IEICE TRANS. INF. & SYST., VOL.E96–D, NO.3 MARCH 2013

PAPER Special Section on Face Perception and Recognition

L1-Norm Based Linear Discriminant Analysis: An Application to Face Recognition

Wei ZHOU[†a)], *Nonmember and* **Sei-ichiro KAMATA**[†b)], *Member*

SUMMARY Linear Discriminant Analysis (LDA) is a well-known feature extraction method for supervised subspace learning in statistical pattern recognition. In this paper, a novel method of LDA based on a new L1-norm optimization technique and its variances are proposed. The conventional LDA, which is based on L2-norm, is sensitivity to the presence of outliers, since it used the L2-norm to measure the between-class and within-class distances. In addition, the conventional LDA often suffers from the so-called small sample size (3S) problem since the number of samples is always smaller than the dimension of the feature space in many applications, such as face recognition. Based on L1-norm, the proposed methods have several advantages, first they are robust to outliers because they utilize the L1-norm, which is less sensitive to outliers. Second, they have no 3S problem. Third, they are invariant to rotations as well. The proposed methods are capable of reducing the influence of outliers substantially, resulting in a robust classification. Performance assessment in face application shows that the proposed approaches are more effectiveness to address outliers issue than traditional ones.

key words: *linear discriminant analysis, L1-norm, linear programming, LDA-L1, 2DLDA-L1, BLDA-L1, face recognition*

1. Introduction

In many data analysis problems, measurements or observation data often lie in a lower dimensional subspace within the original high dimensional data space. Such a subspace, especially the linear subspace, has many important applications in computer vision and pattern recognition, such as motion estimation [1], face recognition [2]. Especially, in face recognition system, the feature vector of face is always located in high dimension and it takes times to classify the faces, so dimensionality reduction is an important step in face recognition task. Among these subspace methods, linear discriminant analysis (LDA) [3] is one of the most popular methods. LDA tries to find a set of projections that maximize the ratio of the between-class (S_w) distance to the within-class (S_w) distance. These projections constitute a low-dimensional linear subspace by which the data structure, such as face features, in the original input space can be effectively captured.

The classical LDA [3], [4] tries to find an optimal discriminant subspace to maximize the S_b separability and the S_w compactness of the data samples in a low-dimensional vector space. However, in many cases, the classical LDA suffers from the so called Small Sample Size (3S) problem, since it needs one of the scatter matrices (S_w) is nonsingular to calculate S_w^{-1}. Unfortunately, the size of the training set is much smaller than the dimension of the feature space in many applications, such as face recognition. In recent years, some LDA extensions have been proposed to deal with such 3S problem. The most famous one is called Fisherface [5], which first applied PCA to reduce the original data and then used LDA to extract the discriminant information. However, this approach may lose important discriminant information in the PCA stage for further face classification process. In [6], null-space linear discriminant analysis (NLDA) was proposed, which projected all the samples onto the null space of S_w and then extracted discriminant information. In [7], direct linear discriminant analysis (DLDA) extracted the discriminant information from the null space of S_w matrix, achieved by diagonalizing first S_b then diagonalizing S_w. A common drawback of the above methods is that they solve the discriminant vectors by focusing on a single data subspace rather than the full data space. Therefore, these methods may lose some useful discriminant information [8], [9] to some extent. Thus, in [8], the researchers proposed a dual-space LDA (DSLDA) approach to take full advantage of the discriminant information of the training samples. The basic idea of the DSLDA method is to divide the whole data space into two complementary subspaces, i.e., the range space of the within-class scatter matrix and its complementary space, and then solve the discriminant vectors in each subspace. On the other hand, from the computational point of view, DSLDA method may not be suitable for online training problems because of its heavy computational cost. In [9], the authors proposed a complete kernel Fisher discriminant analysis (CKFD) algorithm, which can be used to carry out discriminant analysis in both scatters.

However, Frobenius norm (L2-Norm), which is sensitivity to the presence of outliers or noise, is used in all the above LDA approaches to measure S_b and S_w distances. Thus, the process of training may be dominated by outliers since the S_b or S_w distance is determined by the sum of squared distances. Inspired by [10], [11] to reduce the influence of outliers, we propose a novel L1-norm based linear discriminant analysis called LDA-L1 for robust discriminant analysis (and also we noted the traditional LDA based on L2 norm as LDA-L2 in the following). Let $Z = (z_1, z_2, \ldots, z_n)$ be n data points in d-dimensional space. In matrix form $Z = (z_{ji})$, index j sum over spatial dimensions, $j = 1, 2, \ldots, d$ and index i sum over data points, $i = 1, 2, \ldots, n$. L2 norm is

Manuscript received June 8, 2012.
Manuscript revised October 27, 2012.

[†]The authors are with the Graduate School of Information, Production and Systems, Waseda University, Kitakyushu-shi, 808–0135 Japan.
a) E-mail: wei.zhou@fuji.waseda.jp
b) E-mail: kam@waseda.jp

DOI: 10.1587/transinf.E96.D.550

Copyright © 2013 The Institute of Electronics, Information and Communication Engineers

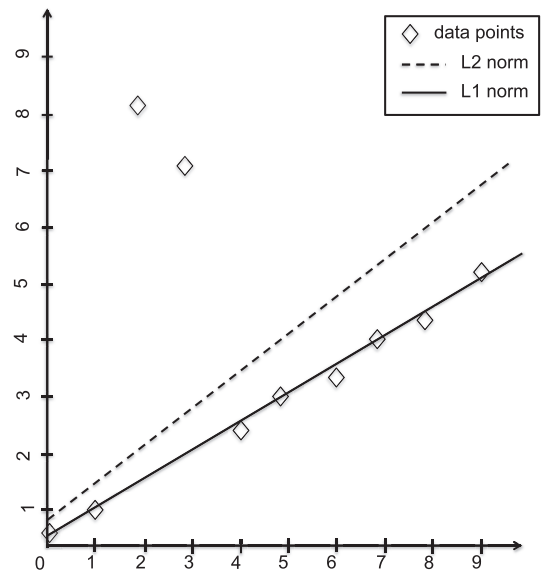


Fig. 1 Fit a line to 10 given data points. The two data points on upper-left are outliers.

