Linear Discriminative Image Processing Operator Analysis

Toru Tamaki, Bingzhi Yuan, Kengo Harada, Bisser Raytchev, Kazufumi Kaneda
Hiroshima University, Japan
tamaki@hiroshima-u.ac.jp, yuan@eml.hiroshima-u.ac.jp

Abstract

In this paper, we propose a method to select a discriminative set of image processing operations for Linear Discriminant Analysis (LDA) as an application of the use of generating matrices *representing image processing operators acting on images. First we show that generating matrices can be used for formulating LDA with increasing training samples, then analyze them as image processing operators acting on 2D continuous functions for compressing many large generating matrices by using PCA and Hermite decomposition. Then we propose Linear Discriminative Image Processing Operator Analysis, an iterative method for estimating LDA feature space along with a discriminative set of generating matrices. In experiments, we demonstrate that discriminative generating matrices outperform a non-discriminative set on the ORL and FERET datasets.*

1. Introduction

Pattern recognition techniques need a lot of training samples for better performance, however, collecting a huge number of samples is usually expensive or impractical. When there are not enough training samples, the problem is sometimes called small sample size problem, or single (training) image per person in face recognition. Many studies have been done for this problem [1, 2, 3], and many of them use an approach to increase the number of training samples by synthetically generating new training samples by various image processing operations, as surveyed in [4].

Focusing on Linear Discriminant Analysis (LDA), the method proposed in this paper finds the most discriminative set of image processing operations to increase training samples. We represent an image processing applied to an image $\boldsymbol{x}$ by a matrix G (called *generating matrix*) to generate a new training sample $\boldsymbol{x}'$ represented by $\boldsymbol{x}' = G\boldsymbol{x}$. This equation is similar to those in [5, 6] in which they have formulated transformation between two images by the equation for parameter estimation such as optical flow, depth [5], or object pose [6]. In contrast, we intend to use the equation for recognition. We first analyze image processing operations that can be represented by linear operators, then implement those operations as a generating matrix. We decompose a matrix implementation of such an operator into two Hermite matrices. Then, we propose a method to find the most discriminative generating matrices by linear combination of *eigen*-generating matrices obtained by Principal Component Analysis (PCA). Those matrices are further compressed with Hermite decomposition. We call the proposed method Linear Discriminative Image Processing Operator Analysis (*LDIPOA*).

1.1. Related work

There are three directions of previous work related to the proposed method. The first is related to tangent propagation [7, 8] or convolutional neural networks [9]. Those approaches make a classifier (neural network) invariant or insensitive to small changes in the original image. Similar to those, our method also provides invariance to image processing operations. However, our approach finds an appropriate image processing for recognition rather than constructing a classifier invariant to image changes.

A second direction of related work is represented by metric learning [10, 11] or discriminative kernels [12] which put class information into classifier metrics. The main difference is that our approach pays attention to image processing operations, in other words feature extraction of holistic image features, while metric learning uses training samples to train a metric or distance.

A third group of related methods learns intra-personal variations using generic training samples [18, 19, 17]. Those methods assume that intra-personal variations can be learned from a separated dataset different from the training and test sets, while our approach emphasizes variations induced by practically possible image processing operations, in particular, those which can be represented by linear operators.

1.2. Contributions

Here we list some of the main contributions of this paper. First, we show that the feature space of LDA with increasing samples can be formulated by using generating matrices,

which means the same feature space is obtained without actually increasing samples if generating matrices are stored. Second, we show that, in general, linear image processing operators are normal operators that can be decomposed into eigenspaces. Therefore, operators can be constructed by a linear combination (like a vector space) rather than multiplication (like a group). Third, analyses of generating matrices are shown: they are almost symmetric or orthogonal, although those are a naive implementation of actual operators. Finally, LDIPOA is proposed to estimate LDA feature space and a set of discriminative generating matrices at the same time.

The organization of the paper is as follows. In section 2, we discuss how generating matrices work on training samples for LDA. Then, in section 3 we analyze image processing operations as linear operators acting on 2D continuous functions. Finally, eigen-generating matrices are obtained and each of them is decomposed into two Hermite matrices. In section 4, we propose an iterative method, LDIPOA, to find a discriminative set of generating matrices for LDA. In section 5, we show some experimental results.

2. Generating matrix with LDA

In this section, we discuss how generating matrices work on training samples for LDA and derive the LDA feature space.

