit, i.e. all data are on a sphere, which is a quite common setting in image recognition; the data normalization is widely used as standard preprocessing as it can effectively reduce the influence of changes in brightness; it also contributes to recent advances of image feature representation in convolutional neural networks [9], [10]. In the following, we focus on image recognition under the data normalization as our primary target task to make the discussion concrete. Nevertheless, we expect that our methods can also work effectively on other data types.

For the simplification, we first introduce a heuristic assumption that the directions of the sample mean vector and the first principal component vector of a class are nearly equivalent, given that the principal component analysis (PCA) without data centering (subtracting the mean) is applied to calculate the principal component vectors. This

- *K. Fukui is with the Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan. E-mail: kfukui@cs.tsukuba.ac.jp.*
- *N. Sogi was with the Department of Computer Science, University of Tsukuba. He is now with the NEC Visual Intelligence Research Laboratories, Kawasaki, Japan. E-mail: naoya-sogi@nec.com.*
- *T. Kobayashi is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. E-mail: takumi.kobayashi@aist.go.jp.*
- *J.-H. Xue is with the Department of Statistical Science, University College London, London, UK. E-mail: jinghao.xue@ucl.ac.uk.*
- *A. Maki is with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: atsuto@kth.se.*

Manuscript received 23 Aug. 2020; revised 27 Dec.2021; accepted 10 Apr. 2022. Date of publication XXXX 2022; date of current version 20 Apr. 2022.
(Corresponding author: Kazuhiro Fukui.)
Recommended for acceptance by XXXXXX.
Digital Object Identifier no. XXX/TPAMI.2022.XXXXX

heuristic relationship enables us to reasonably represent the original Fisher criterion using the principal component vectors and their weights (eigenvalues) of all the classes involved in the classification task. Based on this representation, we simplify the original Fisher criterion in terms of class representation by introducing the assumptions that all the class distributions on a unit sphere have equal prior probability and isotropic variance, and finally approximate the original Fisher criterion compactly with only several principal component vectors for each class. Since a set of the principal component vectors of each class spans a class subspace, our simplified criterion can be considered as a method based on the geometrical relationship between the class subspaces. This new type of Fisher discriminant analysis is hence named geometrical Fisher discriminant analysis (gFDA).

The discriminant criterion of gFDA leads to a generalized eigenvalue problem for the matrix product of between-class and within-class matrices like FDA. This formulation makes it difficult to examine the connection between gFDA and GDS projection. Thus, we transform the generalized eigenvalue problem to a simpler regular eigenvalue problem for the linear combination of between-class and within-class matrices. The formation of the linear combination leads us to an observation that gFDA is equivalent to GDS projection with a small correction term under a condition of no overlaps between class subspaces. As a consequence, we can verify the close connection between FDA and GDS projection via gFDA, as gFDA can be regarded as an approximation of FDA.

The subspace representation also enables gFDA to deal with the situation where only a few samples are available. In this case, the within-class matrix becomes singular so that FDA cannot in principle be computed. This problem is called the small size sample (SSS) problem of FDA [8]. To address the SSS problem, many types of extensions of FDA have been proposed [8], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. gFDA is very different from these conventional methods in that it bypasses the SSS problem of FDA by representing the discriminant criterion in a form of linear combination, which can be solved without depending on the number of samples. gFDA can work even with only one sample without any specific modification, unlike most of the above extensions.

We can find that in many cases it is difficult and costly to collect and label sufficient learning samples for each object class in image recognition tasks such as face and medical image recognitions. In several applications, a class subspace can be stably generated even from few data; for example, in 3D object recognition, it is well known that a set of the images of a 3D convex object with Lambertian reflectance under various illumination conditions can be represented by a subspace with low dimension (from 3 to 9), which is called illumination subspace [21], [22], [23]. This means that an illumination subspace of a 3D object like face can be stably and accurately estimated from only a small number (from 3 to 9) of the object images under different illuminations. This characteristic of subspace representation works effectively to handle the small sample image recognition.

The main goal of the paper is to prove the close relationship between GDS projection and FDA theoretically and experimentally via gFDA, and further verify the robustness of our methods against the small sample size experimentally, as mentioned above. It is highly valuable to rebuild the essential mechanism of FDA by the quite different formulation using a subspace representation, considering high significance of FDA in pattern recognition and machine learning. However, we are also interested in further exploring the framework of gFDA and GDS projection, and thus discuss two useful extensions: 1) a nonlinear extension using kernel trick, and 2) a combination with CNN features.

For the first extension, our previous work [1] on the orthogonalization of class subspaces demonstrated that the nonlinear extension of GDS, Kernel GDS (KGDS), can deal with the case that each class has a nonlinear complicated structure that cannot be naively represented by a linear subspace. Motivated by this, we demonstrate that the nonlinear extension with Gaussian kernel function is also valid in terms of the discriminant ability of gFDA and GDS projection, where all samples are mapped on a unit sphere due to the kernel function.

For the second one, to achieve higher performance in difficult tasks like general object recognition with complicated background and without segmentation, we consider the usage of powerful features called CNN features, which are extracted from a fully connected layer of convolutional neural networks (CNN), as an input of our methods. We show that the combination of our methods and CNN features can achieve competitive performance in comparison with various types of state-of-the-art methods, suggesting the possibility of incorporating the mechanism of GDS projection into the framework of deep neural networks.

Our main contributions are summarized as follows:

- We reveal that the projection of data onto the generalized difference subspace, GDS projection, works as a discriminant analysis through a mechanism similar to the Fisher discriminant analysis. To show the above nature,
  - We propose a new discriminant analysis, geometrical Fisher discriminant analysis (gFDA), which maximizes a simplified Fisher criterion under the common setting that all data are normalized to a unit in terms of length.
  - We prove that gFDA is equivalent to GDS projection with a small correction term.
  - We show the close connection between GDS projection and FDA indirectly by regarding gFDA as an intermediate concept between them.
- We discuss two useful extensions: 1) a nonlinear extension using kernel trick, and 2) a combination with CNN features, in which they are used as an input of our methods.
- We demonstrate that gFDA, GDS projection and their nonlinear extensions have equivalent or better performance than the original FDA and its extensions on various public databases: the extended Yale B+, CMU Multi-PIE face, and ALOI (illumination direction collection), mainly focusing on the small sample image recognition. Besides, we show the effectiveness of the combination with CNN features on ETH80, MNIST, and CIFAR10.

The rest of this paper is organized as follows. Section 2 and Section 3 provide preliminary concepts. In Section 2, we describe the concept and the definition of the generalized difference subspace (GDS). In Section 3, we overview the fundamentals of FDA with the Fisher criterion. In Section 4, we introduce a heuristic assumption on the relationship between the first principal component vector and the mean vector of a class. Then, we simplify the Fisher criterion by using the heuristic relationship and construct the geometrical Fisher discriminant analysis (gFDA) with the simplified criterion. In Section 5, we describe the geometrical mechanism of gFDA and prove that gFDA has dual forms of objective function. In Section 6, we show the close connection between FDA and GDS projection via gFDA. We describe the nonlinear extension of our methods in Section 7. In Section 8, we demonstrate the effectiveness of gFDA through evaluation experiments, focusing on the situation of a small sample size. Section 9 concludes the paper.

2 Generalized difference subspace

In this section, we describe the concept of generalized difference subspace (GDS). As a preliminary to its definition, we describe how to generate a class subspace from the data set for each class. We then define the difference subspace (DS) for two subspaces and extend DS to GDS.

2.1 Generation of class subspace

The principal component vectors of a class are obtained by applying the principal component analysis (PCA) without data centering to a set of data from the class.

Given a set of n_c L-dimensional data $\{\mathbf{x}_i^c\}_{i=1}^{n_c}$ of class c $(c = 1, \ldots, C)$, where an image with $w \times h$ pixels is regarded as an $L(= w \times h)$ dimensional vector $\mathbf{x}$, the principal component vectors $\{\boldsymbol{\phi}_i^c\}_{i=1}^{d_c}$ of class c are obtained by the following procedure:

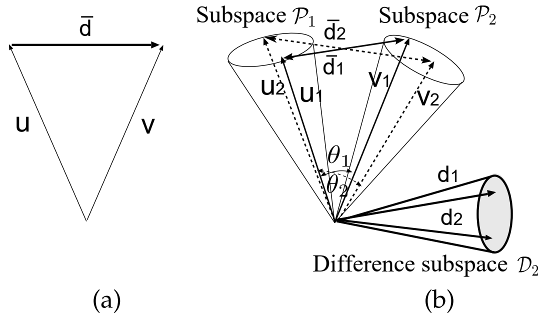
1. An $L \times L$ auto-correlation matrix is computed as $\mathbf{R}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} \mathbf{x}_i^c \mathbf{x}_i^{cT}$ from $\{\mathbf{x}_i^c\}_{i=1}^{n_c}$.
2. The principal component vectors $\{\boldsymbol{\phi}_i^c\}_{i=1}^{d_c}$ of class c are obtained as the unit eigenvectors corresponding to the d_c largest eigenvalues of $\mathbf{R}_c$. If we use all the eigenvalues, we obtain the spectral decomposition of the matrix $\mathbf{R}_c$.

Throughout the whole paper, the principal component vectors of a class are used as the orthonormal basis vector of the corresponding class subspace. In the following, we will interchangeably use the terms of principal component vector and orthonormal basis vector as of the same meaning.

2.2 Geometrical definition of DS

The *difference subspace* (DS) is a natural extension of a difference vector $\bar{\mathbf{d}}$ between two vectors $\mathbf{u}$ and $\mathbf{v}$ as shown in Figs.1a and 1b [1].

We formulate the *difference subspace* between M-dimensional subspace $\mathcal{P}_1$ and N-dimensional subspace $\mathcal{P}_2$ in L-dimensional vector space. In the case that there is no overlap between these subspaces, N canonical angles $\{\theta_i\}_{i=1}^{N}$ (for convenience $N \leq M$) can be obtained between them [24], [25]. Let $\bar{\mathbf{d}}_i$ be the difference vector, $\mathbf{v}_i - \mathbf{u}_i$, between canonical vectors $\mathbf{u}_i \in \mathcal{P}_1$ and $\mathbf{v}_i \in \mathcal{P}_2$, which are obtained through the framework of the canonical correlation analysis and form the ith canonical angle θ_i [24], [25]. All $\bar{\mathbf{d}}_i$ are orthogonal to each other. Thus, after normalizing the length of each difference vector $\bar{\mathbf{d}}_i$ to 1, we regard the normalized difference vectors $\mathbf{d}_i = \frac{\mathbf{v}_i - \mathbf{u}_i}{||\mathbf{v}_i - \mathbf{u}_i||}$ as the orthonormal basis vectors of the *difference subspace* $\mathcal{D}_2$. Thus, $\mathcal{D}_2$ is defined as $< \mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_N >$.



(a) (b)

Fig. 1. Basis concept of difference subspace: (a) difference vector, (b) canonical angles, vectors and difference subspace.

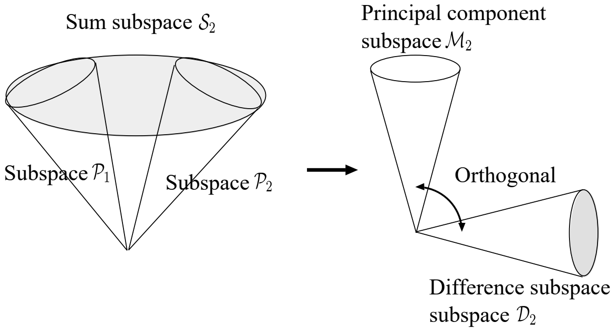


Fig. 2. Direct sum decomposition of sum subspace $\mathcal{S}_2$ into principal component subspace $\mathcal{M}_2$ and difference subspace $\mathcal{D}_2$.

2.3 Analytical definition of DS

The *difference subspace* geometrically defined in Sec.2.2 can also be analytically defined by using the orthogonal projection matrices of two class subspaces [1].

Theorem. *Assuming that there is no overlap between M-dimensional subspace $\mathcal{P}_1$ and N-dimensional subspace $\mathcal{P}_2$, and $\boldsymbol{\phi}_i^{1T}\boldsymbol{\phi}_j^2 \neq 0$ $(i = 1, \ldots, M, j = 1, \ldots, N)$, the i-th basis vector $\mathbf{d}_i$ of the difference subspace $\mathcal{D}_2$ is equal to the normalized eigenvector $\mathbf{x}_i$ of $\mathbf{P}_1 + \mathbf{P}_2$ that corresponds to the i-th smallest eigenvalue smaller than 1, where $\mathbf{P}_1$ and $\mathbf{P}_2 \in \mathbb{R}^{L \times L}$ are the orthogonal projection matrices, defined by $\sum_{i=1}^{M} \boldsymbol{\phi}_i^1 \boldsymbol{\phi}_i^{1T}$ and $\sum_{i=1}^{N} \boldsymbol{\phi}_i^2 \boldsymbol{\phi}_i^{2T}$, respectively.*

- N eigenvectors of matrix $\mathbf{P}_1 + \mathbf{P}_2$ corresponding to eigenvalues smaller than 1 span the *difference subspace* $\mathcal{D}_2$.
- N eigenvectors of matrix $\mathbf{P}_1 + \mathbf{P}_2$ corresponding to eigenvalues larger than 1 span the *principal component subspace* $\mathcal{M}_2$.

The relations lead to the conclusion that the sum subspace $\mathcal{S}_2$ of $\mathcal{P}_1$ and $\mathcal{P}_2$, spanned by all the eigenvectors of matrix $\mathbf{P}_1 + \mathbf{P}_2$, is represented by the orthogonal direct

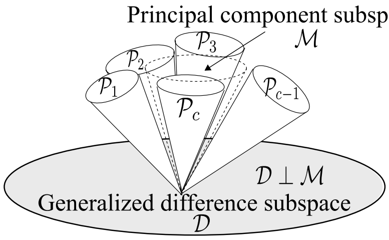


Fig. 3. Conceptual diagram of generalized difference subspace for c subspaces.

sum of the principal component subspace $\mathcal{M}_2$ and the *difference subspace* $\mathcal{D}_2$ as $\mathcal{S}_2 = \mathcal{M}_2 \bigoplus \mathcal{D}_2$. Fig.2 shows the conceptual diagram of this direct sum decomposition. This means that the *difference subspace* $\mathcal{D}_2$ can be defined as the subspace that is produced by removing the principal component subspace $\mathcal{M}_2$ from the sum subspace $\mathcal{S}_2$. Hence, the *difference subspace* can be regarded as the subspace that does not include the principal information of the two subspaces, that is, it contains only the difference component between them.

2.4 Definition of GDS

To deal with the difference between two or more subspaces, the concept of the *difference subspace* was generalized under the analytical definition [1]. Fig.3 shows the conceptual diagram of the *generalized difference subspace* (GDS) $\mathcal{D}$ for C subspaces.