defined as

$$\|Z\|_{L2} = \left(\sum_{i=1}^{n} \sum_{j=1}^{d} z_{ji}^2 \right)^{\frac{1}{2}}, \tag{1}$$

and L1 norm is defined as

$$\|Z\|_{L1} = \sum_{i=1}^{n} \sum_{j=1}^{d} |z_{ji}|. \tag{2}$$

It is well known that L1-norm is much more robust to outliers than L2-norm. Figure 1 shows a simple example of computing the 1D subspace (the straight line) from ten 2D input data points, two of which are outliers. While L2-norm gives erroneous line fitting, L1-norm gives correct result.

Recently, in order to solve the outlier problem, Li [12] proposed rotation invariant L1-norm (notated as R1-norm) based linear discriminant analysis (we call it LDA-R1 in the following). The R1-norm is defined as

$$\|Z\|_{R1} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} z_{ji}^2 \right)^{\frac{1}{2}}. \tag{3}$$

Here, R1-norm is determined by the sum of elements without being squared. Thus, the R1 norm is less sensitive to outliers than L2-norm. However, LDA-R1 takes a lot of time to achieve convergence for a large dimensional input space. In this paper, instead of maximizing variance which is based on L2-norm, L1-norm based linear discriminant analysis is proposed. Based on the reports in [10], the proposed method is expected less sensitive to outliers than L2-norm and R1-norm based approaches, and it also does not have 3S issue since it do not need to calculate the inverse of scatter matrix S_w^{-1}. In addition, the proposed method is simple and easy to implement.

The remainder of this paper is organized as follows: problem formulation will be described in Sect. 2. In Sect. 3, the solution of the proposed method will be introduced. Some variances will be illustrated in Sect. 4 and Sect. 5. Experimental results are presented in Sect. 6. Finally, conclusions and future work are discussed in Sect. 7.

2. Problem Formulation

Assume we have a set of samples $X = \{\{x_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{d \times n}$. N_l of which belong to class ω_l ($l = 1, 2, \ldots, C$), where n and d denote the number of samples and the dimension of the original input space, respectively, and $n = \sum_{l=1}^{C} N_l$.

In LDA-L2, the objective is to seek t projections $Y = \{\{y_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{t \times n}$ by means of t linear transformation vectors $w_i \in \mathbb{R}^{d \times 1}$, which can be arranged by columns into a projection matrix $W = [w_1, w_2, \ldots, w_t] \in \mathbb{R}^{d \times t}$, which embeds the original d dimension into t dimension vector space such that $t < d$. Let S_b be the between-class scatter matrix, and S_w be the within-class scatter matrix. Thus, the between-class and within-class distances can be, respectively, formulated as:

$$S_b = \sum_{l=1}^{C} (m_l - m)(m_l - m)^T, \tag{4}$$

$$S_w = \sum_{l=1}^{C} \sum_{i=1}^{N_l} (x_i^l - m_l)(x_i^l - m_l)^T, \tag{5}$$

where $m_l = (1/N_l) \sum_{i=1}^{N_l} x_i^l$ is the mean of the samples belonging to the l-th class, and $m = (1/n) \sum_{l=1}^{C} N_l m_l$ is the global mean of the samples. LDA-L2 aims to find an optimal transformation W by maximizing the ratio of $\mathrm{Tr}(S_b)$ and $\mathrm{Tr}(S_w)$ as the following problem

$$\max_W J_{L2} = \max_W \frac{\mathrm{Tr}(S_b)}{\mathrm{Tr}(S_w)} = \frac{W^T S_b W}{W^T S_w W}. \tag{6}$$

It is known that the L2-norm is sensitive to outliers and R1-norm approach was presented to solve this problem [12]. In this case, the problem becomes finding W that maximizes the following objective function:

$$\begin{aligned} \max_W J_{R_1} = (1-\alpha) \sum_{l=1}^{C} N_l \sqrt{\|W^T(m_l - m)\|^2} - \\ \alpha \sum_{l=1}^{C} \sum_{i=1}^{N_l} \sqrt{\|W^T(x_i^l - m_l)\|^2}. \end{aligned} \tag{7}$$

Here, the parameter α is a trade-off predefined coefficient such that $0 < \alpha < 1$. However, for a large dimensional input space, it takes a lot of time to achieve convergence, and also it has null space problem, which appears very often in face recognition application. In this paper, we want to maximize the L1 dispersion using the L1-norm in the feature space as the following

$$\begin{aligned} \max_W J_{L_1} &= \max_W \frac{\sum_{l=1}^{C} N_l \|W^T(m_l - m)\|_{L1}}{\sum_{l=1}^{C} \sum_{i=1}^{N_l} \|W^T(x_i^l - m_l)\|_{L1}} \\ &= \max_W \frac{\sum_{l=1}^{C} N_l |W^T(m_l - m)|}{\sum_{l=1}^{C} \sum_{i=1}^{N_l} |W^T(x_i^l - m_l)|}. \end{aligned} \tag{8}$$

The solution of Eq. (8) is invariant to rotations because the maximization is done on the feature space and it is expected to be more robust to outliers than the L2-Norm and R1-Norm solution. Moreover, no Small Sample Size problem will be occurred as described above.