2.1. LDA

Here, we briefly review LDA for c classes ($c \geq 2$): let $\mathcal{X}_i$ be a set of n_i samples of class ω_i in d dimensional space. For each class, the between-class scatter matrix S_B and within-class scatter matrix S_W are defined as follows:

$$S_W = \sum_{i=1}^{c} S_i,\ S_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T, \quad (1)$$

$$S_B = \sum_{i=1}^{c} (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T. \quad (2)$$

A $d \times \tilde{d}$ matrix A is used for dimensionality reduction to make $\tilde{d}$ dimensional features $\boldsymbol{y} = A^T \boldsymbol{x}$. Then, the scatter matrices of $\boldsymbol{y}$ are given as:

$$\widetilde{S}_i = \frac{1}{n_i} \sum_{\boldsymbol{y} \in \mathcal{Y}_i} (\boldsymbol{y} - \tilde{\boldsymbol{m}}_i)(\boldsymbol{y} - \tilde{\boldsymbol{m}}_i)^T = A^T S_i A, \quad (3)$$

$$\widetilde{S}_W = \sum_{i=1}^{c} \widetilde{S}_i = A^T S_W A, \quad (4)$$

$$\widetilde{S}_B = \sum_{i=1}^{c} (\tilde{\boldsymbol{m}}_i - \tilde{\boldsymbol{m}})(\tilde{\boldsymbol{m}}_i - \tilde{\boldsymbol{m}})^T = A^T S_B A. \quad (5)$$

A typical choice of a criterion to be maximized is the Rayleigh quotient $\frac{\mathrm{tr}(\widetilde{S}_B)}{\mathrm{tr}(\widetilde{S}_W)}$ which is equivalent to the following optimization problem: max $\mathrm{tr}(\widetilde{S}_B)$ such that $\widetilde{S}_W = I$. The solution is given by the generalized eigenvalue problem $S_B A = S_W A \Lambda$, where Λ is a diagonal matrix. Hence, the columns of A are the eigenvectors of ${S_W}^{-1} S_B$ which correspond to the largest $\tilde{d}$ eigenvalues.

However S_W becomes singular when d is large because the rank of S_W is smaller than $n - c$ (n is the number of all samples). Therefore, PCA is usually used to project data by a matrix P to a lower dimensional space before LDA, like in Fisherface [13]. A $d \times d'$ matrix P is obtained by eigen-decomposition of the covariance matrix X of all samples:

$$X = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathcal{X}} (\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^T. \quad (6)$$

The columns of P are the eigenvectors of X which correspond to the largest $d' (= n - c)$ eigenvalues.

Finally, the within/between class scatter matrices $\widetilde{S}_W$ and $\widetilde{S}_B$ of d' dimensional feature vectors $\boldsymbol{y} = A^T P^T \boldsymbol{x}$ are:

$$\widetilde{S}_W = A^T P^T S_W P A, \quad \widetilde{S}_B = A^T P^T S_B P A. \quad (7)$$

Then A is obtained by solving the eigenvalue problem $\left(P^T S_W P\right)^{-1} P^T S_B P$.

2.2. LDA with increasing samples

Next, we show that the use of generating matrices produces the same LDA feature space without actually increasing the number of samples.

Suppose J generating matrices $\{G_j\}$ (one of them is the identity matrix) are used to represent new images $\{\boldsymbol{x}_j\}$ from $\boldsymbol{x}$: $\boldsymbol{x}_j = G_j \boldsymbol{x}$. The average of all samples $\boldsymbol{m}'$ and class averages $\boldsymbol{m}'_i$ of ω_i are given by:

$$\boldsymbol{m}'_i = \frac{1}{J n_i} \sum_{j=1}^{J} \sum_{\boldsymbol{x} \in \mathcal{X}_i} G_j \boldsymbol{x} = \frac{1}{J} \sum_{j=1}^{J} G_j \boldsymbol{m}_i = \bar{G} \boldsymbol{m}_i, \quad (8)$$

$$\boldsymbol{m}' = \bar{G} \boldsymbol{m}, \quad (9)$$

where $\bar{G}$ is the average of $\{G_j\}$. By substituting these into the scatter matrices, we have

$$S'_i = \frac{1}{J n_i} \sum_{j=1}^{J} \sum_{\boldsymbol{x} \in \mathcal{X}_i} (G_j \boldsymbol{x} - \boldsymbol{m}'_i)(G_j \boldsymbol{x} - \boldsymbol{m}'_i)^T, \quad (10)$$

$$= \bar{G}(S_i - R_i)\bar{G}^T + \frac{1}{J} \sum_{j=1}^{J} G_j R_i G_j^T, \quad (11)$$

$$S'_W = \bar{G}(S_W - R_W)\bar{G}^T + \frac{1}{J} \sum_{j}^{J} G_j R_W {G_j}^T, \quad (12)$$

$$S'_B = \sum_{i}^{c} (\boldsymbol{m}'_i - \boldsymbol{m}')(\boldsymbol{m}'_i - \boldsymbol{m}')^T = \bar{G} S_B \bar{G}^T, \quad (13)$$

where $R_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{X}_i} \boldsymbol{x}\boldsymbol{x}^T$ is the autocorrelation matrix of class ω_i, and $R_W = \sum_i^c R_i$.

To project data to a $d' (= Jn - c)$ dimensional space, P

is obtained by PCA of

$$X' = \frac{1}{Jn}\sum_{j=1}^{J}\sum_{\boldsymbol{x}\in\mathcal{X}}\left(G_j\boldsymbol{x}-\bar{G}\boldsymbol{m}\right)\left(G_j\boldsymbol{x}-\bar{G}\boldsymbol{m}\right)^T, \quad (14)$$

$$= \bar{G}\left(X - R_{\mathrm{all}}\right)\bar{G}^T + \frac{1}{J}\sum_{j=1}^{J} G_j R_{\mathrm{all}} G_j^T, \quad (15)$$

where $R_{\mathrm{all}} = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}}\boldsymbol{x}\boldsymbol{x}^T$ is the autocorrelation matrix of all samples.