Given $C(\geq 2)$ d_c-dimensional subspaces $\{\mathcal{P}_c\}_{c=1}^{C}$ in L-dimensional vector space, a *generalized difference subspace* $\mathcal{D}$ can be defined as such a subspace that is produced by removing the principal component subspace $\mathcal{M}$, of all the subspaces, from the sum subspace $\mathcal{S}$ of $\{\mathcal{P}_c\}_{c=1}^{C}$. Thus, the *generalized difference subspace* $\mathcal{D}$ is spanned by N_d eigenvectors, $\{\mathbf{d}_i\}_{i=1}^{N_d}$ corresponding to the N_d smallest eigenvalues, of the following sum matrix $\mathbf{G}$:

$$\mathbf{G} = \sum_{c=1}^{C} \mathbf{P}_c = \sum_{c=1}^{C} \sum_{i=1}^{d_c} \boldsymbol{\phi}_i^c \boldsymbol{\phi}_i^{cT} \ , \tag{1}$$

where $\mathbf{P}_c \in \mathbb{R}^{L \times L}$ denotes the orthogonal projection matrix of the class c subspace.

The generalized difference subspace $\mathcal{D}$ contains only the essential component for discriminating all the classes, since it is the orthogonal complement of the principal component subspace $\mathcal{M}$ that represents the principal information of all the class subspaces.

3 Fisher discriminant analysis

Fisher discriminant analysis (FDA) is a method for obtaining a discriminant space $\mathcal{H}$, which can distinguish multiple classes effectively [7], [8]. Such a discriminant space can be found out by maximizing the Fisher criterion of the projected data on the discriminant space $\mathcal{H}$.

The Fisher criterion consists of within-class covariance matrix and between-class covariance matrix. Given C classes, each of which contains the data set $\{\mathbf{x}_i^c\}_{i=1}^{n_c}$ ($c = 1, \ldots, C$), the within-class covariance matrix $\mathbf{\Sigma}_W \in \mathbb{R}^{L \times L}$ is defined as

$$\mathbf{\Sigma}_W = \sum_{c=1}^{C} p(c) \big(\frac{1}{n_c} \sum_{i=1}^{n_c} (\mathbf{x}_i^c - \mathbf{m}_c)(\mathbf{x}_i^c - \mathbf{m}_c)^T \big), \tag{2}$$

where $p(c)$ is the prior probability of class c, n_c and $\mathbf{m}_c$ indicate the number of samples and the mean vector of class c, respectively. The between-class covariance matrix $\mathbf{\Sigma}_B \in \mathbb{R}^{L \times L}$ is defined as

$$\begin{aligned} \mathbf{\Sigma}_B &= \sum_{c=1}^{C} p(c)(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T, & (3) \\ &= \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} p(i)p(j)(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T, & (4) \end{aligned}$$

where $p(i)$ and $p(j)$ are the prior probabilities of the i-th and j-th classes, and $\mathbf{m}$ indicates the mean vector over all the classes.

The Fisher criterion $f(\mathbf{d})$ of the data projected on a 1-dimensional subspace spanned by vector $\mathbf{d}$ is defined as

$$f(\mathbf{d}) = \frac{\mathbf{d}^T \mathbf{\Sigma}_B \mathbf{d}}{\mathbf{d}^T \mathbf{\Sigma}_W \mathbf{d}}, \tag{5}$$

where the vector $\mathbf{d}$ that maximizes function f can be obtained by solving the generalized eigenvalue problem

$$\mathbf{\Sigma}_B \mathbf{d} = \lambda \mathbf{\Sigma}_W \mathbf{d}. \tag{6}$$

Discriminant space $\mathcal{H}$ is spanned by $C-1$ eigenvectors, $\{\mathbf{d}_i\}_{i=1}^{C-1}$, corresponding to the $C-1$ largest eigenvalues of the above eigenvalue problem.

4 Geometrical Fisher discriminant analysis

In this section, we first approximate the Fisher criterion, $f(\mathbf{d})$ in Eq.(5), based on a heuristic relationship between the mean vector $\mathbf{m}_c$ and the first principal component vector $\boldsymbol{\phi}_1^c$ of class c on a unit sphere. We then simplify it by using the following assumptions across class distributions: 1) equal prior probability, 2) approximately equal covariance matrices before the data normalization, and 3) isotropic variance after the data normalization. Finally, we construct the proposed gFDA by maximizing the simplified Fisher criterion.

4.1 Equivalence between the class mean and first principal component vector

Heuristic relationship: For each class subspace, the first principal component vector $\boldsymbol{\phi}_1^c$ and the mean vector $\mathbf{m}_c$ can be in a very close correspondence with each other in terms of their directions, under the condition that $||\mathbf{m}_c||^2$ is comparatively larger than σ_{max}^2, where σ_{max}^2 is the maximum variance of the class distribution among all the dimensions.

The heuristic relationship can be explained as follows. Given a data set on a unit sphere, $\{\mathbf{x}_i^c\}_{i=1}^{n_c}$, of class c, the autocorrelation matrix $\mathbf{R}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$ are defined as

$$\mathbf{R}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{x}_i^c\mathbf{x}_i^{cT}, \tag{7}$$

$$\boldsymbol{\Sigma}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}(\mathbf{x}_i^c-\mathbf{m}_c)(\mathbf{x}_i^c-\mathbf{m}_c)^T. \tag{8}$$

Between the two matrices and the mean vector there is a relationship that $\mathbf{R}_c = \boldsymbol{\Sigma}_c + \mathbf{m}_c\mathbf{m}_c^T$.

It holds that $\|\mathbf{m}_c\|^2$ is more than four times larger than σ^2_{max} for a class distribution in almost all object classes on standard large scale datasets such as ALOI [26], CIFAR10 [27], CMU Multi-PIE Face Database [28] and MNIST [29]. This condition ensures that $\mathbf{m}_c$ is dominant in calculating the first principal component vector $\boldsymbol{\phi}_1^c$ corresponding to the largest eigenvalue (mean squared projection) of $\mathbf{R}_c$, as the projection of $\mathbf{m}_c$ on the direction of $\mathbf{m}_c$ is sufficiently larger than the projections of $\{\mathbf{x}_i^c-\mathbf{m}_c\}$ on any remaining directions measured with reference to the origin. Thus, the direction of the first principal component vector $\boldsymbol{\phi}_1^c$ almost coincides with that of the class mean vector $\mathbf{m}_c$.

We confirmed this via simulation, in which we randomly generated Gaussian distributions with various mean vector $\mathbf{m}_c$ and $\boldsymbol{\Sigma}_c$ on the condition that each simulated multivariate Gaussian data has only nonnegative elements to represent an image. Then, we projected them onto a unit sphere in vector spaces. According to the simulation under the above condition that $\frac{\|\mathbf{m}_c\|^2}{\sigma^2_{max}} = 4$, and changing the dimension L, from 2 to 500, of the vector space, the directions of $\boldsymbol{\phi}_1^c$ and $\mathbf{m}_c$ coincided with high correlation of over 0.999 for all the dimensions. Moreover, the simulation demonstrated that the heuristic assumption can still work with high correlation of over 0.999 even under the extreme condition that $\frac{\|\mathbf{m}_c\|^2}{\sigma^2_{max}} = 1.5$. In fact, the heuristic relationship work with high correlation of over 0.999 on all the datasets used in this paper, as experimentally confirmed in Section 8.

4.2 Simplification of the Fisher criterion

The within-class covariance matrix $\boldsymbol{\Sigma}_W \in \mathbb{R}^{L\times L}$ defined in Eq.(2) can be rewritten by the autocorrelation matrix $\mathbf{R}_c$ and the mean vectors $\mathbf{m}_c$ of the c-th class as follows:

$$\boldsymbol{\Sigma}_W = \sum_{c=1}^{C}p(c)(\mathbf{R}_c - \mathbf{m}_c\mathbf{m}_c{}^T), \tag{9}$$

By using the spectral decomposition of $\mathbf{R}_c$, $\boldsymbol{\Sigma}_W$ can be rewritten as

$$\boldsymbol{\Sigma}_W = \sum_{c=1}^{C}p(c)\Big((\sum_{i=1}^{d_c^{all}}\lambda_i^c\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}) - \mathbf{m}_c\mathbf{m}_c{}^T\Big), \tag{10}$$

where $d_c^{all} = \min(n_c, L)$, and λ_i^c and $\boldsymbol{\phi}_i^c$ indicate the i-th eigenvalue of the autocorrelation matrix $\mathbf{R}_c$ of the class c and its corresponding eigenvector, respectively.

Furthermore, by using the heuristic relationship, $\mathbf{m}_c \approx \bar{m}_c\boldsymbol{\phi}_1^c$, where $\|\boldsymbol{\phi}_1^c\| = 1$ and $\bar{m}_c = \|\mathbf{m}_c\|$, we replace $\boldsymbol{\Sigma}_W$ with $\boldsymbol{\Sigma}_{W1}$:

$$\begin{aligned}\boldsymbol{\Sigma}_{W1} &= \sum_{c=1}^{C}p(c)\Big((\sum_{i=1}^{d_c^{all}}\lambda_i^c\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}) - \bar{m}_c^2\boldsymbol{\phi}_1^c\boldsymbol{\phi}_1^{cT}\Big), &(11)\\ &= \sum_{c=1}^{C}p(c)\sum_{i=1}^{d_c^{all}}\sigma^2_{c,i}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}, &(12)\end{aligned}$$

where $\sigma^2_{c,i}$ represents the variance of the data projected on the i-th principal component vector $\boldsymbol{\phi}_i^c$, $\sigma^2_{c,1} = \lambda_1^c - \bar{m}_c^2$ and $\sigma^2_{c,i} = \lambda_i^c (i \geq 2)$.

With the heuristic relationship, the between-class covariance $\boldsymbol{\Sigma}_B$ can be represented with $\boldsymbol{\phi}_1^c$ as follows:

$$\boldsymbol{\Sigma}_{B1} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}p(i)p(j)(\bar{m}_i\boldsymbol{\phi}_1^i - \bar{m}_j\boldsymbol{\phi}_1^j)(\bar{m}_i\boldsymbol{\phi}_1^i - \bar{m}_j\boldsymbol{\phi}_1^j)^T. \tag{13}$$

We refer to an FDA based on the Fisher criterion of $\frac{\mathbf{d}^T\boldsymbol{\Sigma}_{B1}\mathbf{d}}{\mathbf{d}^T\boldsymbol{\Sigma}_{W1}\mathbf{d}}$ as approximated FDA (aFDA). We simplify the representation of $\boldsymbol{\Sigma}_{B1}$ and $\boldsymbol{\Sigma}_{W1}$ in the following two steps.

Simplification-I: We assume that the prior probabilities $p(c)$ of all the classes are equal to $\frac{1}{C}$. Besides, we use only d_c principal component vectors corresponding to the eigenvalues larger than a specified threshold:

$$\boldsymbol{\Sigma}_{W2} = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{d_c}\sigma^2_{c,i}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}. \tag{14}$$

We also assume that all the classes have approximately equal covariance matrices. Under the data normalization, this assumption leads to that the mean vectors $\{\mathbf{m}_i\}$ of all the classes have almost the same length, $\forall c, \bar{m}_c \approx \bar{m}$. The data normalization can project the data of a class onto a small local region of the unit sphere. As a result, the mean vectors of the normalized data sets can have similar lengths across different classes. Based on this, we replace the norms $\bar{m}_i$ of the mean vectors of all the classes with $\bar{m}$ as follows:

$$\boldsymbol{\Sigma}_{B2} = \frac{\bar{m}^2}{C^2}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)^T. \tag{15}$$

We refer to the FDA based on the simplified Fisher criterion of $\frac{\mathbf{d}^T\boldsymbol{\Sigma}_{B2}\mathbf{d}}{\mathbf{d}^T\boldsymbol{\Sigma}_{W2}\mathbf{d}}$ as simplified FDA (sFDA).

Simplification-II: Next, assuming that all the class distributions have the same isotropic variance on a unit sphere, we replace all the values of $\{\sigma^2_{c,i}\}$ with $\bar{\sigma}^2_{max}$, where $\bar{\sigma}^2_{max}$ is the maximum variance of the class distribution across all the dimensions. With this assumption, we further simplify $\boldsymbol{\Sigma}_{W2}$ to $\boldsymbol{\Sigma}_{W3}$ as

$$\boldsymbol{\Sigma}_{W3} = \frac{\bar{\sigma}^2_{max}}{C}\sum_{c=1}^{C}\sum_{i=1}^{d_c}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}. \tag{16}$$

We can regard this variance replacement as the estimation of the within-class variance on the safe side, because $\boldsymbol{\Sigma}_{W3}$ is an upper bound of $\boldsymbol{\Sigma}_{W2}$ as $\mathbf{x}^T\boldsymbol{\Sigma}_{W3}\mathbf{x} > \mathbf{x}^T\boldsymbol{\Sigma}_{W2}\mathbf{x}$, $\forall\mathbf{x}$, which works on the safe side for the variance minimization in FDA. Here, we should note that $\boldsymbol{\Sigma}_{W3}$ is compactly represented by a set of C d_c-dimensional subspaces. Such