3. L1 Norm Based Linear Discriminant Analysis

Generally, classical LDA used L2 norm (Frobenius norm) to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. However, L2 norm is sensitive to outliers and is not satisfied for robust discriminant analysis. In order to address this key issue, L1 norm is applied into the objective function and we call this novel approach as L1 norm based Linear Discriminant Analysis (LDA-L1).

Suppose that $W = [w_1, w_2, \ldots, w_t] \in \mathbb{R}^{d\times t}$ is the orthogonal projection matrix which want to be obtained. Here, we use greedy iteration algorithm to find $w_k (k = 1, 2, \ldots, t)$ one by one. $w_k(r)$ stands for the result w_k of the iteration r. In order to remove the absolute value operators in Eq. (8), polarity functions $p_l(r)$ and $q_l(r)$ are introduced as

$$p_l(r) = \begin{cases} 1 & \text{if} \quad w_k^T(r)(m_l - m) > 0 \\ -1 & \text{if} \quad w_k^T(r)(m_l - m) \le 0, \end{cases} \tag{9}$$

$$q_l(r) = \begin{cases} 1 & \text{if} \quad w_k^T(r)(x_i^l - m_l) > 0 \\ -1 & \text{if} \quad w_k^T(r)(x_i^l - m_l) \le 0, \end{cases} \tag{10}$$

with the help of $p_l(r)$ and $q_l(r)$, for a special w_k, Eq. (8) can be rewritten as

$$\max_{w_k} J_{L_1} = \max_{w_k} \frac{\sum_{l=1}^{C} N_l p_l w_k^T (m_l - m)}{\sum_{l=1}^{C}\sum_{i=1}^{N_l} q_l w_k^T (x_i^l - m_l)}. \tag{11}$$

An important property to notice about the objective Eq. (11) is that it is invariant with respect to rescaling of the vectors $w_k \to \rho w_k$. Hence, we can always choose w_k such that the denominator is simply $\sum_{l=1}^{C}\sum_{i=1}^{N_l} q_l w_k^T (x_i^l - m_l) = 1$, since it is a scalar itself. For this reason we can transform the problem of objective Eq. (11) into the following constrained optimization problem:

$$\begin{aligned} &\max_{w_k} J_{L_1}^* = \max_{w_k} \textstyle\sum_{l=1}^{C} N_l p_l w_k^T (m_l - m), \\ &s.t. \quad \textstyle\sum_{l=1}^{C}\sum_{i=1}^{N_l} q_l w_k^T (x_i^l - m_l) = 1, \\ &\quad\; w_k^T w_k = 1. \end{aligned} \tag{12}$$

Let

$$f(w_k(r)) = \textstyle\sum_{l=1}^{C} N_l p_l(r) w_k(r)^T (m_l - m). \tag{13}$$

First, we want to prove if $w_k(r+1) = arg\max_{w_k} f(w_k)$ then $f(w_k(r+1)) \ge f(w_k(r))$. According to Eq. (9), we can get $p_l(r+1)w_k^T(r+1)(m_l - m) \ge p_l(r)w_k^T(r+1)(m_l - m)$, then

$$\begin{aligned} f(w_k(r+1)) &= \textstyle\sum_{l=1}^{C} N_l p_l(r+1) w_k^T(r+1)(m_l - m) \\ &\ge \textstyle\sum_{l=1}^{C} N_l p_l(r) w_k^T(r+1)(m_l - m), \end{aligned} \tag{14}$$

since $w_k(r+1) = arg\max_{w_k} f(w_k)$, then

$$\begin{aligned} f(w_k(r+1)) &\ge \textstyle\sum_{l=1}^{C} N_l p_l(r) w_k^T(r)(m_l - m) \\ &= f(w_k(r)). \end{aligned} \tag{15}$$

Thus, what we want to do in the next step is to find $w_k(r+1) = arg\max_{w_k} f(w_k)$ where $w_k(r+1)$ stands for the result w_k of the iteration $r+1$. Inspired by the solution of LDA-L2, this problem can be converted into

$$\begin{aligned} &\max_{w_k} f(w_k) = \textstyle\sum_{l=1}^{C} N_l p_l w_k^T (m_l - m), \\ &s.t. \quad \textstyle\sum_{l=1}^{C}\sum_{i=1}^{N_l} q_l w_k^T (x_i^l - m_l) = 1, \\ &\quad\; w_k^T w_k = 1. \end{aligned} \tag{16}$$

Denote $\theta = \sum_{l=1}^{C} N_l p_l (m_l - m)$ and $\phi = \sum_{l=1}^{C}\sum_{i=1}^{N_l} q_l (x_i^l - m_l)$. Consequently, based on augmented Lagrangian method, we can have the following optimization problem:

$$\max_{w_k} \mathcal{L} = w_k^T\theta - \lambda(w_k^T\phi - 1) - \tfrac{1}{2}\beta(w_k^T w_k - 1). \tag{17}$$

According to the Karush-Kuhn-Tucker (KKT) conditions for the optimal solution, we can get

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_k} = \theta - \lambda\phi - \beta w_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = w_k^T\phi - 1 = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta} = w_k^T w_k - 1 = 0. \end{cases} \tag{18}$$

Then based on Eq. (18), we can obtain the solution of λ, β and w_k as follows:

$$\begin{cases} \lambda = \frac{\theta^T\phi}{\phi^T\phi} \\ \beta = (\theta^T - \lambda\phi^T)\phi \\ w_k = \frac{1}{\beta}(\theta - \lambda\phi). \end{cases} \tag{19}$$

Since $p_l(r) \le 1$ and each element of $w_k(r)$ is less equal to 1, then

$$\begin{aligned} f(w_k(r)) &= \textstyle\sum_{l=1}^{C} N_l p_l(r) w_k(r)^T (m_l - m) \\ &\le \textstyle\sum_{l=1}^{C} N_l \|m_l - m\|_{L_1}. \end{aligned} \tag{20}$$