Now, scatter matrices $\widetilde{S}'_W$ and $\widetilde{S}'_B$ in the feature space are given with $\{G_j\}$ as follows:

$$\widetilde{S}'_i = \frac{1}{Jn_i}\sum_{\boldsymbol{y}_j\in\mathcal{Y}_i}\left(\boldsymbol{y}_j-\tilde{\boldsymbol{m}}'_i\right)\left(\boldsymbol{y}_j-\tilde{\boldsymbol{m}}'_i\right)^T = A^TP^TS'_iPA,$$

$$\widetilde{S}'_W = \sum_i^c \widetilde{S}'_i = A^TP^TS'_WPA,$$

$$\widetilde{S}'_B = \sum_i^c \left(\tilde{\boldsymbol{m}}'_i-\tilde{\boldsymbol{m}}'\right)\left(\tilde{\boldsymbol{m}}'_i-\tilde{\boldsymbol{m}}'\right)^T = A^TP^TS'_BPA,$$

where $\boldsymbol{y}_j = A^TP^T\boldsymbol{x}_j, \tilde{\boldsymbol{m}}'_i = A^TP^T\boldsymbol{m}'_i, \tilde{\boldsymbol{m}}' = A^TP^T\boldsymbol{m}'$.

Then the eigen-problem $\left(P^TS'_WP\right)^{-1}P^TS'_BP$ is solved, similar to normal LDA.

It should be noted that we don't need to increase training samples to calculate S'_W and S'_B because S_W, S_B, R_W can be computed from the given set of samples, and $\{G_j\}$ are given in advance. Therefore, increasing samples by image processing can be replaced with computing $\bar{G}$ and autocorrelation matrices, which requires less computational cost but more storage memory. In the next section we give a method to reduce the memory cost . However, the most important point is that an equivalent process can be reproduced by using generating matrices without actually increasing the samples.

3. Analysis of image processing operators

In this section, we analyze image processing operations as linear operators.

3.1. Hermite, unitary, and normal operators

Definition 1 *Let $f(\boldsymbol{x}), g(\boldsymbol{x}) \in L^2(R^2)$ be complex-valued 2D functions where $\boldsymbol{x} \in R^2$. The* inner product *is defined as*

$$(f,g) \equiv \int_{R^2} f(\boldsymbol{x})\overline{g(\boldsymbol{x})}d\boldsymbol{x}, \quad (16)$$

where $\bar{g}$ is the complex conjugate of g.

An operator $G : f \mapsto g$ is linear *if it satisfies $G(af + bg) = aG(f) + bG(g),\ \forall a, b \in R$.*

G^* is the adjoint operator *of G if it satisfies $(Gf, g) = (f, G^*g)$.*

We suppose that f, g are images and G is an operator which represents an image processing. A liner operator is the most interesting and important one because many image processing operations belongs to this type [1] as shown below.

At first, we choose a filtering (*e.g.* averaging, blur, or motion blur) which is represented by convolution with a filter kernel.

Proposition 1 *A filtering is defined as*

$$Gf(\boldsymbol{x}) = \int G(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y})d\boldsymbol{y}, \quad (17)$$

where the kernel is symmetric $G(\boldsymbol{x},\boldsymbol{y}) = G(\boldsymbol{y},\boldsymbol{x})$ and real valued. G is an Hermite *operator which satisfies $G^* = G$.*

Proof

$$(Gf,g) = \iint G(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y})\overline{g(\boldsymbol{x})}d\boldsymbol{y}d\boldsymbol{x}, \quad (18)$$

$$= \iint f(\boldsymbol{y})\overline{G(\boldsymbol{y},\boldsymbol{x})g(\boldsymbol{x})}d\boldsymbol{x}d\boldsymbol{y} = (f,Gg). \quad (19)$$

Therefore $G^* = G$. ∎

Many filters satisfy the symmetric assumption even if not being rotationally symmetric. In our experiments, motion blur is modeled by an anisotropic Gaussian-like kernel.

A second type of commonly used image processing we choose is the affine transform.

Proposition 2 *A geometric (affine) transformation G is defined as*

$$Gf(\boldsymbol{x}) = |A|^{1/2}f(A\boldsymbol{x}+\boldsymbol{t}), \quad (20)$$

where $|A| \neq 0$. G is a unitary *operator which satisfies $G^*G = I$.*

Proof

$$(Gf,g) = \int |A|^{1/2}f(A\boldsymbol{x}+\boldsymbol{t})\overline{g(\boldsymbol{x})}d\boldsymbol{x}, \quad (21)$$

$$= \int |A|^{1/2}f(\boldsymbol{y})\overline{g(A^{-1}(\boldsymbol{y}-\boldsymbol{t}))}|A|^{-1}d\boldsymbol{y}, \quad (22)$$

$$= (f,G^*g), \quad (23)$$

therefore the adjoint of G is the inverse transformation:

$$G^*f(\boldsymbol{x}) = |A|^{-1/2}f(A^{-1}(\boldsymbol{x}-\boldsymbol{t})), \quad (24)$$

hence $G^*Gf(\boldsymbol{x}) = f(\boldsymbol{x})$ and therefore $G^*G = I$. ∎

In addition to the affine transformation, many other geometric transformations can be represented as unitary operators if the transformation has an inverse.