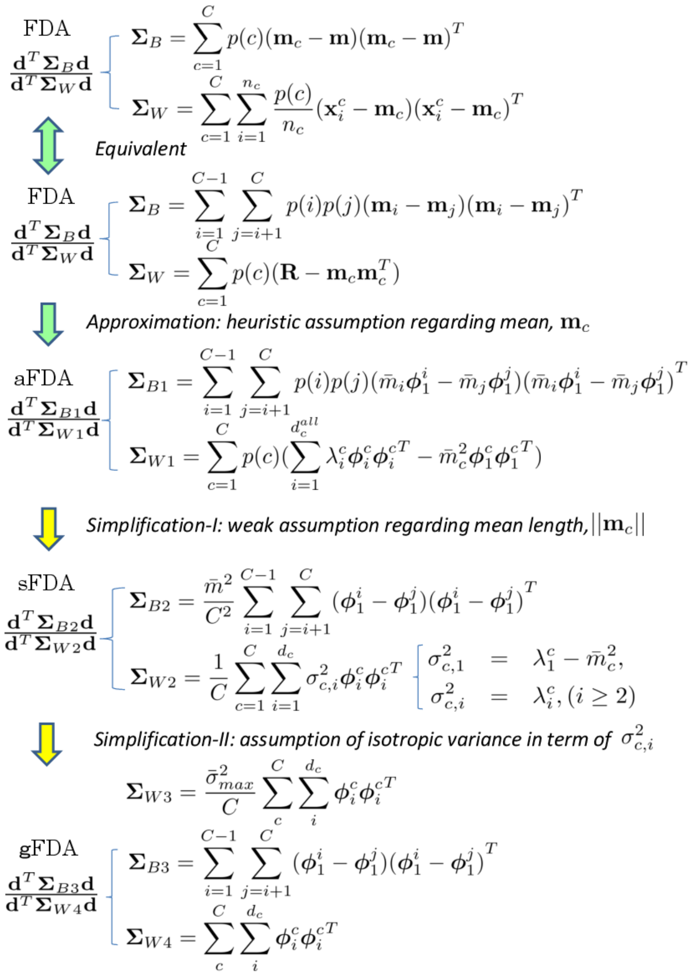

FDA
$\frac{\mathbf{d}^T\mathbf{\Sigma}_B\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_W\mathbf{d}}$
$\mathbf{\Sigma}_B = \sum_{c=1}^{C} p(c)(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$
$\mathbf{\Sigma}_W = \sum_{c=1}^{C}\sum_{i=1}^{n_c} \frac{p(c)}{n_c}(\mathbf{x}_i^c - \mathbf{m}_c)(\mathbf{x}_i^c - \mathbf{m}_c)^T$
Equivalent
FDA
$\frac{\mathbf{d}^T\mathbf{\Sigma}_B\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_W\mathbf{d}}$
$\mathbf{\Sigma}_B = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p(i)p(j)(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$
$\mathbf{\Sigma}_W = \sum_{c=1}^{C} p(c)(\mathbf{R} - \mathbf{m}_c\mathbf{m}_c^T)$
Approximation: heuristic assumption regarding mean, $\mathbf{m}_c$
aFDA
$\frac{\mathbf{d}^T\mathbf{\Sigma}_{B1}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W1}\mathbf{d}}$
$\mathbf{\Sigma}_{B1} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p(i)p(j)(\bar{m}_i\boldsymbol{\phi}_1^i - \bar{m}_j\boldsymbol{\phi}_1^j)(\bar{m}_i\boldsymbol{\phi}_1^i - \bar{m}_j\boldsymbol{\phi}_1^j)^T$
$\mathbf{\Sigma}_{W1} = \sum_{c=1}^{C} p(c)(\sum_{i=1}^{d_c^{all}} \lambda_i^c\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT} - \bar{m}_c^2\boldsymbol{\phi}_1^c\boldsymbol{\phi}_1^{cT})$
Simplification-I: weak assumption regarding mean length, $||\mathbf{m}_c||$
sFDA
$\frac{\mathbf{d}^T\mathbf{\Sigma}_{B2}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W2}\mathbf{d}}$
$\mathbf{\Sigma}_{B2} = \frac{\bar{m}^2}{C^2}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)^T$
$\mathbf{\Sigma}_{W2} = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{d_c}\sigma_{c,i}^2\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT} \quad \begin{cases} \sigma_{c,1}^2 = \lambda_1^c - \bar{m}_c^2, \\ \sigma_{c,i}^2 = \lambda_i^c, (i \geq 2) \end{cases}$
Simplification-II: assumption of isotropic variance in term of $\sigma_{c,i}^2$
$\mathbf{\Sigma}_{W3} = \frac{\bar{\sigma}_{max}^2}{C}\sum_{c}^{C}\sum_{i}^{d_c}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}$
gFDA
$\frac{\mathbf{d}^T\mathbf{\Sigma}_{B3}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W4}\mathbf{d}}$
$\mathbf{\Sigma}_{B3} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)^T$
$\mathbf{\Sigma}_{W4} = \sum_{c}^{C}\sum_{i}^{d_c}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}$

Fig. 4. The simplification process from FDA toward gFDA. the simplification process from Eq.(5) to Eq.(21) is summarized.

subspaces can be stably estimated from even a small number of samples. As a result, this assumption makes sense in practice and effectively contributes to bypassing the SSS problem of FDA and improving sFDA's robustness against the small sample size, as demonstrated in the paper.

Several types of Fisher-like criteria are available to define as combinations of the above simplified matrices. In this paper, we are interested in the simplest criterion defined by $\mathbf{\Sigma}_{B2}$ and $\mathbf{\Sigma}_{W3}$ and consider the objective function f_1:

$$f_1(\mathbf{d}) = \frac{\mathbf{d}^T\mathbf{\Sigma}_{B2}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W3}\mathbf{d}}, \quad (17)$$

$$= \frac{\bar{m}^2}{C\bar{\sigma}_{max}^2}\frac{\mathbf{d}^T\mathbf{\Sigma}_{B3}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W4}\mathbf{d}}, \quad (18)$$

where

$$\mathbf{\Sigma}_{B3} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)(\boldsymbol{\phi}_1^i - \boldsymbol{\phi}_1^j)^T, \quad (19)$$

$$\mathbf{\Sigma}_{W4} = \sum_{c=1}^{C}\sum_{i=1}^{d_c}\boldsymbol{\phi}_i^c\boldsymbol{\phi}_i^{cT}. \quad (20)$$

Since the term $\frac{\bar{m}^2}{C\bar{\sigma}_{max}^2}$ is constant, we ignore it and define our final objective function $f_g(\mathbf{d})$ as

$$f_g(\mathbf{d}) = \frac{\mathbf{d}^T\mathbf{\Sigma}_{B3}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W4}\mathbf{d}}. \quad (21)$$

We can finally obtain vector $\mathbf{d}$ by solving the following generalized eigenvalue problem:

$$\mathbf{\Sigma}_{B3}\mathbf{d} = \lambda\mathbf{\Sigma}_{W4}\mathbf{d}. \quad (22)$$

The process of the set of simplifications is summarized in Fig.4. We define the FDA based on the above simplified Fisher criterion as geometrical FDA (gFDA).

4.3 Criterion based on class subspaces

Our Fisher-like criterion $f_g(\mathbf{d})$ is defined by using only the principal component vectors $\{\boldsymbol{\phi}_i^c\}_{i=1}^{d_c}$. This can be interpreted as that $f_g(\mathbf{d})$ is determined based on the geometry of the class subspaces, which are spanned by the principal component vectors $\{\boldsymbol{\phi}_i^c\}_{i=1}^{d_c}$ of each class c.

More specifically, the denominator of $f_g(\mathbf{d})$ indicates the sum of the orthogonal projection matrices of all the class subspaces and the numerator indicates the autocorrelation matrix of all the difference vectors among the first orthogonal basis vectors, namely, their mean vectors. This indicates that the maximization of $f_g(\mathbf{d})$ can be realized, by minimizing the sum of projections of all the class subspaces while maximizing the projections of the differences between the mean vectors at the same time. Reflecting this mechanism, we name the discriminant analysis based on our Fisher-like criterion *geometrical FDA (gFDA)*.

5 Discrimination Mechanism of gFDA

5.1 Two-steps process

It is well known that the whole process of FDA consists of two steps: whitening and PCA. The process of gFDA in the form of $\frac{\mathbf{d}^T\mathbf{\Sigma}_{B3}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W4}\mathbf{d}}$ can be also divided into these two steps as shown in Fig.5.

We consider the case that C N-dimensional class subspaces in $\mathbb{R}^L$ $(L \gg N)$ are given, assuming that there is no overlap between class subspaces. For the simplicity of discussion, to make the matrix $\mathbf{\Sigma}_{W4}$ full rank, we assume that the dimensionality of the vector space can be reduced from L to $\hat{L} = CN$ by applying PCA-based dimensionality reduction. Thus, in the following, we consider C N-dimensional class subspaces in $\mathbb{R}^{\hat{L}}$ (Input space $\mathcal{I}$). The details of each step are as follows:

1) In the first step, whitening $\mathbf{A}$ such that $\mathbf{A}^T\mathbf{\Sigma}_{W4}\mathbf{A} = \mathbf{I}$ is applied to CN orthonormal basis vectors $\{\boldsymbol{\phi}_i^c\}$ of C N-dimensional class subspaces. As a result, the orthonormal basis vectors of all the classes are orthogonalized to each other. A subspace spanned by these orthogonalized basis vectors in the first step is called normalized space $\mathcal{N}$ in contrast with the original input space $\mathcal{I}$. Let the orthogonalized basis vectors be $\{\hat{\boldsymbol{\phi}}_i^c\}$ in the normalized space $\mathcal{N}$.
2) In the second step, PCA is applied to a set of difference vectors $\{\mathbf{z}_1^{ij}\}(i = 1, \ldots, C-1, j = i+1, \ldots, C)$ between the first principal component vectors, $\{\hat{\boldsymbol{\phi}}_1^c\}_{c=1}^C$, where $\mathbf{z}_1^{ij}$=$\hat{\boldsymbol{\phi}}_1^i - \hat{\boldsymbol{\phi}}_1^j$. We obtain

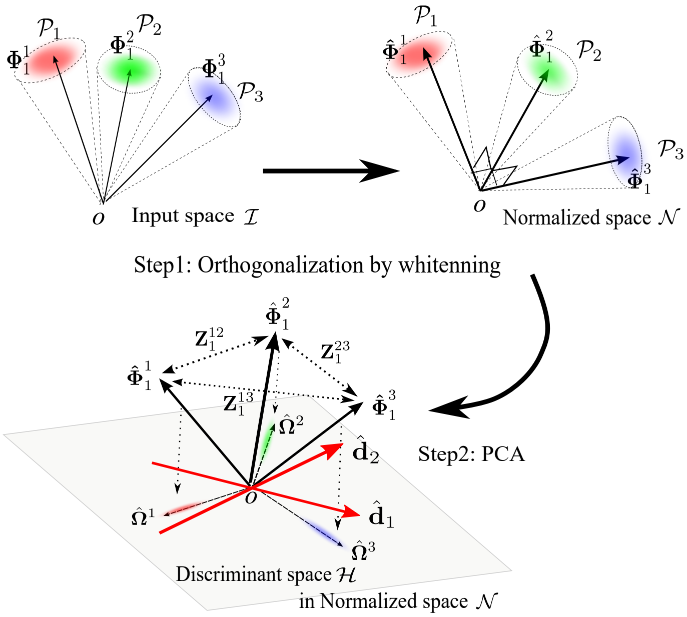


Fig. 5. Discrimination mechanism of gFDA, which consists of two processes: 1) whitening: all the orthogonal basis vectors $\{\hat{\phi}_i^c\}$ of three class subspaces are orthogonalized to each other in the normalized space $\mathcal{N}$, 2) PCA: only three first basis vectors $\{\hat{\phi}_1^c\}_{c=1}^3$ of each class are extracted and the difference vectors $\mathbf{Z}_1^{ij}$ between them are calculated. Then, the orthogonal basis vectors,$\hat{\mathbf{d}}_1$ and $\hat{\mathbf{d}}_2$, of the discriminant $\mathcal{H}$ are obtained by applying PCA to a set of the three difference vectors. Finally, all the data are projected onto the discriminant space $\mathcal{H}$.

$C-1$ principal component vectors $\{\hat{\mathbf{d}}_i\}_{i=1}^{C-1}$ from $\mathbf{\Sigma}_A = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} \mathbf{z}_1^{ij}\mathbf{z}_1^{ij\,T}$, since the rank of $\mathbf{\Sigma}_A$ is $C-1$. Note that $\mathbf{\Sigma}_A$ can be also represented by $\mathbf{A}^T\mathbf{\Sigma}_{B3}\mathbf{A}$. $\{\hat{\mathbf{d}}_i\}_{i=1}^{C-1}$ span the discriminant space $\mathcal{H}$ in the normalized space $\mathcal{N}$.

In the input space $\mathcal{I}$, the discriminant space $\mathcal{H}$ is spanned by the linearly transformed principal component vectors, $\{\mathbf{A}\hat{\mathbf{d}}_i(=\mathbf{d}_i)\}_{i=1}^{C-1}$.

5.2 Dual forms of the objective function

The objective function f_g of our simplified Fisher criterion is represented as a generalized eigenvalue problem for the matrix product ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}$. In the following, we prove that the objective function can also be represented as a simpler regular eigenvalue problem for the linear combination of $\mathbf{\Sigma}_{B3}$ and $\mathbf{\Sigma}_{W4}$ under the same setting as in the previous section. We consider a set of C N-dimensional class subspaces in $\mathbb{R}^{\hat{L}}$.

The flow of our proof is summarized as follows:

C1. $C-1$ eigenvalues of matrix ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3} \in \mathbb{R}^{\hat{L}\times\hat{L}}$ are all equal to C without depending on the dimensionality of each class subspace.

C2. The characteristic C1 above leads to the following equivalent relationship: ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}\mathbf{d} = C\mathbf{d} \;\Leftrightarrow\; (\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3})\mathbf{d} = 0\mathbf{d}$, where we note that in the former equation we need to take the eigenvectors corresponding to $C-1$ largest eigenvalues, while in the latter we need to take the eigenvectors corresponding to $C-1$ smallest eigenvalues (zero).

The two sets of eigenvectors obtained from the two eigenvalue problems in C2 are different. In fact, those in the first set are not orthogonal, since the matrix ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}$ is not symmetric. In contrast, the eigenvectors in the latter are orthogonal to each other, since matrix $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$ is symmetric. However, the two subspaces spanned by the respective sets of the eigenvectors coincide completely. Therefore, we will confirm that gFDA has dual forms of objective function.

Proof of C1. The characteristic C1 can be proved as follows: ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}$ has the same eigenvalues as $\mathbf{\Sigma}_A = \mathbf{A}^T\mathbf{\Sigma}_{B3}\mathbf{A}$, where $\mathbf{A}$ is the whitening such that $\mathbf{A}^T\mathbf{\Sigma}_{W4}\mathbf{A} = \mathbf{I}$, as described in the previous section. $\mathbf{\Sigma}_A$ is represented with the difference vectors between the C first orthonormal basis vectors, $\{\hat{\phi}_1^c\}$, which are orthogonalized by whitening $\mathbf{A}$:

$$\mathbf{\Sigma}_A = \sum_{i,j=1,i<j}^{C} (\hat{\phi}_1^i - \hat{\phi}_1^j)(\hat{\phi}_1^i - \hat{\phi}_1^j)^T . \tag{23}$$

Let $\hat{\mathbf{\Sigma}}_A \in \mathbb{R}^{C\times C}$ be the autocorrelation matrix of the difference vectors among the standard bases $\{\mathbf{e}_1,\ldots,\mathbf{e}_C\}$ of $\mathbb{R}^C$:

$$\hat{\mathbf{\Sigma}}_A = \sum_{i,j=1,i<j}^{C} (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T . \tag{24}$$

Since both $\{\hat{\phi}_1^c\}_{c=1}^C$ and $\{\mathbf{e}_1,\ldots,\mathbf{e}_C\}$ are orthonormal bases of $\mathbb{R}^{\hat{L}}$ and $\mathbb{R}^C$, respectively, they span two C-dimensional subspaces with the same geometrical structure. Therefore, the two autocorrelation matrices $\mathbf{\Sigma}_A$ and $\hat{\mathbf{\Sigma}}_A$ have the same C eigenvalues, though their matrix sizes are different.