Thus, apparently, $f(.)$ function has an upper bound and increases monotonically. Then, by iteration we can get the optimal w_k in Eq. (11). So far, the previous equations can be used to compute the principal vector w_k of $W \in \mathbb{R}^{d\times t}$. And in order to calculate the following vector $w_{k+1} (k = 1, 2, \ldots, t-1)$, x_i^l should be updated as

$$(x_i^l)^{k+1} = (x_i^l)^k - w_k w_k^T (x_i^l)^k, \tag{21}$$

and then be followed by Eq. (11), for $k = 1$, we have $(x_i^l)^1 = x_i^l$. This kind of updating rule guarantees that w_{k+1} is orthogonal to w_k and then W is an orthogonal matrix. To justify $w_k^T w_{k+1} = 0$, first we prove that $w_k^T (x_i^l)^{k+1} = 0$ holds. Multiply w_k^T on both sides of Eq. (21), then

$$w_k^T (x_i^l)^{k+1} = w_k^T (x_i^l)^k - w_k^T w_k w_k^T (x_i^l)^k, \tag{22}$$

since $w_k^T w_k = 1$, then we can have

$$w_k^T (x_i^l)^{k+1} = w_k^T (x_i^l)^k - w_k^T (x_i^l)^k = 0. \tag{23}$$

On the other side, w_{k+1} can be represented by $(x_i^l)^{k+1}$,

Algorithm 1 LDA-L1

Require: $X = \{\{x_i^l\}_{i=1}^{N_l}\}_{l=1}^{C} \in \mathbb{R}^{d\times n}, t \le d$
Initialization: $W = [w_1, w_2, \ldots, w_t] \in \mathbb{R}^{d\times t}, W^T W = I, X_1 = X$
for $k = 1 \to t$ **do**
$w_k(1) = w_k, r = 1$
while not converge **do**
1. Calculate $p_l(r)$ and $q_l(r)$ according to Eq. (9) and Eq. (10)
2. Update w_k according to Eq. (19)
3. r++
end while
Update $X_{k+1} = X_k - w_k w_k^T X_k$
end for
return $W \in \mathbb{R}^{d\times t}$

such as $w_{k+1} = \sum_{l=1}^{C}\sum_{i=1}^{N_l}\gamma_{il}(x_i^l)^{k+1}$, where γ_{il} are coefficients. Then we can have $w_k^T w_{k+1} = 0$.

Finally, we can get the algorithm of Eq. (8) as Algorithm 1.

Note that this procedure tries to find a local maximum solution and there is a possibility that it may not be the global solution. However, considering that the initial vector w_1 can be set arbitrarily, by setting w_1 appropriately, we can run the LDA-L1 procedure several times with different initial vectors and output the projection vector that gives the maximum L1 dispersion.

4. L1 Norm Based Two Dimensional Linear Discriminant Analysis

L1 Norm Based Two Dimensional Linear Discriminant Analysis (2DLDA-L1) aims to find two optimal transformation matrices $W_1 = [w_1^1, w_2^1, \ldots, w_{t1}^1] \in \mathbb{R}^{w\times t_1}$ and $W_2 = [w_1^2, w_2^2, \ldots, w_{t2}^2] \in \mathbb{R}^{h\times t_2}$ to maximize the following objective function

$$\max_{W_1,W_2} J_{L_1}^{2D} = \max_{W_1,W_2} \frac{\sum_{l=1}^{C} N_l \|W_1^T(m_l^{2D} - m^{2D})W_2\|_{L1}}{\sum_{l=1}^{C}\sum_{i=1}^{N_l} \|W_1^T((x_i^l)^{2D} - m_l^{2D})W_2\|_{L1}} = \max_{W_1,W_2} \frac{\sum_{l=1}^{C} N_l |W_1^T(m_l^{2D} - m^{2D})W_2|}{\sum_{l=1}^{C}\sum_{i=1}^{N_l} |W_1^T((x_i^l)^{2D} - m_l^{2D})W_2|}. \quad (24)$$

In 2DLDA-L1, the sample is treated as a matrix instead of the vector in LDA-L1. $(x_i^l)^{2D} \in \mathbb{R}^{w\times h}$, where w and h denote the width and height of samples, respectively. $m_l^{2D} = (1/N_l)\sum_{i=1}^{N_l}(x_i^l)^{2D}$ is the mean of the samples belonging to the l-th class, and $m^{2D} = (1/n)\sum_{l=1}^{C} N_l m_l^{2D}$ is the global mean of the samples. Based on the solution of LDA-L1, projection matrix W_1 and W_2 can be solved one by one. More specifically, firstly, we can fix W_2 to solve W_1 and then fix W_1 to solve W_2.

5. L1 Norm Based Block Linear Discriminant Analysis

In this section, the L1-norm based optimization is applied to Block LDA, which is called L1 Norm Based Block Linear Discriminant Analysis (BLDA-L1). This procedure consists of dividing each sample into several small blocks to build several sub-datasets and combining them to optimize basis vectors with respect to the L1-dispersion. Instead of using row vectors as computational units, this approach is based on the observation that the pixels within a small block usually have strong correlation.

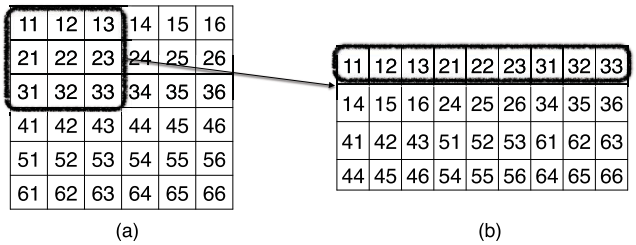


Fig. 2 Sample of BLDA-L1 (Block size is 3 by 3).