Corollary 1 *Filtering or geometric transformation operators G are* normal operators *which satisfy $G^*G = GG^*$.*

[1] Operations which involve intensity change such as histogram equalization or gamma correction are not linear in this sense.

(a) $\boldsymbol{x}$ (b) $G\boldsymbol{x}$ (c) $G^TG\boldsymbol{x}$ (d) $G^T\boldsymbol{x}$

Figure 1. Examples of a generating matrix G for counter-clockwise rotation by 30 degrees. (a) An image $\boldsymbol{x}$ (size 64×64, from [14]). (b) $G\boldsymbol{x}$. (c) $G^TG\boldsymbol{x}$. (d) $G^T\boldsymbol{x}$. In (c), G^TG is not identical to the identity matrix due to aliasing and corners cropped as black regions, however most pixels in the center region remain by the successive transformations.

A normal operator can be orthogonally decomposed into eigenspaces as $G = \sum \lambda_i P_i$, where P_i is a projection operator onto an eigenspace corresponding to eigenvalue λ_i.

This is important to show because (1) many of the commonly used image processing operators are normal operators; (2) operators can be decomposed into eigenspaces and expressed by a linear combination rather than multiplication of operators; (3) eigenspaces with small eigenvalues can be eliminated for approximation in order to reduce the memory cost.

3.2. Naive implementation

To implement these operators, we use generating matrices. In our current naive implementation a generating matrix corresponding to Hermite or unitary operator is not guaranteed to be symmetric or orthogonal matrix. However, all generating matrices G of filters (*i.e.* Hermite) we have used for experiments are *almost* symmetric: $||G - G^T|| < 10^{-6}$. Generating matrices G of the affine transform do not satisfy $GG^T = I$, but results of $G^T\boldsymbol{x}$ are quite promising: if G represents a rotation, $G^T\boldsymbol{x}$ gives an inversely rotated image (shown in Fig.1). Therefore, we assume that the properties of the operators shown above are still valid for the matrix implementation.

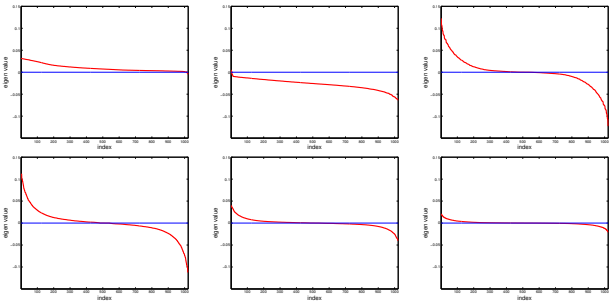
One problem when approximating a generating matrix by eigenspaces is that it has complex eigenvalues when G is almost orthogonal (in the sense mentioned above). Therefore, we use the decomposition of an operator G into two Hermite operators H_1, H_2 as follows:

$$G = H_1 + \mathrm{i}H_2,\ H_1 = \frac{G + G^T}{2},\ H_2 = \frac{G - G^T}{2\mathrm{i}}, \qquad (25)$$

where $\mathrm{i} = \sqrt{-1}$. The benefit of using this *Hermite decomposition* is that any generating matrix can be approximated by a pair of eigen-decompositions because both H_1 and H_2 are Hermite and all eigenvalues are real (not complex) numbers.

$\boldsymbol{x}$ $E_1\boldsymbol{x}$ $E_2\boldsymbol{x}$ $E_3\boldsymbol{x}$ $E_4\boldsymbol{x}$ $E_5\boldsymbol{x}$ $E_6\boldsymbol{x}$

Figure 2. Images obtained when the first six eigen-generating matrices E_i are applied to an image $\boldsymbol{x}$ (32×32, from [14]).



(a) H_{11}, H_{21} (b) H_{12}, H_{22} (c) H_{13}, H_{23}

Figure 3. Eigenvalues of H_{1j} (top row) and H_{2j} (bottom row) in descending order.

3.3. PCA of generating matrices

Combining PCA and Hermite decomposition, we propose a two–step approximation of the generating matrices. First, *eigen*-generating matrices E_j are obtained by PCA. Second, each of E_j is decomposed into two Hermite matrices H_{1j}, H_{2j}. Finally, H_{1j}, H_{2j} are approximated by their eigenspaces:

$$G \simeq \sum_j a_j E_j = \sum a_j (H_{1j} + \mathrm{i}H_{2j}), \qquad (26)$$

$$\simeq \sum_j a_j \sum_i (\lambda_{1ji} P_{1ji} + \mathrm{i}\lambda_{2ji} P_{2ji}), \qquad (27)$$

where P_{kji} is a projection matrix to the eigenspace of eigenvalue λ_{kji} of H_{kj}.