$\hat{\mathbf{\Sigma}}_A$ can be written as

$$\hat{\mathbf{\Sigma}}_A = \begin{pmatrix} C-1 & -1 & \cdots & -1 \\ -1 & C-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & C-1 \end{pmatrix}, \tag{25}$$

$$= C\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{26}$$

In the above equation, the first matrix has C eigenvalues of C and the second one has one C and $C-1$ zeros as the eigenvalues. Hence, matrix $\hat{\mathbf{\Sigma}}_A$ has $C-1$ eigenvalues of C as non-zero eigenvalue. Therefore, we can confirm that $\mathbf{\Sigma}_A$ has $C-1$ eigenvalues of C as well.

Proof of C2. Next, we shall prove characteristics C2. By substituting $\lambda = C$ into Eq.(22), we obtain

$$\mathbf{\Sigma}_{B3}\mathbf{d} = C\mathbf{\Sigma}_{W4}\mathbf{d}. \tag{27}$$

Further, we can rewrite the equation as

$$(\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3})\mathbf{d} = 0 = 0\mathbf{d}, \tag{28}$$

where, by considering that $\mathbf{d}$ is not a zero vector, $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$ has $C-1$ zero eigenvalues.

This characteristic means that the eigenspace (null space) of $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$ corresponding to zero eigenvalue is equivalent to that of ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}$ corresponding to the eigenvalue of C. In other words, the null space spanned by the $C-1$ eigenvectors of $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$ corresponding to zero eigenvalues coincides with the discriminant space $\mathcal{H}$ spanned by the $C-1$ eigenvectors of ${\mathbf{\Sigma}_{W4}}^{-1}\mathbf{\Sigma}_{B3}$.

In summary, we can generate a discriminant space $\mathcal{H}$ of gFDA by solving a simpler regular eigenvalue problem of $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$. Here, we reiterate that the linear combination form can mitigate the SSS problem of FDA, since it can be stably calculated independently of the number of sample data and the dimension of vector space, unlike the matrix product form. In the following, we use $\hat{\mathbf{G}}$ to indicate $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$.

5.3 Geometrical structure of gFDA

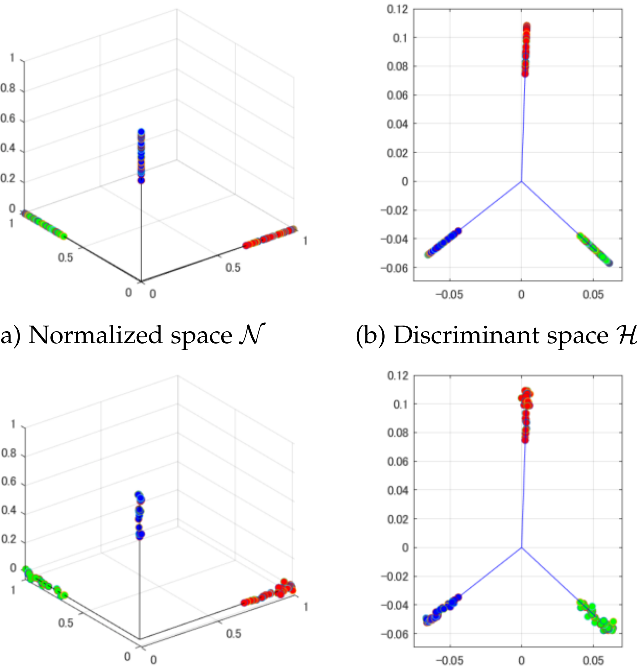
We describe the geometrical structure of gFDA, which can explain the robustness of gFDA in dealing with the small sample image recognition, and lead to natural introduction of data normalization on discriminant space $\mathcal{H}$.

Invariance to variation within class subspace: As described in the two-step process, in a normalized space $\mathcal{N}$, only the first basis vectors $\{\hat{\boldsymbol{\phi}}_1^c\}$ are selected from the orthogonalized basis vectors $\{\hat{\boldsymbol{\phi}}_i^c\}$ of all the class subspaces, and the remaining basis vectors $\{\hat{\boldsymbol{\phi}}_i^c\}_{i=2}^{d_c}$ are discarded. This operation results in that all the data of class c are projected onto only $\{\hat{\boldsymbol{\phi}}_1^c\}$ in the normalized space $\mathcal{N}$ as shown in Fig.6a, when all the data of class c are completely contained within the c-th class subspace spanned by $\{\boldsymbol{\phi}_i^c\}_{i=1}^{d_c}$.

In the process above, the subspace representation provides the robustness against the small sample size to gFDA. Consider that low d-dimensional class subspaces generate the sample data of each class, respectively. In this case, we can determine each class subspace from only d independent samples of each class, which alleviates the difficulty of the small sample image recognition. Such a situation corresponds to that the illumination subspace of an object contains any object images under various illumination conditions as described in Sec.1. For a static object, the minimum number of necessary sample data is from 3 to 9, corresponding to the dimensionality of the illumination subspace of the object [21], [22], [23].

However, as it is in general difficult to generate such a perfect illumination subspace in practical applications, the projected data points of the c-th class in the normalized space $\mathcal{N}$ can have nonzero components on the basis vectors $\{\hat{\boldsymbol{\phi}}_1^{c'}\}(c' \neq c)$ of other classes. As a result, they are projected at a remove from $\{\hat{\boldsymbol{\phi}}_1^c\}$ as shown in Fig.6c. Nevertheless, we should note that the degrees of their deviations are still very small. This geometrical relationship remains in the discriminant space $\mathcal{H}$ as shown in Figs.6b and d.

Moreover, we note that the above geometrical mechanism of gFDA holds without depending on selecting the first basis vector (mean vector) of a class subspace. In other words, the first basis vector does not necessarily need to coincide with the true mean vector. This also contributes to further enhancement of the robustness of gFDA against the small sample size. However, we need data normalization as described below, because the norms of the projections can get shortened when the first basis vector is significantly different from the true mean vector.



(a) Normalized space $\mathcal{N}$ (b) Discriminant space $\mathcal{H}$

(c) Normalized space $\mathcal{N}$ (d) Discriminant space $\mathcal{H}$

Fig. 6. Effectiveness of normalization: (a) and (b) show the projections of data of 3 classes in the normalized spaces and discriminant space, respectively, where all data of each class are completely contained within its class subspace. (c) and (d) show the projections of data of the 3classes, where some component of data are not covered by the class subspaces.

Normalization of projection data: In the above described process, the variation of the projections in the direction of $\hat{\boldsymbol{\phi}}_1^c$ can necessarily remain even if we could generate an exact illumination subspace of class c. Namely, we cannot in principle remove them. A valid way for ignoring this extra variation is to normalize the orthogonal projection $\pi(\mathbf{x}) = ({\mathbf{d}_1}^T\mathbf{x}, {\mathbf{d}_2}^T\mathbf{x}, \cdots, {\mathbf{d}_{C-1}}^T\mathbf{x}) \in \mathbb{R}^{C-1}$ of data $\mathbf{x}$ on the discriminant space $\mathcal{H}$. To get the maximum performance out of gFDA in a classification task, besides the data normalization of input data, we essentially need to incorporate the normalization of orthogonal projection $\pi(\mathbf{x})$ of data $\mathbf{x}$ on the discriminant space $\mathcal{H}$ into the mechanism of gFDA, where the normalization is defined as $\pi(\mathbf{x})/||\pi(\mathbf{x})||$.

5.4 Generation of discriminant space

As stated in the previous section, all of $C-1$ discriminant vectors $\{\mathbf{d}_i\}_{i=1}^{C-1}$ have the same discriminant ability, C; all the valid eigenvalues in Eq.(22) are C according to the characteristic C1. This characteristic suggests that each individual vector of $C-1$ $\mathbf{d}_i$ does not have much meaning, rather a subspace spanned by them should be considered to be essential. Hence, we define a subspace spanned by $C-1$ discriminant vectors as discriminant space $\mathcal{H}$, where the discriminant vectors are orthogonalized to each other by using the Gram-Schmidt orthonormalization.

5.5 Small sample size problem of FDA

In many practical applications of image recognition, the dimension L of data is much larger than the total number of data, n. In such a case, Eqs.(6) and (22) cannot be solved since $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_{W4}$ are singular. This issue is called the small sample size (SSS) problem [8] of FDA, which has been well known as a critical limitation of FDA.

To overcome the SSS problem, various types of extensions of FDA have been proposed [11]. There are two typical solutions widely used due to their simple implementation. One is to use PCA to reduce the dimension before applying FDA [8]. The other is to add a regularization term to matrix $\mathbf{\Sigma}_W$ [12] as $f(\mathbf{d}) = \frac{\mathbf{d}^T\mathbf{\Sigma}_B\mathbf{d}}{\mathbf{d}^T(\mathbf{\Sigma}_W+\delta\mathbf{I})\mathbf{d}}$, where δ is a parameter that controls the strength of the regularization and $\mathbf{I}$ is the identity matrix.

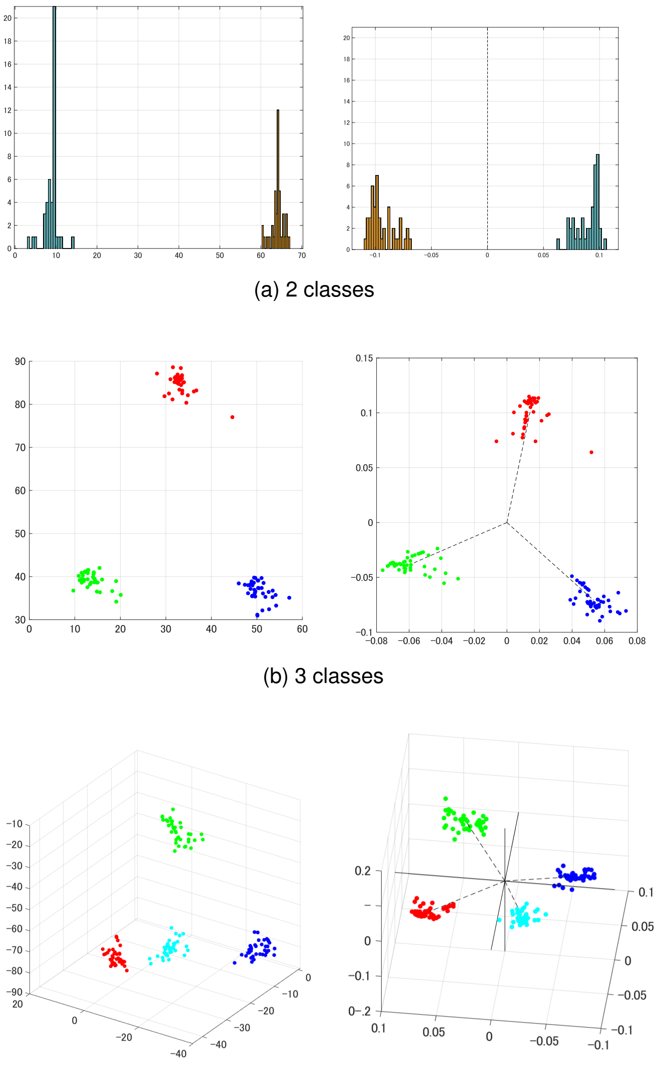
In addition to the above simple methods, many other extensions based on the null and range spaces of the within-class and between-class scatter matrices have been proposed to circumvent the SSS problem [11], [14], [15], [16], [17], [19], [20]. In particular, among them, nullLDA [13] has been known as a useful method to avoid the SSS problem. In this method, all the data are first projected onto the null space of the within-class scatter matrix, and then a between-class scatter matrix is calculated from the projections. Finally, a discriminant space is obtained by solving the eigenvalue problem of the between-class scatter matrix.

For gFDA, the objective function f_g can be rewritten in the linear combination form of the two symmetric matrices in Eq.(28). This enables gFDA to avoid the SSS problem and work even with only one sample without any modification. However, in terms of computational cost, it is desirable to use the PCA based dimensionality reduction together, as it can largely reduce the data dimension. For gFDA, the dimension of the original dimension can be in fact reduced to the number of the orthonormal basis vectors without losing any structural information of the class subspaces, since the orthonormal basis vectors over all the classes are linearly independent, assuming no overlap among class subspaces.

5.6 Comparison of FDA and gFDA

Fig.7 shows the comparisons of projections onto discriminant spaces generated by FDA (left) and gFDA (right), where we used sets of face images from the Yale face database. In this database, each subject class contains 45 frontal face images which were collected under different lighting conditions. It is known that all the possible images of a face under various lighting conditions are contained in an illumination cone [30]. The illumination cone of a subject can be accurately approximated by a convex cone formed by a set of nine frontal face images of the subject under nine specific lighting conditions. These nine images are called the 9PL images [30] in the Yale face database. Further, the illumination cone is contained in a 9-dimensional illumination subspace, which can be generated by applying PCA to a set of the 9PL images. Hence, a 9-dimensional illumination subspace can in principle contain other 36 images under different illumination conditions. For more details of the Yale database, see Section 8.

We used the 9PL images as the learning data, and used the remaining face images as the test data. The dimension of each class subspace was set to 9. In Fig.7, a row represents the case of 2, 3 or 4 classes. We used FDA with PCA dimensionality reduction [8], since the original FDA cannot be used under this setting due to the SSS problem. In contrast, gFDA avoids the SSS problem by using the linear combination form. We can see that the distributions of the projections by FDA and gFDA are similar in all the cases.



(a) 2 classes

(b) 3 classes

(c) 4 classes

Fig. 7. The projections of face classes from the Yale face database by FDA (left) and gFDA (right), where a row represents the case of 2, 3 or 4 classes.

6 Connection of gFDA and GDS projection

In this section, we show a close connection between gFDA and GDS projection. To this end, we prove that gFDA is equivalent to GDS projection with a small correction item.

6.1 GDS projection with a small correction term

According to the new form of $\mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$ for gFDA presented in the previous section, we notice that gFDA is closely related to GDS projection [1] that uses $C-1$ smallest eigenvector of only $\mathbf{G}(=\mathbf{\Sigma}_{W4})$, because $||\mathbf{\Sigma}_{W4}||_F \gg ||\frac{1}{C}\mathbf{\Sigma}_{B3}||_F$.