Figure 2 gives an example. Assume that we have original 6 by 6 image as Fig. 2 (a) (the number is denoted the location of each pixel), and the block size B is set 3 by 3. Then we can non-overlapping divide the original image into 4 blocks and convert it into 9 by 4 image as Fig. 2 (b). Each row in Fig. 2 (b) stands for one block in Fig. 2 (a). That is, BLDA-L1 can be converted into 2DLDA-L1, and then we can use the algorithm for 2DLDA-L1 to solve it.

Some points should be noted. If the block is selected as each row of the original image, then BLDA-L1 is same as 2DLDA-L1. If the whole image is treated as one block, then BLDA-L1 is same as LDA-L1. Thus, BLDA-L1 can be considered as a general framework of LDA-L1 and 2DLDA-L1.

6. Experiments

In this section, we will apply the proposed LDA-L1 algorithm and its variances to face recognition problems based on ORL [13], AR [14] and FERET [15] dataset. The performance is compared with those of LDA-L2 and LDA-R1. In our study, several initialization of w_1 is used: the first type is several arbitrary values; the second type is the solution of LDA-L2; the third type is a data vector in the training set, which has a maximum mean value. According to the following experiments, we can conclude that in general, the third type data is enough to get our solution converges to global optimization. And most times, the second type data and the third type data can get the same solution, while the first type data sometimes suffers the local optimization problem.

6.1 ORL Dataset

The first experiment over the ORL dataset [13] is to compare the classification performances and reconstruction errors of LDA-L2, LDA-R1 and the proposed methods. The ORL database consists of face images of 40 different people, each individual providing 10 different images. For some subjects, the images were taken at different times. The facial expressions open or closed eyes, smiling or non-smiling and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there is also some variation in the scale of up to about 10 percent. All images are

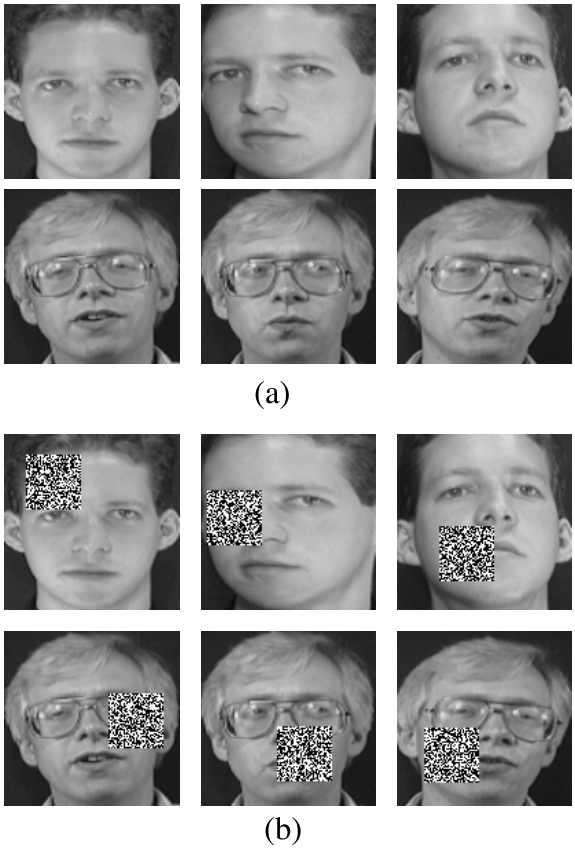


Fig. 3 ORL dataset (a) Original Images (b) Corresponding Images with occlusion.

gray scale and normalized to a resolution of 32 × 32 pixels. Among these 400 images, 30 percent were randomly selected and occluded with a rectangular noise consisting of random black and white dots whose size was 10 × 10, located at a random position. For a better illustration, some training samples are shown in Fig. 3. 3 images per person were used for training and others were for testing. Simple 1-nearest-neighbor(1NN) classifier was used for the final classification. The performance is shown in Fig. 4, where x-axis corresponds to the reduced dimension and y-axis is associated with the accuracy. The average number of iterations for LDA-L1 is 5.1 while 9.7 for LDA-R1. From this figure, we can see that the proposed method is the outstanding one and can obtain about 10 percent or 40 percent than LDA-R1 and LDA-L2, respectively. Moreover, in this figure, when the reduced dimension is very small, the proposed method can get significant performance. In order to see how the accuracy changes in small dimension, another experiment is carried out and the result is shown in Fig. 5. From this figure, we can see more clear about the effectiveness of the proposed method.

Figure 6 and Fig. 7 shows the accuracy with the change of selected feature dimension and block size by 2DLDA-L1 and BLDA-L1, respectively. From these two figures, we can see that the accuracy of two dimensional based LDA is higher than one dimension based LDA, and in Fig. 7, smaller

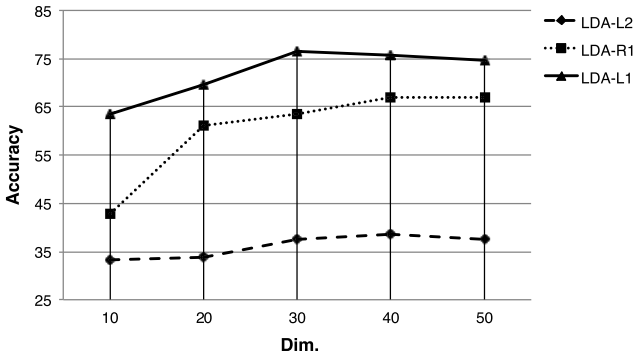


Fig. 4 Classification results on occluded ORL Dataset.