Fig. 2 shows results when applying E_i to $\boldsymbol{x}$, which shows how eigen-generating matrices work on images (matrices are difficult to visualize). E_i are obtained by PCA of 567 generating matrices (3 scaling, 7 rotations, 3 Gaussian filters and 9 motion blurs).

Fig. 3 shows the eigenvalues of two Hermite matrices for E_1, E_2, and E_3. There are few small eigenvalues of H_{1j}, however, many eigenvalues of H_{2j} are relatively small in magnitude and hence can be removed for approximation.

Fig. 4 illustrates eigenspaces of Hermite matrices (again, the figure shows how matrices work on an image). Each column shows the results of $P_{kji}\boldsymbol{x}$. It is worth noting that the sum of $P_{kji}\boldsymbol{x}$ provides $E_j\boldsymbol{x}$, a face, even though each of $P_{kji}\boldsymbol{x}$ looks like a geometric pattern (like Gabor or Fourier basis) rather than a face.

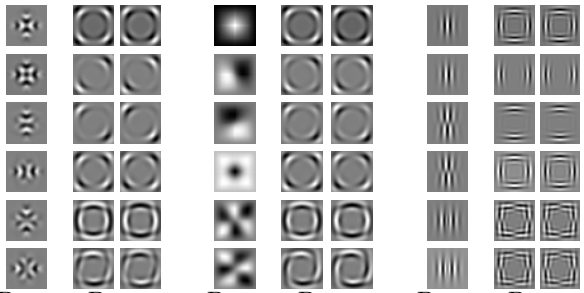


Figure 4. Images obtained when P_{kij} are applied (for $i = 1, \ldots, 6$ from top to bottom) to $\boldsymbol{x}$ (same $\boldsymbol{x}$ in Fig. 2). The two columns of $P_{2ji}\boldsymbol{x}$ correspond to the real (left) and imaginary (right) parts because P_{2ji} are unitary (complex) matrices, while P_{1ji} are orthogonal (real) matrices.

4. Discriminative generating matrix for LDA

In this section, we propose Linear Discriminative Image Processing Operator Analysis (LDIPOA), a method for combining generating matrices and LDA. In previous work, just the number of training samples is increased by a lot of image processing operations, however, those increased samples may not be useful for recognition. Our method allows to select *discriminative* image processing operations suitable for LDA from the given generating matrices.

4.1. Our approach

We estimate a single generating matrix $G^{(k)}$ at each step. To this end, the Rayleigh-Ritz variational technique is used: *i.e.*, $G^{(k)}$ is represented by the linear combination of J given generating matrices $\{G_j\}$, $G^{(k)} = \sum_j^J \alpha_j^{(k)} G_j$, where $(\alpha_1^{(k)}, \alpha_2^{(k)}, \ldots, \alpha_J^{(k)})^T = \boldsymbol{\alpha}^{(k)}$ are the coefficients to be estimated by maximizing the same criterion with LDA, the Rayleigh quotient E.

The proposed algorithm maximizes E in terms of A, P, and $\boldsymbol{\alpha}$: $\max_{A,P,\boldsymbol{\alpha}} E(A, P, \boldsymbol{\alpha})$, where A, P are the same as those defined in section 2. Since this problem is difficult to solve in a closed form, we propose a solution based on two maximization steps (Algorithm 1). In the following subsections, we give the details of the algorithm.

Algorithm 1 LDIPOA

1: Compute PCA P and LDA A. $G^0 \leftarrow I$.
2: **for** $k = 1, \ldots,$ **do**
3: **repeat**
4: $\boldsymbol{\alpha}$ step: $\boldsymbol{\alpha}^{(k)} = \mathrm{argmax}_{\boldsymbol{\alpha}} E(A, P, \boldsymbol{\alpha})$
5: PCA step: Compute P with $\boldsymbol{\alpha}^{(k)}$.
6: LDA step: $A = \mathrm{argmax}_A E(A, P, \boldsymbol{\alpha}^{(k)})$
7: **until** E converges
8: **end for**

4.2. α step

Suppose that a sample $\boldsymbol{x}$ in class ω_i is transformed to $\boldsymbol{x}^{(k)}$ by $G^{(k)}$ at the kth step as $\boldsymbol{x}^{(k)} = G^{(k)}\boldsymbol{x}$. Let $\bar{G}^{(k)} = \frac{1}{k+1}\sum_{l=0}^{k} G^{(l)}$. Now $D = PA$ are known, and the scatter matrices $\widetilde{S}_W^{\prime(k)}$, $\widetilde{S}_B^{\prime(k)}$ are given as follows:

$$\widetilde{S}_W^{\prime(k)} = D^T \bar{G}^{(k)} \left(S_W - R_W\right) \bar{G}^{(k)^T} D + \frac{1}{k+1} \sum_{l=0}^{k} D^T G^{(l)} R_W G^{(l)^T} D \quad (28)$$

$$\widetilde{S}_B^{\prime(k)} = D^T \bar{G}^{(k)} S_B \bar{G}^{(k)^T} D \quad (29)$$