To prove this relationship, we introduce a pair of vectors, $\mathbf{z}_i^{jk}$ and $\mathbf{z'}_i^{jk}$, between the i-th orthonormal basis vectors

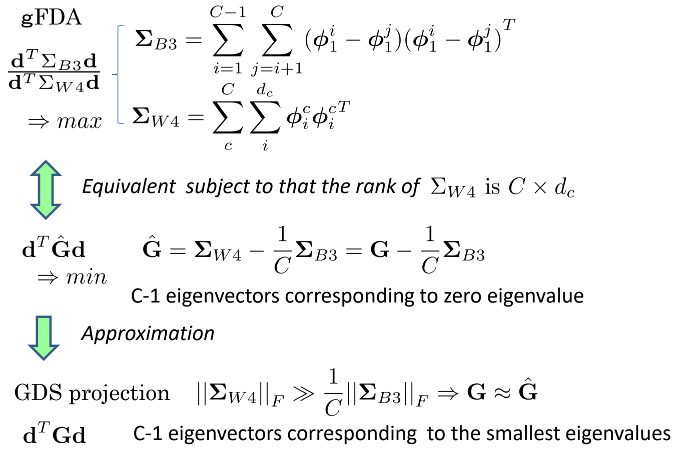

gFDA
$\frac{\mathbf{d}^T\Sigma_{B3}\mathbf{d}}{\mathbf{d}^T\Sigma_{W4}\mathbf{d}} \Rightarrow max$
$\mathbf{\Sigma}_{B3} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C}(\phi_1^i - \phi_1^j)(\phi_1^i - \phi_1^j)^T$
$\mathbf{\Sigma}_{W4} = \sum_{c}^{C}\sum_{i}^{d_c}\phi_i^c\phi_i^{cT}$
Equivalent subject to that the rank of Σ_{W4} is $C \times d_c$
$\mathbf{d}^T\hat{\mathbf{G}}\mathbf{d} \Rightarrow min$
$\hat{\mathbf{G}} = \mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3} = \mathbf{G} - \frac{1}{C}\mathbf{\Sigma}_{B3}$
C-1 eigenvectors corresponding to zero eigenvalue
Approximation
GDS projection
$||\mathbf{\Sigma}_{W4}||_F \gg \frac{1}{C}||\mathbf{\Sigma}_{B3}||_F \Rightarrow \mathbf{G} \approx \hat{\mathbf{G}}$
$\mathbf{d}^T\mathbf{G}\mathbf{d}$
C-1 eigenvectors corresponding to the smallest eigenvalues

Fig. 8. Connection between gFDA and GDS projection.

ϕ_i^j and ϕ_i^k of classes j and k, where $\mathbf{z}_i^{jk} = \phi_i^j - \phi_i^k$ and $\mathbf{z'}_i^{jk} = \phi_i^j + \phi_i^k$. Note that $\mathbf{z}_i^{jk}\mathbf{z}_i^{jk^T} + \mathbf{z'}_i^{jk}\mathbf{z'}_i^{jk^T} = 2(\phi_i^j\phi_i^{j^T} + \phi_i^k\phi_i^{k^T})$.

With $\{\mathbf{z}\}$ and $\{\mathbf{z'}\}$, we rewrite matrix $\mathbf{G}(= \mathbf{\Sigma}_{W4})$ in Eq.(1) for GDS projection as follows:

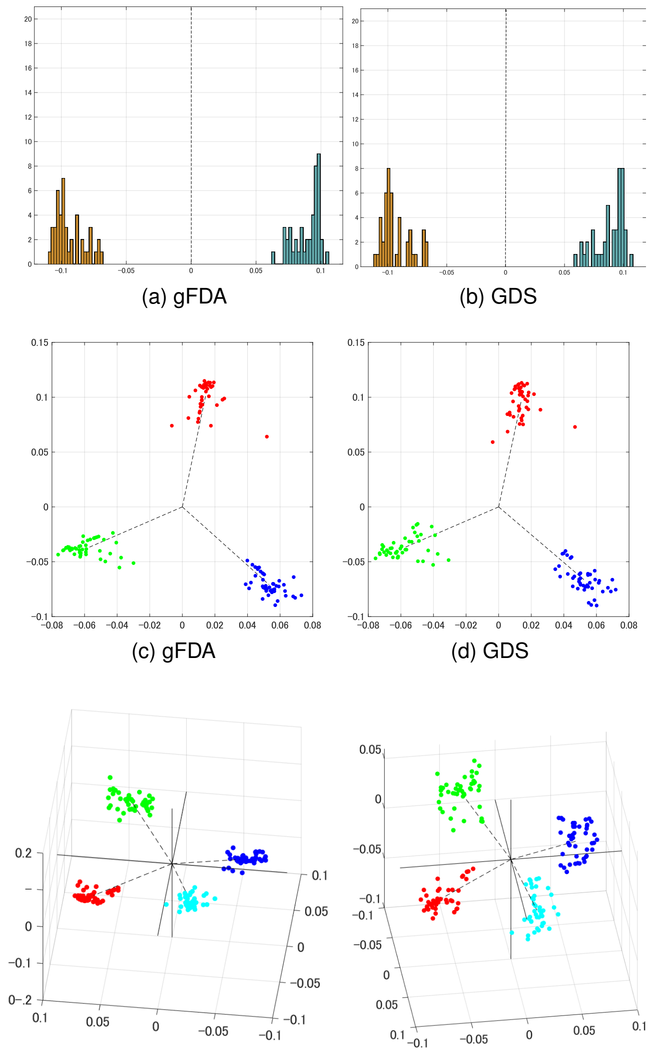
$$\begin{aligned}\mathbf{G} &= \mathbf{\Sigma}_{W4} = \sum_{j=1}^{C}\sum_{i=1}^{d_j}\phi_i^j\phi_i^{j^T}, &(29)\\ &= \frac{1}{2(C-1)}\Big(\sum_{j,k,j<k}^{C}(\mathbf{z}_1^{jk}\mathbf{z}_1^{jk^T} + \mathbf{z'}_1^{jk}\mathbf{z'}_1^{jk^T}\Big)\\ &\quad + \sum_{j=1}^{C}\sum_{i=2}^{d_j}\phi_i^j\phi_i^{j^T}, &(30)\\ &= \frac{1}{2(C-1)}\mathbf{\Sigma}_{B3} + \mathbf{\Sigma}_A. &(31)\end{aligned}$$

In Eq.(31), $\mathbf{\Sigma}_{B3} = \sum_{j,k,j<k}^{C}\mathbf{z}_1^{jk}\mathbf{z}_1^{jk^T}$ (Eq.(19)) and $\mathbf{\Sigma}_A = \frac{1}{2(C-1)}\sum_{j,k,j<k}^{C}\mathbf{z'}_1^{jk}\mathbf{z'}_1^{jk^T} + \sum_{j=1}^{C}\sum_{i=2}^{d_j}\phi_i^j\phi_i^{j^T}$. $||\mathbf{\Sigma}_A||_F \gg ||\mathbf{\Sigma}_{B3}||_F$ and $||\mathbf{\Sigma}_{W4}||_F > ||\mathbf{\Sigma}_A||_F$. Hence, $||\mathbf{\Sigma}_{W4}||_F \gg \frac{1}{C}||\mathbf{\Sigma}_{B3}||_F$. For real data sets, for example, in the case with three 9-dimensional subspaces of three face classes shown in Fig.9, $||\mathbf{\Sigma}_{W4}||_F = 7.2383$ is about 143 times as $\frac{1}{3}||\mathbf{\Sigma}_{B3}||_F = 0.0504$. They support the above magnitude relationship.

From the standpoint of GDS projection, $\frac{1}{C}||\mathbf{\Sigma}_{B3}||$ can be regarded as a small correction on itself. Thus, we can regard gFDA as GDS projection with a small correction term of $\frac{1}{C}\mathbf{\Sigma}_{B3}$. Fig.8 summarizes the whole flow of the simplification from gFDA to GDS projection that has been discussed so far. The close connection suggests that GDS projection has a discriminant ability and the robustness against the SSS problem as well as gFDA. Fig.9 shows the comparison between gFDA and GDS projection on the examples that were used for the comparison of FDA and gFDA in Fig.7. We can see high similarity between the results of these two methods.

6.2 Geometry gap between gFDA and GDS

We now discuss the relationship between gFDA and GDS projection in more detail. In the same form as Eq.(31), we rewrite $\hat{\mathbf{G}}$ for gFDA as follows:



(a) gFDA (b) GDS

(c) gFDA (d) GDS

(e) gFDA (f) GDS

Fig. 9. Visualization of the projections by gFDA and GDS projection in the cases of two, three and four classes from the Yale face database.

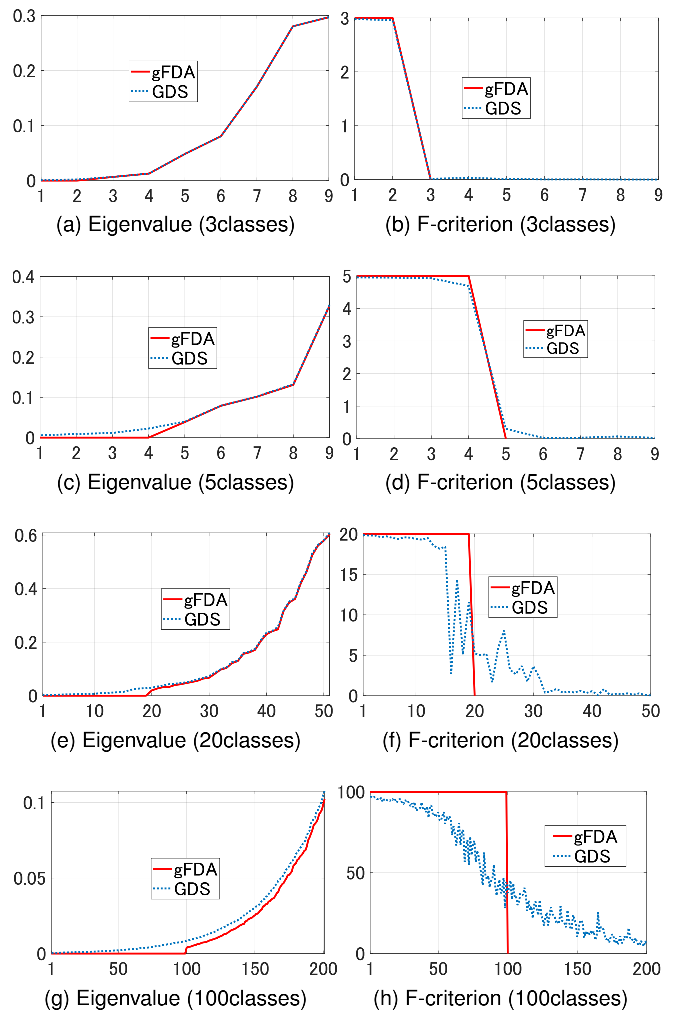
$$\hat{\mathbf{G}} = \Big(\frac{1}{2(C-1)} - \frac{1}{C}\Big)\mathbf{\Sigma}_{B3} + \mathbf{\Sigma}_{A1}. \quad (32)$$

We notice that only the weights on $\mathbf{\Sigma}_{B3}$, that is, on the difference vectors $\{\mathbf{z}_1^{jk}\}$, are different between Eq.(31) and Eq.(32), which are $\frac{1}{2(C-1)} > 0$ and $\frac{1}{2(C-1)} - \frac{1}{C} \leq 0$, respectively. This difference in the weights produces a geometrical gap between gFDA and GDS projection. We measure the gap by using an index σ, which is defined as follows:

$$\sigma = \frac{\frac{1}{2(C-1)} - \big(\frac{1}{2(C-1)} - \frac{1}{C}\big)}{\frac{1}{2(C-1)}} = 2(1 - \frac{1}{C}). \quad (33)$$

The value of σ becomes larger starting from 1.0 in the case of C=2 toward 2.0 as the class number C increases. Thus, we can see that the gap increases as C gets larger.

To show this characteristic more clearly, we compared the distributions of eigenvalues of $\mathbf{G}$ and $\hat{\mathbf{G}}$, which were generated from a set of C 3-dimensional class subspaces. The left column of Fig.10 shows the eigenvalues from $\mathbf{G}$



(a) Eigenvalue (3classes) (b) F-criterion (3classes)

(c) Eigenvalue (5classes) (d) F-criterion (5classes)

(e) Eigenvalue (20classes) (f) F-criterion (20classes)

(g) Eigenvalue (100classes) (h) F-criterion (100classes)

Fig. 10. Comparison of gFDA and GDS projection in terms of the distribution of eigenvalues and our Fisher-like criterion.

and $\hat{\mathbf{G}}$ in the ascending order. We can see that the $C-1$ smallest eigenvalues are zero for gFDA as described earlier. The eigenvalue distributions in both cases are almost the same when the number of classes, C, is small as three or five. However, we can see that the difference between the two distributions becomes larger as the number of class increases. In conjunction with this observation, the value of σ also increases from 1.33 (3 classes) to 1.98 (100 classes).

We also compared the gap between these two methods by using another index. The right column of Fig. 10 indicates the Fisher's ratio on each basis vector $\mathbf{d}_i$ of the discriminant spaces $\mathcal{H}$, which was calculated from the projections of the basis vectors of all the class subspaces onto $\mathbf{d}_i$. gFDA has the discriminant ability of $C\times(C-1)$, since gFDA has $C-1$ eigenvalues of C according to the characteristic C1, which corresponds to the area under the line over the interval from 1 to $C-1$. In the same way, we regard the area under the curve over the interval from 1 to $C-1$ as the discriminant ability of GDS projection. We can see that the discriminant ability of gFDA and GDS projection are almost equivalent in terms of this index, when the class number C is small. However, as C becomes larger, the gap between them gets larger, that is, the discriminant ability of GDS decreases. This suggests that the discriminant ability of GDS projection could be insufficient when the dimension of GDS is set to $C-1$ as in gFDA, especially in the case with a large number of classes. Thus, we propose to use a larger dimension than $C-1$ for GDS. To be concrete, we take the N_d basis vectors such that the total sum of the Fisher's ratios over them gets larger than a specified threshold value, $\beta = C(C-1)\times\gamma$, where γ is empirically set to a value in the range between 0.8 and 0.95.

7 Nonlinear extension by using kernel function

In this section, we first review the nonlinear subspace with a Gaussian kernel function. Then, we extend gFDA to kernel gFDA (KgFDA).

7.1 Generation of nonlinear class subspace

Let ψ be the nonlinear function that maps a feature vector $\mathbf{x} \in \mathbb{R}^L$ onto a high dimensional feature space $\mathcal{F}$. We generate the c-th class subspace on the feature space $\mathcal{F}$ by applying PCA without data centering to a set of n_c data of the c-th class, $\{\mathbf{x}_l^c\}_{l=1}^{n_c}$, mapped onto $\mathcal{F}$.