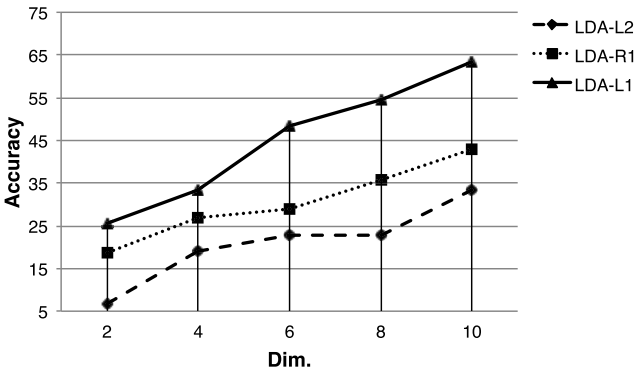


Fig. 5 Classification results for small dimension on occluded ORL Dataset.

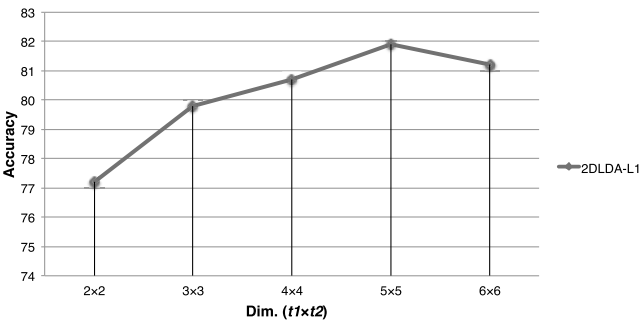


Fig. 6 Classification results on occluded ORL Dataset by 2DLDA-L1 method.

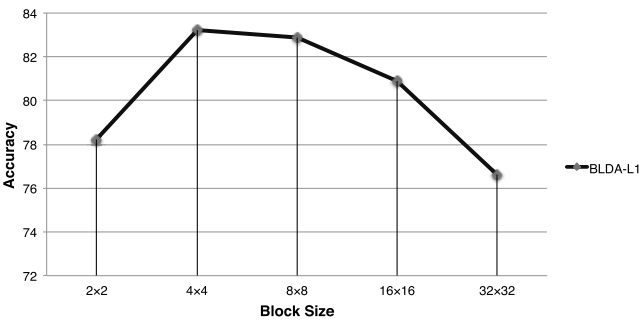


Fig. 7 Classification results with different block size by BLDA-L1 method on occluded ORL Dataset.

Table 1 Recognition rate on occluded ORL dataset.

method	Recognition Rate
LDA-L2	38.6
LDA-R1	67.1
2DPCA-L1 [11]	79.1
LDA-L1	76.6
2DLDA-L1	81.9
BLDA-L1	83.2

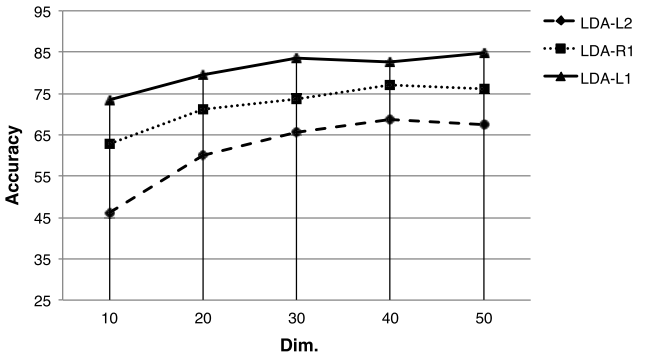


Fig. 8 Classification results on un-occluded ORL Dataset.

Table 2 Recognition rate on un-occluded ORL dataset.

method	Recognition Rate
LDA-L2	67.5
LDA-R1	77.1
2DPCA-L1 [11]	85.1
LDA-L1	84.7
2DLDA-L1	89.3
BLDA-L1	91.4

or larger block size will decrease the recognition rate. The reason may be that if the block size is small, then the correlation within each block is weak, and the correlation between blocks will be weak if large block size is applied. Finally, the accuracy on ORL dataset is concluded in Table 1, and our proposed methods has higher performance than traditional ones.

In order to show the effectiveness of the proposed methods in un-occluded case, several evaluations are carried out. Figure 8 illustrated the relationship between the accuracy and the reduced dimension. And Table 2 shows the comparison precision among several traditional approaches and the proposed methods. From these experiments, we can generate that the proposed methods are not only effective to deal with occlusion problem, but also powerful in un-occlusion case.

In next experiment, the proposed methods are applied to face reconstruction problems and the performances are compared with those of other methods. We applied LDA-L2, LDA-R1, LDA-L1, 2DLDA-L1 and BLDA-L1 and extracted various numbers of features. By using only a fraction of features, we could compute the average reconstruction error with respect to the original un-occluded images as Eq. (25) and Eq. (26) for one dimension based and two

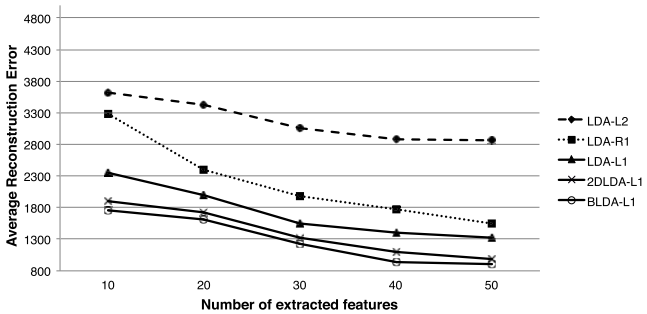


Fig. 9 Average reconstruction errors for ORL dataset.