Now the criterion $E(A, P, \boldsymbol{\alpha}^{(k)})$ can be rewritten [2] as the ratio of

$$\mathrm{tr}\left(\widetilde{S}_W^{\prime(k)}\right) = \boldsymbol{\alpha}^{(k)^T} H_W^{\prime(k)} \boldsymbol{\alpha}^{(k)} + 2\boldsymbol{q}_W^{(k)^T} \boldsymbol{\alpha}^{(k)} + \pi_W^{(k)}, \quad (30)$$

$$\mathrm{tr}\left(\widetilde{S}_B^{\prime(k)}\right) = \boldsymbol{\alpha}^{(k)^T} H_B^{\prime(k)} \boldsymbol{\alpha}^{(k)} + 2\boldsymbol{q}_B^{(k)^T} \boldsymbol{\alpha}^{(k)} + \pi_B^{(k)}, \quad (31)$$

where $H', \boldsymbol{q}, \pi$ are variables defined in the supplemental materials. Since this is not a usual form of the Rayleigh quotient, we prove the following proposition.

Proposition 3 *Let $\boldsymbol{\beta}^{(k)}$ be a $J+1$ dimensional vector $\boldsymbol{\beta}^{(k)} = (\boldsymbol{\alpha}^{(k)^T}, 1)^T$. Then, the solution that maximizes the ratio of the equations above is given by the solution to the eigenvalue problem for $Q_W^{(k)^{-1}} Q_B^{(k)}$, which maximize*

$$\frac{\boldsymbol{\beta}^{(k)^T} Q_B^{(k)} \boldsymbol{\beta}^{(k)}}{\boldsymbol{\beta}^{(k)^T} Q_W^{(k)} \boldsymbol{\beta}^{(k)}}, \quad (32)$$

where

$$Q_B^{(k)} = \begin{bmatrix} H_B^{(k)} & \boldsymbol{q}_B^{(k)} \\ \boldsymbol{q}_B^{(k)^T} & \pi_B^{(k)} \end{bmatrix}, \quad Q_W^{(k)} = \begin{bmatrix} H_W^{(k)} & \boldsymbol{q}_W^{(k)} \\ \boldsymbol{q}_W^{(k)^T} & \pi_W^{(k)} \end{bmatrix}. \quad (33)$$

Proof See the supplemental materials.

4.3. PCA step

The covariance matrix at the kth step is given as follows:

$$X^{\prime(k)} = \bar{G}^{(k)} \left(X - R_{\mathrm{all}}\right) \bar{G}^{(k)^T} + \frac{1}{k+1} \sum_{l=0}^{k} G^{(l)} R_{\mathrm{all}} G^{(l)^T}, \quad (34)$$

and then the eigenspace P of dimension $d' = (k+1)\,n - c$ can be obtained.

[2] Details of this derivations are given in the supplemental materials due to the page limitation.

4.4. LDA step

The scatter matrices are given by

$$S_W'^{(k)} = \bar{G}^{(k)} \left(S_W - R_W \right) \bar{G}^{(k)^T} + \frac{1}{k+1} \sum_{l=0}^{k} G^{(l)} R_W G^{(l)^T}, \tag{35}$$

$$S_B'^{(k)} = \bar{G}^{(k)} S_B \bar{G}^{(k)^T}. \tag{36}$$

Then the LDA feature space A is obtained by solving the eigenvalue problem for $\left(P^T S_W'^{(k)} P \right)^{-1} P^T S_B'^{(k)} P$.

5. Experimental results

Experimental results obtained by the proposed method are shown here. We compared recognition rates by using the ORL [15] and the FERET [16] datasets. The ORL dataset includes 10 images of 40 people. The first five images were used for training (200 images in total), and the other five images for testing (200 images in total). The FERET dataset used in this experiment includes the subset of the target set 'fa' and the query set 'fb'. There are 1002 people in 'fa' and 1001 in 'fb', and person has only one image (this setting is the same as in [17]). Therefore, this experiment can demonstrate the ability of the proposed method to handle the challenging problem called *learning from a single image per person*. For this case, we need to extend the proposed method because LDA can not be performed on a dataset with a single sample for each class. A simple strategy we employed here was to just blur each image and add the blurred images to the training set for computing and storing S_W, S_B, and R_W. Face regions were extracted and resized to 32×32 pixels. We prepared 567 generating matrices, from which eigen-generating matrices were obtained and used for computing $G^{(k)}$ by PCA that gives 80% and 95% cumulative contribution rates ("G-PCA" in the figures). For recognition, the nearest neighbor classifier was used.

First, by using ORL, we confirmed that the two steps can really maximize the Rayleigh quotient E. Figure 5 shows the values of E at each α and LDA step for estimating the first generating matrix $G^{(1)}$. We can see that it converges after only a few iterations. Figure 5 also shows the corresponding recognition rate.