The d_c orthonormal basis vectors $\{\mathbf{e}_i^c\}_{i=1}^{d_c}$ of the d_c-dimensional nonlinear subspace $\mathcal{V}_c$ are represented by the linear combination of $\{\psi(\mathbf{x}_l^c)\}_{l=1}^{n_c}$ as $\mathbf{e}_i^c = \sum_{l=1}^{n_c} \mathrm{a}_{il}^c\,\psi(\mathbf{x}_l^c)$, where the coefficient a_{il}^c is the l-th component of the eigenvector $\mathbf{a}_i^c$ corresponding to the i-th largest eigenvalue λ_i of the Gram matrix $\mathbf{K} \in \mathbb{R}^{n_c\times n_c}$. The vector of $\mathbf{a}_i^c \in \mathbb{R}^{n_c}$ is normalized to satisfy $\lambda_i(\mathbf{a}_i^c\cdot\mathbf{a}_i^c)$=1. The elements $[\mathrm{k}_{ll'}]$ of matrix $\mathbf{K}$ are defined as $(\psi(\mathbf{x}_l^c)\cdot\psi(\mathbf{x}_{l'}^c)) = k(\mathbf{x}_l^c,\mathbf{x}_{l'}^c)$, where we use a Gaussian kernel function $k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma^2}\right)$ so that the mapped data have a unit length.

7.2 Generation of Kernel gFDA

To extend gFDA to kernel gFDA (KgFDA), we consider the eigenvalue problem of the matrix, $\hat{\mathbf{G}} = \mathbf{\Sigma}_{W4} - \frac{1}{C}\mathbf{\Sigma}_{B3}$, on the feature space $\mathcal{F}$.

For the extension, we define $\mathbf{D} \in \mathbb{R}^{N_T\times N_T}$ as $\mathbf{E}^T\mathbf{E}$, where $\mathbf{E}$ is the matrix that contains all the basis vectors as columns: $\mathbf{E} = [\mathbf{e}_1^1 \ldots \mathbf{e}_{d_1}^1 \ldots \mathbf{e}_1^C \ldots \mathbf{e}_{d_C}^C]$ and $N_T = \sum_{c=1}^C d_c$. The element, $\mathrm{D}(i,j)$, of matrix $\mathbf{D}$ is denoted by the inner product $(\mathbf{e}_i^c\cdot\mathbf{e}_j^{c'})$ between the i-th orthonormal basis vector $\mathbf{e}_i^c$ of the subspace of class c and the j-th orthonormal basis vector $\mathbf{e}_j^{c'}$ of the subspace of class c'. The value of this inner product can be calculated as $\sum_{l=1}^{n_c}\sum_{l'=1}^{n_{c'}} \mathrm{a}_{il}^c\mathrm{a}_{jl'}^{c'}k(\mathbf{x}_l^c,\mathbf{x}_{l'}^{c'})$. With the matrices $\mathbf{D}$ and $\mathbf{E}$, the eigenvalue problem of $\hat{\mathbf{G}}$ can be written as the following generalized eigenvalue problem (see the supplemental material for detailed derivation):

$$\mathbf{H}\hat{\mathbf{b}} = \hat{\beta}\mathbf{D}\hat{\mathbf{b}}, \quad (34)$$

$$\mathbf{H} = \mathbf{D}\mathbf{D}^T - \frac{1}{C}\hat{\mathbf{B}}, \quad (35)$$

$$\hat{\mathbf{B}} = \sum_{i=1}^{C-1}\sum_{j=i}^{C}(\mathbf{D}_{\omega(i)}-\mathbf{D}_{\omega(j)})(\mathbf{D}_{\omega(i)}-\mathbf{D}_{\omega(j)})^T, \quad (36)$$

where $\mathbf{D}_{\omega(i)} = \mathbf{E}^T\mathbf{e}_1^i \in \mathbb{R}^{N_T}$, $\hat{\mathbf{B}} \in \mathbb{R}^{N_T\times N_t}$, $\mathbf{D}\mathbf{D}^T \in \mathbb{R}^{N_T\times N_T}$ and the eigenvector $\hat{\mathbf{b}} \in \mathbb{R}^{N_T}$ is normalized to satisfy that $\hat{\mathbf{b}}^T\mathbf{D}\hat{\mathbf{b}} = 1$. By comparing this eigenvalue problem with that of KGDS [1], $\mathbf{D}\mathbf{b} = \beta\mathbf{b} \Rightarrow \mathbf{D}\mathbf{D}^T\mathbf{b} = \beta\mathbf{D}\mathbf{b}$, we can see that the difference between them is the part of $\frac{1}{C}\hat{\mathbf{B}} \in \mathbb{R}^{N_T \times N_T}$, which can be regarded as a correction term against KGDS.

The i-th orthonormal basis vector $\hat{\mathbf{d}}_i^{\psi}$ of a discriminant space $\hat{\mathcal{D}}^{\psi}$ produced by KgFDA can be represented as $\hat{\mathbf{d}}_i^{\psi} = \sum_{j=1}^{N_T} \hat{\mathrm{b}}_{ij}\mathbf{E}_j$, where $\hat{\mathrm{b}}_{ij}$ indicates the j-th element of the eigenvector $\hat{\mathbf{b}}_i$ corresponding to i-th smallest eigenvalue (zero) of Eq.(34).

7.3 Projection onto KgFDA

Let $\mathbf{E}_j$ be the $\eta(j)$-th basis vector of class $\zeta(j)$ in matrix $\mathbf{E}$. The projection of the mapped data $\psi(\mathbf{x})$ onto the basis vectors $\hat{\mathbf{d}}_i^{\psi}$ can be calculated from an input data $\mathbf{x}$ as

$$\begin{aligned}
(\hat{\mathbf{d}}_i^{\psi} \cdot \psi(\mathbf{x})) &= \sum_{j=1}^{N_T} (\hat{\mathrm{b}}_{ij}\mathbf{E}_j \cdot \psi(\mathbf{x})), && (37) \\
&= \sum_{j=1}^{N_T} \sum_{l=1}^{n_{\zeta(j)}} \hat{\mathrm{b}}_{ij}\mathrm{a}_{\eta(j)l}^{\zeta(j)} (\psi(\mathbf{x}_l^{\zeta(j)}) \cdot \psi(\mathbf{x})), && (38) \\
&= \sum_{j=1}^{N_T} \sum_{l=1}^{n_{\zeta(j)}} \hat{\mathrm{b}}_{ij}\mathrm{a}_{\eta(j)l}^{\zeta(j)} k(\mathbf{x}_l^{\zeta(j)}, \mathbf{x}), && (39)
\end{aligned}$$

where we can easily compute $k(\mathbf{x}_l^{\zeta(j)}, \mathbf{x})$ through $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$. Finally, the projection $\pi(\psi(\mathbf{x}))$ of the mapped $\psi(\mathbf{x})$ onto the $C-1$ dimensional discriminant space produced by KgFDA is represented as $\pi(\psi(\mathbf{x})) = (z_1,\ z_2, \dots,\ z_{C-1})^T,\ z_i = (\hat{\mathbf{d}}_i^{\psi} \cdot \psi(\mathbf{x}))$.

8 Evaluation experiments

In this section, we first verify the validity of our heuristic relationship on real data. We then evaluate the effectiveness of gFDA, GDS projection and their nonlinear extensions from the following aspects: 1) Fisher-like discriminant ability, 2) their performance on face and object recognition under small samples, and 3) the validity of the combination of our methods and CNN features. In all the subsequent experiments, the length of data is always normalized to a unit before applying a classification method.

8.1 Validity of our heuristic relationship

We verify the validity of the heuristic relationship: the equivalence in terms of direction between the first orthonormal basis vector $\boldsymbol{\phi}_1$ and the mean vector $\mathbf{m}$ of a class distribution. For this purpose, we measured the normalized correlation coefficient between $\boldsymbol{\phi}_1$ and $\mathbf{m}$ in each set of 9PL images of 29 subjects from the Yale face database B+, which were used in Sec.5.6. The average of the correlation coefficients of all the subjects was 0.99932. For the CMU face database, which will be used for evaluation later, the average value of 120 subject classes was 0.99998, where each class consists of 20 frontal face images under different illuminations. We confirmed that CNN features can also satisfy the heuristic relationship with higher correlations than 0.999 in all the cases with ResNet18 and ResNet50 in Secs.8.5 and 8.6. These high correlations support the validity of our heuristic relationship.

TABLE 1
Comparison of the three discriminant spaces by FDA, sFDA and gFDA.

2 classes		3 classes	
sFDA↔FDA	gFDA↔FDA	sFDA↔FDA	gFDA↔FDA
$\cos\theta_1$=0.988	$\cos\theta_1$=0.937	$\cos\theta_1$=0.970	$\cos\theta_1$=0.904
-	-	$\cos\theta_2$=0.904	$\cos\theta_2$=0.888

In addition, we measured the degree of coincidence between the discriminant spaces that were generated by the original FDA, sFDA with $\frac{\mathbf{d}^T\mathbf{\Sigma}_{B2}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W2}\mathbf{d}}$, and gFDA with $\frac{\mathbf{d}^T\mathbf{\Sigma}_{B3}\mathbf{d}}{\mathbf{d}^T\mathbf{\Sigma}_{W4}\mathbf{d}}$. As the original FDA cannot work on our experimental setting due to the small sample size problem, we used regLDA [12] instead. Table 1 shows the cosines of the canonical angles between them. We can see that gFDA can be still regarded as a reasonable approximation of the original FDA despite the considerable simplification of the original Fisher criterion.

8.2 Fisher-like discriminant ability

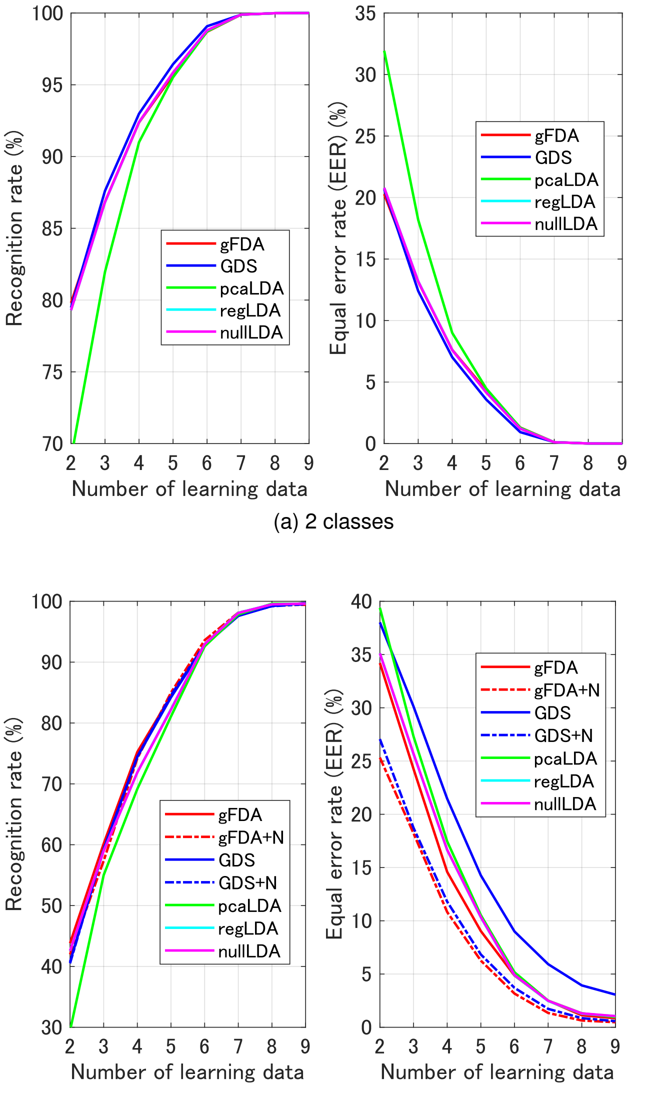
We verify that gFDA and GDS projection have inherited the high discriminant ability from FDA on the Yale face database B+.

Experimental settings: The Yale face database B+ consists of face images of 38 subjects, where these images were acquired under 64 different lighting conditions in nine different poses [31]. We selected 29 individuals from the database; these individuals' images appear across the four subsets. In the evaluation, we used only the frontal face images, so that our data set contains 1,035 images of 29 subjects under 45 different lighting conditions. We converted the cropped images of 640 × 480 pixels to images of 32 × 24 pixels and normalized the image vectors.

We conducted evaluation experiments on this database. The 9PL images of each subject, which were described in Sec.5.6, were used for learning and the remaining 36 images were used for testing. To verify the robustness of the methods against few sample data, we changed the number of learning data from two to nine, where they were randomly selected from the nine 9PL images. We repeated this sampling 60 times and calculated the averages of all the results obtained as the final one.

We conducted experiments under the above condition in the cases of two and 29 classes, using a nearest neighbor classifier with the l^2 norm between an input data and each class mean. In the case of two classes, we randomly selected 25 pairs of two classes from 29 subject classes and used the average results as the final performance. We evaluated the performances of the methods in terms of the recognition rate (%) and equal error rate (EER) (%).

We used three typical variants of modified FDA: pcaLDA [8], regLDA [12] and nullLDA [13], since the original FDA cannot work on this experimental setting due to the SSS problem. In the sequel, we will refer to these modified LDAs as original LDAs for simplicity. We also evaluated the performances of gFDA and GDS with the normalization of the projected data, which we denote as gFDA+N and GDS+N, respectively. The dimension of class subspace of



(a) 2 classes

(b) 29 classes

Fig. 11. Performances of the different methods on the Yale face database in terms of recognition rate and equal error rate (%).

TABLE 2
Mean performances (%) of different methods in terms of recognition rate and equal error rate (%).

	gFDA	GDS	pcaLDA	regLDA	nullLDA
Rate	94.16	94.43	92.91	94.13	94.13
EER	5.85	5.59	8.13	5.89	5.88

(a) 2 classes

	gFDA	gFDA+N	GDS	GDS+N	pcaLDA	regLDA	nullLDA
Rate	81.67	81.13	81.00	81.03	78.04	80.74	80.72
EER	11.43	8.28	15.72	8.91	13.08	12.18	12.20

(b) 29 classes

gFDA and GDS projection was set to the number of learning data of each class. For GDS projection, the value of γ was set to 0.90 to determine the dimension of GDS. For pcaLDA, the sum of squared residuals in PCA used for dimension reduction were 1e-2 and 1e-9 for two and 29 classes, respectively. The value of δ in regLDA was set to 1e-4 for both. The parameters were empirically determined using the learning data set. We used a nearest neighbor classifier with the l^2 norm between an input data and each class mean as a classifier.

Experimental results and consideration: Figs.11a and b show the results of the different methods in the case of two and 29 classes, respectively. In the figures, the horizontal axis indicates the number of learning data and the vertical axes in the left and right panels indicate the recognition rate (%) and EER (equal error rate) (%), respectively. Table 2 shows the mean performances of different methods.