dimension based LDA, respectively,

$$e_1(m) = \frac{1}{n}\sum_{i=1}^{n} \|x_i^{org} - \sum_{k=1}^{t} w_k w_k^T x_i\|_2, \tag{25}$$

$$e_2(m) = \frac{1}{n}\|(X^{2D})^{org} - W_1 W_1^T X^{org} W_2 W_2^T\|_2. \tag{26}$$

Here, n is the number of samples, which is 400 in this case, x_i^{org} and x_i are the i-th original un-occluded image and the i-th image used in the training, respectively, and t is the number of extracted features. $(X^{2D})^{org}$ and X^{org} are original un-occluded matrix based image and occluded matrix based image used in the training, respectively. Figure 9 shows the average reconstruction errors for various numbers of extracted features. In this figure, even when the number of extracted features is small, the average reconstruction error of the proposed method is much smaller than LDA-L2 and LDA-R1 approaches. The difference between the proposed method and traditional methods is apparent and BLDA-L1 is the most outstanding one.

6.2 AR Dataset

The AR [14] dataset consists of over 3,200 color images of the frontal images of faces of 126 subjects. There are 26 different images for each subject, including frontal views with different facial expressions, lighting conditions and occlusions. For each subject, these images were recorded in two different sessions separated by two weeks, each session consisting of 13 images. For the experiments reported in this section, 60 different individuals were randomly selected from this database. Then there are 1560 images in our experiments. All images were manually cropped and resized to 80 by 60. Some example images of one person are shown in Fig. 10.

In this experiment, we compare the recognition performances of the different algorithms on AR database. We randomly selected six samples of each individual for training, and the remaining ones were used for testing. We performed 10 times to randomly choose the training set and calculate the average recognition rates. Some classification results are listed in Fig. 11, where we can see that the proposed methods have higher performance than LDA-L2 and LDA-R1, especially for 2DLDA-L1 and BLDA-L1. (Note that in

Fig. 10 Some samples from AR dataset.

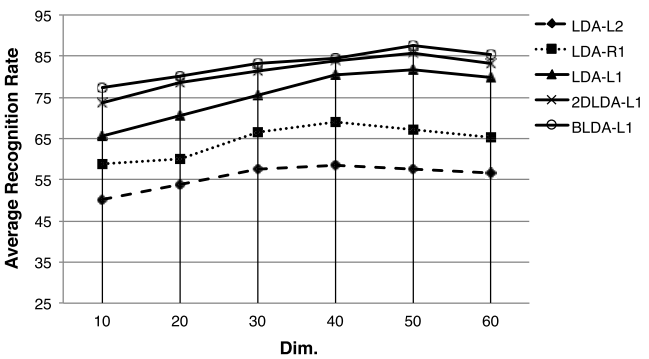


Fig. 11 Classification results on AR Dataset.

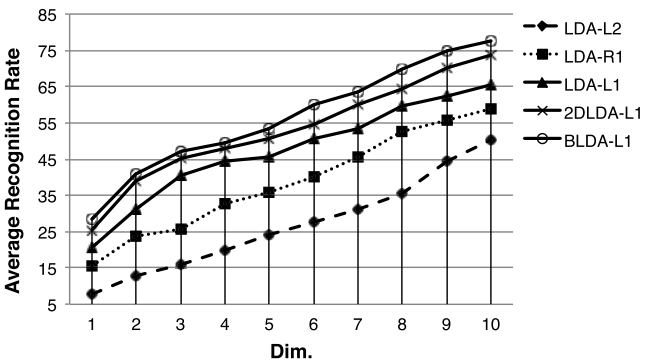


Fig. 12 Classification results for small dimension on AR Dataset.

2DLDA-L1 and BLDA-L1, the Dim. is equal to $t1 \times t2$) In general, LDA-L1 can obtain about 10 percent or 20 percent than LDA-R1 and LDA-L2, respectively. And the average number of iterations for LDA-L1 is 7.5 while 25.3 for LDA-R1. Thus, we can see clearly that LDA-R1 takes much more computation cost to achieve convergence in larger dimensional input space, such as face recognition application, than LDA-L1. Base on this evaluation, our proposed methods are more effective and efficient than the traditional approaches to solve facial expression, illumination or occlusions issues.

In Fig. 12, we focus on only low-dimensional spaces because we want to make a comparison of the most discriminant features for the proposed methods and some related algorithms. Same as Fig. 11, the proposed methods can extract more discriminant features.

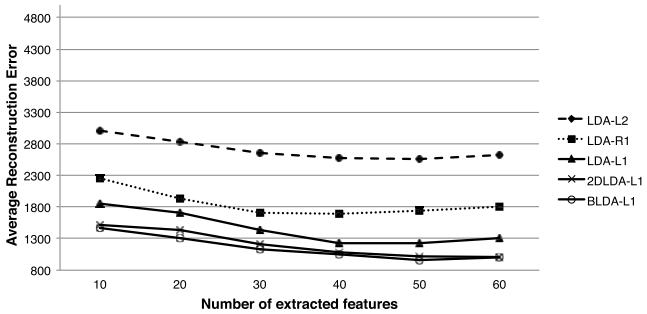


Fig. 13 Average reconstruction errors for AR dataset.

Table 3 Recognition rate on AR dataset.

method	Recognition Rate
LDA-L2	58.6
LDA-R1	69.1
2DPCA-L1 [11]	82.2
LDA-L1	81.6
2DLDA-L1	85.7
BLDA-L1	87.5

Fig. 14 Some samples from FERET dataset.

The average reconstruction errors for AR dataset are shown in Fig. 13 with various numbers of extracted features. From this figure, we can see that the average reconstruction error of the proposed method is much smaller than the traditional approaches while BLDA-L1 is a little better or comparable than 2DLDA-L1.