Figure 6 (thick solid line shown as "PCA 95% (no blur)") shows recognition rates for ORL when increasing the number k of generating matrices from 0 to 10: $k = 0$ means that no generating matrices were used, but only $G^{(0)} = I$ was used, which is just normal LDA. $k = 1$ uses $G^{(0)}$ and $G^{(1)}$, and $k = 10$ uses $G^{(0)}$ and $G^{(1)} + \cdots + G^{(10)}$. The maximum recognition rate of 91.5% was achieved at $k = 5$, which is 1% higher than that of the normal LDA. Moreover, this result outperformed the case that used all original 567 generating matrices ("567 Gs" in Tab. 1) to increase the training samples to 113,400 images (2,835 for each class). This clearly demonstrates our concept: a discriminative set of image processing operations outperforms a non-discriminative one.

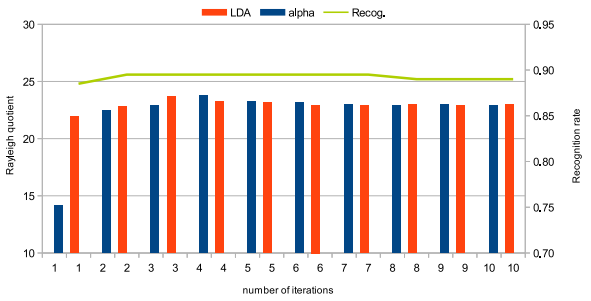


Figure 5. Rayleigh quotient (for α and LDA steps) and recognition rate over iterations for $G^{(1)}$.

Table 1. Recognition rates on the ORL dataset. Numbers in braces are corresponding values of k.

G-PCA	95	95	80	80
LDA-PCA	95	80	95	80
LDIPOA				
(no blur)	90.5 (0,5)	87.5 (10)	**93.0** (9)	87.0 (2)
(blur)	91.5 (5)	87.5 (9)	92.0 (10)	87.5 (2,5)
LDA				
(no blur)	90.5 (0)	87.5 (0)	90.5 (0)	87.0 (0)
(blur)	90.5 (0)	86.5 (0)	90.5 (0)	86.5 (0)
567 Gs (no blur)	78.0			

We also explored the effect of changing the cumulative contribution rate of PCA for recognition. Figure 6 shows the resulting performance when changing the value from 80% to 95% ("LDA-PCA" in the figures). This figure also compares the performance for the cases when blur (Gaussian $\sigma^2 = 0.5$) is either added or not added to the test images. The maximum recognition rate was 93% , which is much better than that of the normal LDA. Performances are summarized in Tab. 1.

Next, the proposed method was applied to the FERET dataset. Figure 7 shows the change in performance when increasing the number k of generating matrices $\{G^{(k)}\}$. In this case, the combination of cumulative contribution rates (G-PCA and LDA-PCA) of 95% perform much better. In particular, after $k = 5$ the recognition rate improves by about 10% compared to the normal LDA ($k = 0$), and the maximum performance achieves 82.72% ($k = 6$, with blur) and 82.62% ($k = 6$, no blur). Table 2 shows a comparison of performances with state–of–the–art results of single image per person which extend LDA and use a whole image as a feature. Since each paper uses different experimental settings, we compare how the performance can be improved from the baseline LDA. The results show that the highest relative improvement is achieved by our method.

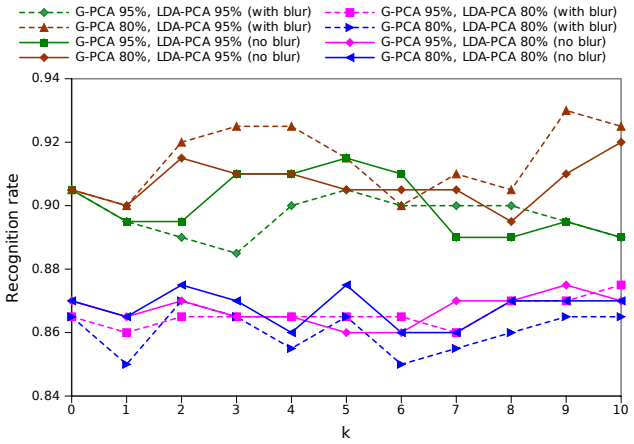


Figure 6. Performance change on ORL when increasing the number k of generating matrices. Note that $k = 0$ is the normal LDA.

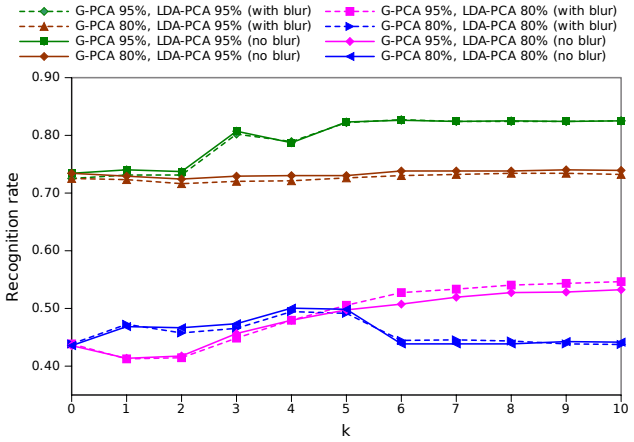


Figure 7. Performance change on the FERET dataset.