In the case of two classes, the performances of gFDA and GDS projection are almost the same as those of regLDA and nullLDA, as shown in Fig.11a and Table 2a. This supports that gFDA certainly inherits the discriminant ability of the original FDA and furthermore GDS projection inherits the discriminant ability from gFDA. Thanks to the characteristic of the illumination subspaces generated from the 9PL images, the recognition rates of all the methods were nearly perfect when using all the 9PL images as learning data.

The performance of pcaLDA is slightly lower than the other methods, especially when the number of learning data is small. This can be ascribed to the fact that the dimension reduction based on the PCA could not estimate a meaningful within-class covariance from very few learning data. For example, only six learning data were used for conducting PCA when the number of learning data is three for each class.

The close relationship among the methods can be also observed in the case of 29 classes as shown in Fig.11b and Table 2b. gFDA+N and GDS+N outperform the other methods in terms of EER according to a statistical t-test with a significance level of 0.01. Although the performance of GDS projection was lower than those of the other methods in terms of EER, it has been visibly improved by the normalization.

8.3 Performance evaluation on face recognition

We conducted the classification on larger scale data from the CMU Multi-PIE face database.

Experimental settings: The CMU Multi-PIE face database consists of face images of 337 subjects, captured from 15 viewpoints with 20 different lighting conditions in four recording sessions [28]. In the experiment, we used frontal face images of 128 subjects across all four sessions. We took a sub-sampled image of size 36×36 pixels from an original image, where we cropped this image by reference to the two inner corners of the eyes and the tip of the nose. The vectorized images were normalized. In classification, we used a nearest neighbor classifier with the l^2 norm between an input and each class mean.

For each subject on a session, n images randomly sampled from 20 images were used for learning and the remaining 20 $-n$ images for testing. We evaluated the performance of the methods while increasing the number of learning data, n, from two to ten. We repeated this evaluation 60 times for each n and then calculated the average of the results. Further, we conducted the same evaluation on the three remaining sessions and took the average of the results on the four sessions as a final recognition performance. For

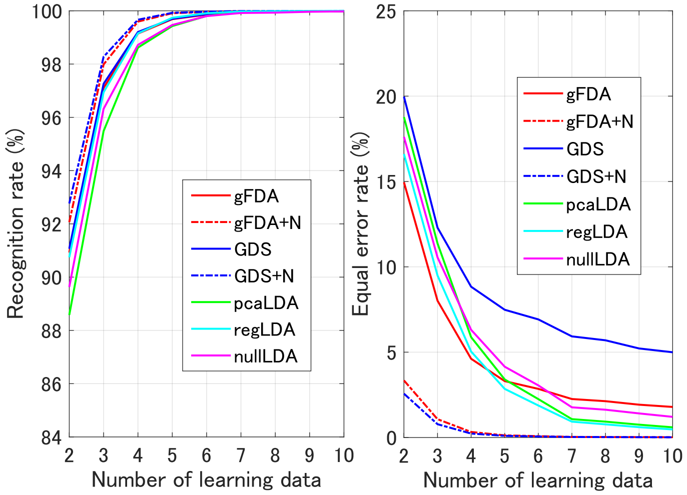


Fig. 12. Comparison of different methods in terms of recognition rate and equal error rate (%) in 128 classes from the CMU face database.

TABLE 3
Mean performances (%) of different methods on CMU dataset.

	gFDA	GDS	pcaLDA	regLDA	nullLDA
Rate	98.51	98.55	97.98	98.49	98.19
EER	4.65	8.59	5.00	4.29	5.30

(a) Without the normalization of the projected data

	gFDA+N	GDS+N	pcaLDA+N	regLDA+N	nullLDA+N
Rate	98.83	98.95	97.90	97.34	98.17
EER	0.56	0.43	0.99	1.23	0.92

(b) With the normalization of the projected data

gFDA and GDS projection, we set the dimension of class subspace to the number of learning data, n. For GDS, the value of γ was set to 0.90. The sum of squared residuals of PCA used in pcaLDA was 2e-4. For regLDA, δ was 1e-4. The other parameters were set to the same values that were used in the previous experiment.

Experimental results and consideration: Table 3 shows the mean performances of the methods in terms of recognition rate and equal error rate (EER). The overall trend remains more or less the same as the previous experiments: gFDA is comparatively superior to FDAs when the number of learning data is small (from two to four), the performance of gFDA is slightly lower than those of the FDAs when n is large (over 5), and GDS projection is poorer than the other methods particularly in terms of EER. However, the effectiveness of the normalization of projection data in this case is much clearer in comparison with those in the previous cases; gFDA+N and GDS+N significantly outperform the other methods in both indexes. This result supports that the normalization of projection data is intrinsically required to get the best performance out of gFDA and GDS projection. Moreover, in the extreme case that only one learning data is available, pcaLDA and nullLDA cannot work in principle, but gFDA+N and GDS+N can still work with the recognition rates of 53.2% and 47.8%, and the EERs of 15.9% and 17.1%, respectively.

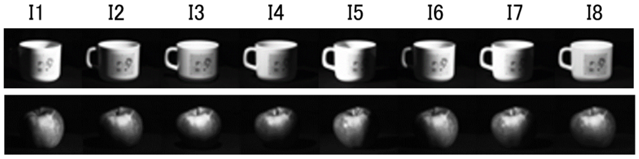


Fig. 13. Examples of two objects under different eight illuminations from our ALOI300 (illumination direction collection).

8.4 Performance evaluation on 3D object recognition

We verify the effectiveness of gFDA and GDS on ALOI database (illumination direction collection), focusing on the robustness against the small sample image recognition.

Experimental settings: The illumination direction collection consists of one thousand small 3D object images captured under different eight illumination conditions generated by turning on five lights [26]. The five illumination conditions of I1-I5 were yielded by turning on only one out of the five lights. The conditions of I6 and I7 were yielded by turning on two right and two left lights at the sides of the object, respectively. I8 was yielded by turning on all the lights. In this way, for each object, a total of eight images were captured under the different illumination conditions as shown in Fig.13. To consider the situation with small sample size, we used only the first 300 objects, and refer to it as ALOI300 in this paper.

The poses of objects were fixed during the capturing, while the illumination condition changed. Thus, we can expect that the subspace representation works effectively as in the previous experiments on front face images, However, the database contains many objects with an image set that cannot be accurately represented by a linear subspace, because they do not satisfy the necessary conditions of 3D convex shape and Lambertian reflection for subspace representation. To address this issue, we introduced KgFDA and KGDS based on nonlinear subspace with more flexible and richer representation ability.

We evaluated the performances of our methods with the normalization including their nonlinear extensions in comparison with FDA and its various extensions for addressing the SSS problem: regularized LDA (regLDA) [12], pcaLDA [8], nullLDA [13], eigenfeature regularization method (EFR) [16], maximum uncertainty LDA (mLDA) [18], improved Direct LDA (idLDA) [17], approximate LDA (aLDA) [15]. For EFR, mLDA, idLDA and aLDA, we used the MATLAB codes from the LDA-SSS package [11]. Besides the linear methods, we evaluated the performance of kernel Fisher discriminant analysis (KFDA) [32] to verify the advantage of subspace representation in the nonlinear extensions. We used a nearest neighbor classifier with the l^2 norm between an input data and each class mean in the classification.

In the experiment, for each object, n images of I1-In were used for learning and the remaining $8-n$ images were used for testing, where n changed from two to five. Accordingly, the dimensions of the subspaces of each object class was n. The value of σ^2 for the Gaussian kernel function was set to 30 for both KgFDA and KGDS. For KFDA, the value of σ^2 and the regularization coefficient of δ were set to 1.0 and 1e-6, respectively. These parameters were empirically

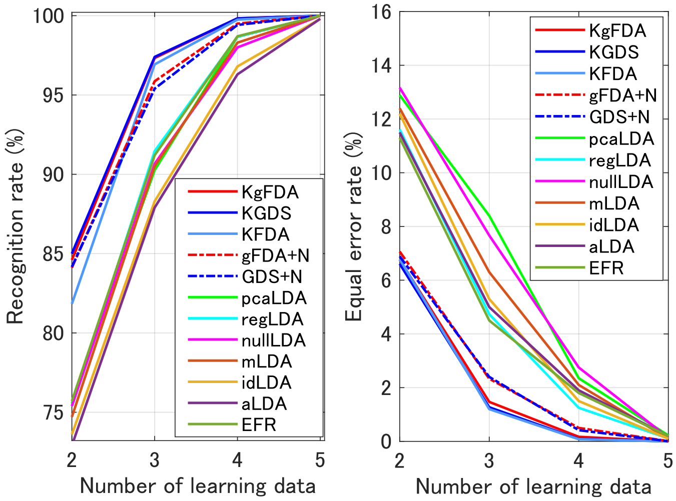


Fig. 14. Performance shifts of different methods, depending on the number of learning data on ALOI300.

TABLE 4
Performances of different methods on ALOI300, which are ranked in descending order of the mean recognition rates.

	Rates (%)				EER (%)			
	2	3	4	mean	2	3	4	mean
(1) **KGDS** [1]	85.0	97.4	99.8	**94.1**	6.7	1.3	0.1	**2.7**
(2) **KgFDA**	84.6	97.3	99.8	**93.9**	6.8	1.5	0.2	**2.8**
(3) **gFDA+N**	84.1	95.9	99.5	**93.2**	7.1	2.3	0.5	**3.3**
(4) **GDS+N** [1]	84.2	95.4	99.4	**93.0**	6.9	2.4	0.4	**3.2**
(5) KFDA [32]	81.8	96.9	99.8	92.8	6.8	1.2	0.1	2.7
(6) regLDA [12]	75.4	91.5	98.7	88.5	11.6	4.7	1.3	5.9
(7) EFR [16]	75.7	91.2	98.7	88.5	12.4	6.3	2.1	6.9
(8) pcaLDA [8]	75.6	90.3	98.7	88.2	12.9	8.4	2.4	7.9
(9) nullLDA [13]	75.4	90.7	98.0	88.0	13.1	7.7	2.8	7.9
(10) mLDA [18]	74.7	90.5	98.3	87.8	12.2	5.3	1.5	6.3
(11) idLDA [17]	73.6	88.3	96.8	86.2	11.5	5.0	1.9	6.1
(12) aLDA [15]	72.9	87.9	96.3	85.7	11.3	4.5	1.8	5.9

determined using the learning set. KgFDA and KGDS used the normalization of the projected data in all the subsequent experiments, although we will indicate them without "+N".

Experimental results and consideration: Figure 14 shows the performances of the different methods in terms of recognition rates and equal error rate (EER). The overall trend of the comparison result looks almost the same as that of the front face images. The figure also clearly shows the advantage of gFDA+N and GDS+N against FDA and its various extensions. This indicates that our methods based on the subspace representation can work effectively against 3D objects with more complicated shapes than face.

For the robustness against the small sample image recognition, we can confirm the superiority of our method in terms of both indexes, particularly when the number of learning data is extremely small, 2 and 3, although all the methods have achieved the perfect performance when using five learning data. Moreover, we can see that KgFDA and KGDS further improve the performances of gFDA+N and GDS+N as expected. This improvement shows high representation ability of nonlinear subspace over linear subspace.

TABLE 5
Performances of different methods when using image sets of five objects for learning in terms of recognition rate and EER (%).

	ResNet50	KFDA	KgFDA	KGDS	FDA	gFDA	GDS
Rate	90.81	91.98	92.08	92.09	90.56	82.37	91.14
EER	3.68	3.53	3.67	3.46	6.21	11.52	5.95

8.5 3D object recognition with CNN features

We have shown the validity of our methods in face and 3D object classifications under a relatively simple setup where each object has a fixed pose under varying illumination condition. To achieve high performance in more complex tasks like 3D object recognition from multi-view images, however, it is preferred to utilize more powerful features instead of raw images. For this purpose, we employed CNN features extracted from a fully connected layer of convolutional neural networks (CNN), ResNet50, which is one of the popular CNNs, as an input of our methods.

We evaluated the performances of our methods and KFDA using CNN features in comparison with that of ResNet50 for classification of object category on ETH80 dataset [33]. ETH80 dataset consists of eight different object categories, each of which has ten types of objects. The images of each object were captured from 41 viewpoints. We resized the gray images to 32 × 32 pixels and converted them to 1024-dimensional vectors. For each category, we used randomly selected five objects for learning and the remaining five for testing. That is, each category subspace was generated from $205 = (41\times5)$ images of the five objects selected. The total number of testing images was $205 = (41\times5)$. We repeated this process for 20 times.

To extract more discriminative CNN features from images, we slightly modified the architecture of the original ResNet50 trained by the ImageNet database [34] as follows: we replaced the final 1000-way fully connected (FC) layer with a 1024-way FC layer and added a ReLU activation layer behind the fc1024. Furthermore, we added a 8-way FC layer that matches the number of categories and a softmax layer to the 1024-way FC layer. We then fine-tuned the modified ResNet50 using our learning set. Finally, by feeding an image to our fine-tuned ResNet50, we extracted the output of the 1024-way FC layer of it as a 1024-dimensional CNN feature vector.

The parameters of the methods were empirically determined using the learning data set. First of all, for KFDA, the parameter of Gaussian kernel function, σ^2, and the regulation value, δ, were set to 2.4 and 1e-6, respectively. For KgFDA and KGDS, σ^2 was also set to 2.4 so that our focus remains on evaluating the effectiveness of subspace representation by using the same kernel mapping. Finally, the dimensions of class subspaces were determined to be 20 for KgFDA and KGDS, and the dimension of the generalized difference subspace of KGDS was set to 20. Note that the dimensions of the discriminant spaces of KFDA and KgFDA were automatically set to 7 (=the number of categories-1). The classification was performed by using a nearest neighbor classifier with the l^2 norm between an input data and each class mean.

Table 5 shows the results of the different methods by using five objects for learning. We can see that the nonlinear

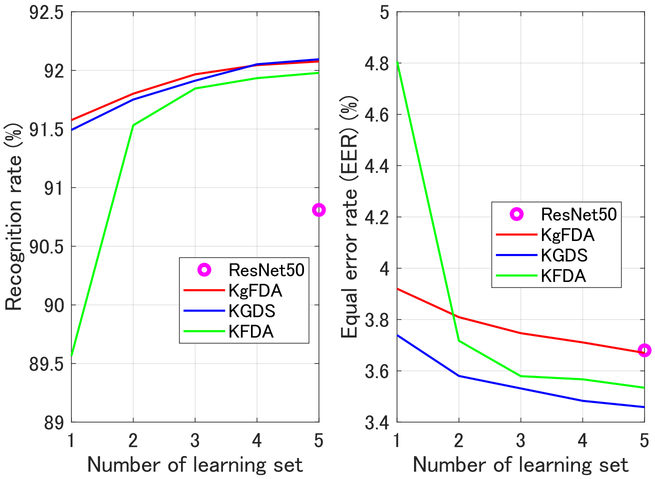


Fig. 15. Performances of KFDA, KgFDA and KGDS using CNN features on ETH80 in terms of recognition rate and EER (%) for different numbers of training objects.

methods, KFDA, KgFDA, and KGDS, have enhanced the performance of the ResNet50. In contrast, the linear methods, FDA, gFDA and GDS could not noticeably improve the performance of the ResNet50.

These results can be explained as follows: the prediction layer of the fc1024 and fc8 layers with the softmax layer work as a linear classification based on the least mean square error when applying cross-entropy loss. Such a classification method is known to correspond to Fisher discriminant analysis [35]. This correspondence implies that CNN features extracted from the module are already optimized with respect to the Fisher discriminant criterion. Thus, there is little room for improving the performance of the ResNet50 even if we replaced the prediction layer with any other type of linear classification module based on Fisher discriminant criterion including our linear methods.

Next, we evaluated the robustness of our methods against the small sample image recognition by changing n, the number of objects used for learning from one to five. In this case, $n\times41$ images of randomly selected n objects were used for generating each category class subspace. We repeated this classification for 30 times. Figure 15 shows the mean recognition rates and equal error rates (EER) of different methods according to the number of objects, n. The small circles there indicate the mean performance of our fine-tuned ResNet50 using five objects, which was obtained in the same way as the other methods. The difference between the three methods is not remarkable when n is two or larger. However, KgFDA and KGDS significantly outperform KFDA for both indexes when using only one object for learning. This supports that the subspace representation also provides high stability against the small sample image recognition in the case of using CNN features.

8.6 General classification ability with CNN features

Finally, we verify the general versatility of our methods with CNN features on a few standard datasets, MNIST [29] and CIFAR10 [27], beyond the small sample image recognition.

TABLE 6
Performances of different methods on MNIST.

	ResNet18	FDA	gFDA	GDS	KFDA	KgFDA	KGDS
Rate(%)	99.17	99.14	98.83	98.83	99.57	99.56	99.61
EER(%)	0.28	0.36	0.58	0.60	0.16	0.16	0.15

(a) Performance comparison of FDA based methods with ResNet18. For the parameters, gFDA and GDS: d_c=2; KFDA: σ^2=0.7 and δ=1e-6; KgFDA and KGDS: σ^2=0.6 and d_c=700.

	KFDA	KgFDA	KGDS	CKN [36]	DSN [37]	CapsNet [38]
Rate (%)	99.70	99.67	99.72	99.60	99.60	99.75
EER (%)	0.18	0.12	0.11	-	-	-

(b) Performances of our kernel methods with data compression and typical DNNs without data augmentation. '-' indicates not available. For the parameters, KFDA: σ^2=0.67 and δ=1e-6; KgFDA and KGDS: σ^2=0.45 and d_c=490.

The MNIST dataset consists of 70,000 images with 28 × 28 pixels of handwritten digits in 10 classes, where each has 7,000 images. The sets of 6,000 images and 1,000 images were used for training and testing, respectively. The CIFAR10 dataset consists of 60,000 color images with 32×32 pixels in ten object classes with complex background and without segmentation. Each class has 6,000 images. The sets of 5,000 and 1,000 images are used for learning and testing, respectively.

To obtain the CNN features, we replaced the final 1000-way FC layer of the pre-trained ResNet by using the ImageNet with a 10-way FC layer. We used the ResNet18 and ResNet50 for MNIST and CIFAR10, respectively. Then, we fine-tuned our ResNets using the training images. In this way, we extracted 512-dimensional and 2048-dimensional CNN feature vectors by feeding an image to our fine-tuned ResNet18 and ResNet50, respectively. The classification was performed by using a nearest neighbor classifier with the l^2 norm between an input data and each class mean.

Table 6a shows the performances of different methods on MNIST. We can observe that the performances of our linear methods are lower than that of the ResNet18, while that of FDA is at the similar level as the ResNet18. This can be explained in the aspect of the optimization as described earlier. On the other hand, our nonlinear methods, KFDA, KgFDA, and KGDS have succeeded in further enhancing the performance of ResNet18 in terms of both recognition rate and EER. The recognition rates of 99.56% and 99.61% of KgFDA and KGDS, respectively, are competitive with typical methods based on DNN without data augmentation, such as convolutional kernel network (CKN) [36] and deeply-supervised nets (DSN) [37], as shown in Table 6b.

Moreover, to reduce the computational cost of our kernel methods, we decreased the number of learning data from 50,000 to 5,000 by using a method [39] based on the k-means clustering before the kernel mapping. Interestingly, their performances have further increased and achieved the state-of-the-art accuracy [38] as shown in Table 6b.

Table 7 shows the performances of the different methods on CIFAR10. The overall trend is similar to that on MNIST; the linear methods do not work on this general object classification as the handwriting digits classification. However, we can see the superiority of the nonlinear methods against

TABLE 7
Performances of different methods on CIFAR10. For the parameters, gFDA and GDS: d_c=2; KFDA: σ^2=1.0 and δ=1e-6; KgFDA and KGDS: σ^2=0.7 and d_c=700.

	ResNet50	FDA	gFDA	GDS	KFDA	KgFDA	KGDS
Rate (%)	93.21	93.12	91.84	91.78	93.87	93.72	93.92
EER(%)	2.35	0.261	0.340	0.368	2.21	2.58	2.34

the ResNet50.

The results on both MNIST and CIFAR10 imply that there still remains a room for further enhancing the prediction layer module based on the cross-entropy loss. This also suggests that it is a promising research direction to incorporate a combination of a nonlinear mapping and the mechanism of gFDA/GDS projection into the architecture of CNN in an end-to-end fashion. In fact, the validity of this approach has been partly demonstrated in the work of [40] where GDS projection is used as filter banks in a semi-supervised shallow network without end-to-end learning. More concrete discussion about how to implement this idea in a form of end-to-end learning is beyond the scope of this paper. However, a valid direction which we could take in the future is suggested in several work [41], [42], [43] where the discriminant mechanism based on Fisher criterion is incorporated into learning of deep neural networks in an end-to-end fashion.

On MNIST and CIFAR10, the performances of KgFDA and KGDS are more or less on a par with that of KFDA. This is because the amount of learning data is sufficient to estimate the within-class variation, unlike in the previous experiments with the small sample size. Nevertheless, we should recall that our methods can work stably even with a small number of samples, as clearly shown in the previous experiments. This characteristic can be viewed as a significant advantage in practical applications where collecting a large amount of data is difficult.

9 Conclusion

In this paper, we revealed that the orthogonal projection of data onto a generalized difference subspace (GDS), called GDS projection, can function as a discriminant feature extraction through a similar mechanism as the Fisher discriminant analysis (FDA). In this process, we introduced geometrical Fisher discriminant analysis (gFDA), which is a discriminant analysis based on a simplified Fisher criterion. We then proved that gFDA is equivalent to GDS projection with a small correction term. This equivalence ensures GDS projection to inherit the discriminant ability from FDA by regarding gFDA as an intermediate concept between them. To further enhance the performances of gFDA and GDS projection, we proposed to normalize the projected vectors onto the discriminant spaces. Moreover, we discussed two useful extensions of our methods: 1) nonlinear extension by using kernel trick, 2) the combination with CNN features.

Extensive experiments using the extended Yale B+ database,the CMU face database, and the ALOI with small sample size showed that gFDA and GDS projection have high discriminant ability as well as FDA, and further their extensions with normalization have equivalent or higher performance than various types of variants of modified FDA. Furthermore, experiments on ETH80, MNIST and CIFAR10 demonstrated the effectiveness of using CNN features as input of our methods.

As future work, we shall evaluate how much our methods are robust against imbalanced data in comparison with FDA and its variants. Secondly, we shall explore how to reduce the high computational cost for the kernelization effectively. We consider the reduction method based on k-means clustering used in Sec.8.6 as a good starting basis.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 16H02842. We thank Lincon Souza, Bernardo Gatto, Erica Shimomoto and Rui Zhu for their helpful comments. We also thank the anonymous reviewers for their constructive comments which greatly helped improve the paper.

References

[1] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 2164–2177, 2015.

[2] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Proc. 11th International Symposium of Robotics Research (ISRR2003)*, 2003, pp. 192–201.

[3] K. Fukui, B. Stenger, and O. Yamaguchi, "A framework for 3D object recognition using the kernel constrained mutual subspace method," in *Proc. Asian Conference on Computer Vision*, vol. 3852, 2006, pp. 315–324.

[4] K. Fukui, "Subspace methods," in *Computer Vision: A Reference Guide*. Springer International Publishing, 2020.

[5] L. S. de Souza, B. B. Gatto, J.-H. Xue, and K. Fukui, "Enhanced grassmann discriminant analysis with randomized time warping for motion recognition," *Pattern Recognition*, vol. 97, 2020.

[6] B. B. Gatto, E. M. dos Santos, A. L. Koerich, K. Fukui, and W. S. S. Júnior, "Tensor analysis with n-mode generalized difference subspace," *Expert Systems with Applications*, vol. 171:114559, 2021.

[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[8] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed. Academic Press, 1990.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[11] A. Sharma and K. K. Paliwal, "Linear discriminant analysis for the small sample size problem: an overview," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 3, pp. 443–454, 2015.

[12] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165–175, 1987.

[13] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, pp. 1713–1726, 2000.

[14] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.

[15] K. K. Paliwal and A. Sharma, "Approximate lda technique for dimensionality reduction in the small sample size case," *Journal of Pattern Recognition Research*, vol. 6, no. 2, pp. 298–306, 2011.

[16] X. Jiang, B. Manda, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, 2008.

[17] K. K. Paliwal and A. Sharma, "Improved direct lda and its application to dna microarray gene expression data," *Pattern Recognition Letter*, vol. 31, no. 16, pp. 2489–2492, 2010.

[18] C. E. Thomaz and D. F. Gillies, "A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition," in *Proc. Brazilian Symposium on Computer Graphics and Image Processing*, 2005, pp. 89–96.

[19] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of lda," in *International Conference on Pattern Recognition*, vol. 3, 2002, pp. 29–32.

[20] W. Zhao, R. Chellappa, and P. Phillips, "Subspace linear discriminant analysis for face recognition," University of Maryland, Tech. Rep., 1999.

[21] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 218–233, 2003.

[22] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible lighting conditions?" *International Journal of Computer Vision*, vol. 28, pp. 1–16, 1998.

[23] J. R. Beveridge, B. A. Draper, J.-M. Chang, M. Kirby, H. Kley, and C. Peterson, "Principal angles separate subject illumination spaces in YDB and CMU-PIE," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 351–363, 2009.

[24] H. Hotelling, "Relation between two sets of variables," *Biometrica*, vol. 28, pp. 322–377, 1936.

[25] S. N. Afriat, "Orthogonal and oblique projectors and the characteristics of pairs of vector spaces," *Mathematical Proc. Cambridge Philosophical Society*, vol. 53, no. 4, pp. 800–816, 1957.

[26] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The amsterdam library of object images," *International Journal of Computer Vision*, vol. 61(1), pp. 103–112, 2005.

[27] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[28] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, 1998, pp. 2278–2324.

[30] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 684–698, 2005.

[31] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 643–660, 2001.

[32] S. Mika, G. Rätsch, J. Weston, B. Schälkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, 1999, pp. 41–48.

[33] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, pp. 409–415.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] C. H. Park and H. Park, "A relationship between linear discriminant analysis and the generalized minimum squared error solution," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 2, pp. 474–492, 2005.

[36] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 2627–2635.

[37] C.-Y. Lee, S. Xie, P. Gallagher, ZhengyouZhang, and Z. Tu, "Deeply-supervised nets," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 38, 2015, pp. 562–570.

[38] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3859–3869.

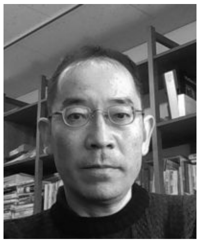
[39] Y. Ohkawa and K. Fukui, "Hand shape recognition using the distributions of multi-viewpoint image sets," *IEICE Transactions on Information and Systems*, vol. E95-D, no. 6, pp. 1619–1627, 2012.

[40] B. B. Gatto, L. S. Souza, E. M. dos Santos, K. Fukui, W. S. S. Júnior, and K. V. dos Santos, "A semi-supervised convolutional neural network based on subspace representation for image classification," *EURASIP J. Image Video Process.*, vol. 2020, no. 1:22, 2020.

[41] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 4, pp. 596–608, 2012.

[42] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

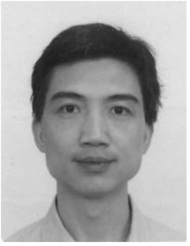
[43] D. Díaz-Vico and J. R. Dorronsoro, "Deep least squares fisher discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2019.



Kazuhiro Fukui received his M.E. (Mechanical Engineering) from the Kyushu University in 1988. He joined Toshiba Corporate R&D Center and served as a senior research scientist at Multimedia Laboratory. He received his PhD degree from the Tokyo Institute of Technology in 2003. He is currently a professor in the Faculty of Engineering, Information and Systems at the University of Tsukuba. His research interests include the theory of machine learning, computer vision, pattern recognition and their applications. He has served as a program committee member at many pattern recognition and computer vision conferences, including as an Area Chair of ICPR'12, 14, 16 and 18. He is a member of the SIAM.

Naoya Sogi received Ms. Eng. and Dr. Eng. from the University of Tsukuba in 2019 and 2022, respectively. He is currently a researcher at the NEC Visual Intelligence Research Laboratories. His interests include the theory of computer vision, pattern recognition, machine learning and applications of these theories.

Takumi Kobayashi received Ms. Eng. from the University of Tokyo in 2005 and Dr. Eng. from the University of Tsukuba in 2009. He was a researcher at Toshiba Corporation in 2006 and then joined National Institute of Advanced Industrial Science and Technology (AIST), Japan, in 2007. He is also a professor in the Department of Computer Science, Graduate School of Systems and Information Engineering at the University of Tsukuba. His research interest includes pattern recognition, machine learning and computer vision.



Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a professor in the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He is an Associate Editor of IEEE TCSVT, IEEE TCYB, and IEEE TNNLS.

Atsuto Maki is a Professor of Computer Science at KTH Royal Institute of Technology, Sweden. He obtained BEng and MEng in electrical engineering from Kyoto University and the University of Tokyo, respectively, and his PhD degree in computer science from KTH in 1996. Previously he was an associate professor at the Graduate School of Informatics, Kyoto University, and then a senior researcher at Toshiba's Cambridge Research Lab in the UK. His research interests cover a broad range of topics in computer vision and machine learning, including representation learning. He has been serving as a program committee member at major computer vision conferences, e.g. as an area chair of ICCV and ECCV.