Table 2. Recognition rates on the FERET dataset.

Ref.	method	recognition rate	improvement from LDA
ours (no blur)	LDIPOA	82.6	**9.2**
	LDA	73.4	
ours (with blur)	LDIPOA	82.7	**10.2**
	LDA	72.5	
[18]	Adapted FLD	88.5	4.4
	LDA (Generic FLD)	84.1	
[19]	KNN ADA	90.1	6.0
	Lasso ADA	91.2	7.1
	LDA (Generic FLD)	84.1	

6. Conclusions

In this paper, we proposed a method for finding the most discriminative set of image processing operations to increase the number of training samples for LDA. By representing linear image processing operators as generating matrices, a two–step method was proposed that estimates LDA feature space and the most discriminative generating matrices at the same time. Experiments on the ORL and FERET datasets demonstrated that the proposed method can provide a set of discriminative generating matrices. Future work includes applying the generating matrix to other types of dimensionality reduction methods and classifiers such as the Mutual Subspace Method and Support Vector Machines.

Most derivations have been omitted due to the page limitation. Those will be given in the supplemental materials.

Acknowledgments

We thank to Jun Fujiki at AIST for the suggestion about the proposition 3, and Masashi Nishiyama at Toshiba for the discussion about FERET dataset.

References

[1] H. Ishida, T. Takahashi, I. Ide, Y. Mekada, H. Murase, “Identification of Degraded Traffic Sign Symbols by a Generative Learning Method”, ICPR2006, 1, 531–534, 2006.

[2] Q. Gao, L. Zhang, D. Zhang, “Face Recognition using FLDA with Single Training Image per Person”, Applied Mathematics and Computation, 205(2), 726–734, 2008.

[3] H. Yin, P. Fu, S. Meng, “Sampled FLDA for Face Recognition with Single Training Image per Person”, Neurocomputing, 69(16–18), pp. 2443–2445, 2006.

[4] X. Tan, S. Chen, Z. Zhou, and F. Zhang, “Face Recognition from a Single Image per Person: A Survey”, Pattern Recognition, 39(9), 1725–1745, 2006.

[5] S. Seitz, S. Baker, “Filter Flow”, ICCV2009, 143–150, 2009.

[6] T. Tamaki, T. Amano, K. Kaneda, “Representing images of a rotating object with cyclic permutation for view-based pose estimation”, Computer Vision and Image Understanding, 113(12), 1210–1221, 2009.

[7] P. Simard, B. Victorri, Y. Le Cun, J. Denker, “Tangent Prop – a Formalism for Specifying Selected Invariances in an Adaptive Network”, NIPS, 895–903, 1991.

[8] P. Y. Simard, D. Steinkraus, J. Platt, “Best Practice for Convolutional Neural Networks Applied to Visual Document Analysis”, ICDAR2003, 958–962, 2003.

[9] Y. Le Cun, L. Bottou, Y. Bengio, P. Haffner, “Gradient–based Learning Applied to Document Recognition”, Proc. of the IEEE, 86, 2278–2324, 1998.

[10] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, “Learning a Mahanalobis metric from equivalence constraints,” JMLR, 6, 937–965, 2005.

[11] M. Guillaumin, J. Verbeek, C. Schmid, “Is that you? Metric learning approaches for face identification,” CVPR2009, 498–505, 2009.

[12] S. X. Zhang, M. W. Mak, “Optimized Discriminative Kernel for SVM Scoring and Its Application to Speaker Verification,” Neural Networks, 22(2), 173–185, 2011.

[13] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection”, PAMI, 19(7), 711–720, 1997.

[14] A. Martinez, R. Benavente, “The AR Face Database”, CVC Technical Report, No. 24, 1998.

[15] The ORL face database, `http://www.cam-orl.co.uk/facedatabase.html`, AT&T Laboratories Cambridge.

[16] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, Patrick J. Rauss, “The FERET Evaluation Methodology for Face-Recognition Algorithms,” PAMI, 22(10), 1090-1104, 2000.

[17] Masashi Nishiyama, Abdenour Hadid, Hidenori Takeshima, Jamie Shotton, Tatsuo Kozakaya, Osamu Yamaguchi, “Facial Deblur Inference Using Subspace Analysis for Recognition of Blurred Faces,” PAMI, 33(4), 838-845, 2011.

[18] Meina Kan, Shiguang Shan, Yu Su, Xilin Chen, Wen Gao, “Adaptive Discriminant Analysis for face recognition from Single Sample per Person,” FG11, 2011.

[19] Yu Su, Shiguang Shan, Xilin Chen, Wen Gao, “Adaptive Generic Learning for Face Recognition from a Single Sample per Person,” CVPR2011, 2011.