Finally, the accuracy on AR dataset is concluded in Table 3, and our proposed methods are superior to the traditional approaches.

6.3 FERET Dataset

The FERET dataset is a standard face image set to test and evaluate face recognition algorithms [15]. In this evaluation, we chose a subset of the FERET database. It includes 900 images of 150 individuals (each individual has six images). The six images of each individual consist of two or three front images with varied facial expressions and illuminations, and other images ranging from -15° to $+15^\circ$ pose. The facial portion of each original image was cropped to a size of 100 by 80 and no preprocessing method was applied. Figure 14 illustrates the six cropped images of one person. In our experiments, three images of each individual (450 images) are randomly selected for training. These training images were also used as gallery images. The remaining 450 images were used as probe images. The nearest neighbor classifier was used to match probe images and gallery images, and the average recognition rate was adopted by cal-

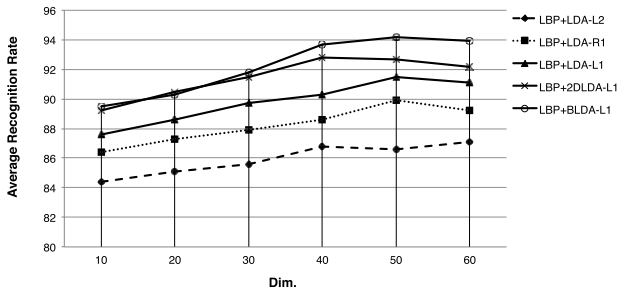


Fig. 15 Recognition Rate by combine LBP and some feature reduction methods from FERET dataset.

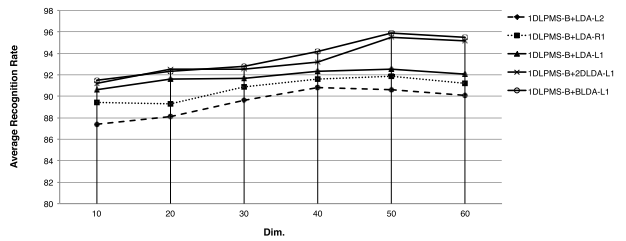


Fig. 16 Recognition Rate by combine 1DLPMS-B and some feature reduction methods from FERET dataset.

culating the mean value of recognition rates across 5 runs.

In this section, several experiments are conducted to judge whether the proposed methods are powerful or not when some other feature spaces are applied. Here, we selected two local patterns: uniform LBP [16] (the number of sampling points is set 8) and our previous proposed 1DLPMS-B [17] (Here, four kinds of scans are used and the number of sampling points for each scan is set 4). LBP first assigns a label to every pixel of an image by thresholding the 3 ×3-neighborhood of each pixel with the center pixel value and considering the result as a binary number. Second, all the binary numbers generate a histogram to be treated as the feature vector of this face. In 1DLPMS-B, multi-scans are used to capture the different spatial information on the facial image. Compared to LBP, which only uses a circle to encode the neighborhood pixels, multi-scans can keep more spatial information, reduce the effect of illumination and noise problem and 1DLPMS-B is rotation invariant. First, each facial image is equally divided into 80 blocks, and then the above mentioned local patterns is applied into each block. Thus, the total feature dimension of each facial image is 59×80 = 4720 and 64×80 = 5120, respectively and the average recognition rate by LBP is 85.6% while 88.5% for 1DLPMS-B. Figure 15 shows some classification results corresponding to LBP combined with some feature reduction methods and the results by combing 1DLPMS-B and related feature reduction methods are shown in Fig. 16.

From these two figures, we can see that the feature reduction methods are also effective and efficient in some other feature spaces, which are extracted by local patterns. Our proposed methods can improve accuracy by about 8 percent while 4 percent for LDA-R1 and 1 percent for LDA-L2. Note that the feature vector was firstly converted into two

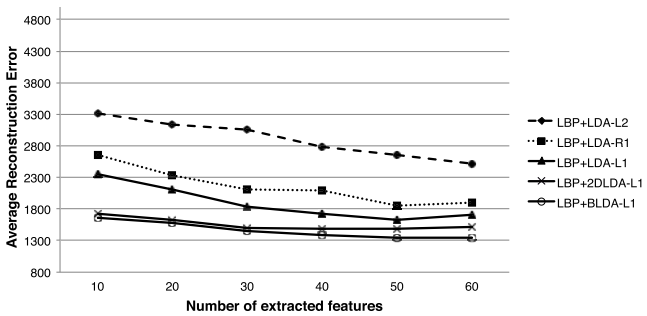


Fig. 17 Average reconstruction errors for FERET dataset by LBP.

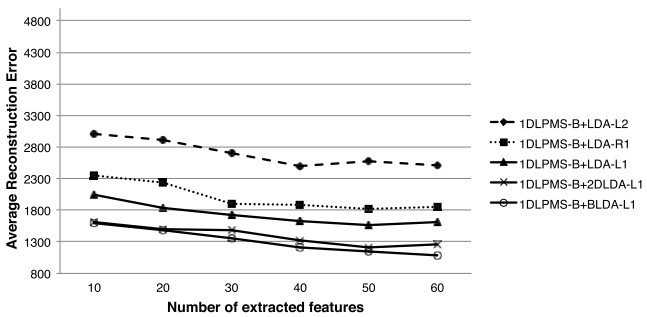


Fig. 18 Average reconstruction errors for FERET dataset by 1DLPMS-B.

dimensions when 2DLDA-L1 and BLDA-L1 was applied.

In the second estimation, the average reconstruction errors for FERET dataset with LBP and 1DLPMS-B are shown in Fig. 17 and Fig. 18 with various numbers of extracted features. From these figures, we can also see that the proposed methods are more suitable for reconstruction in local patterns based feature space, while BLDA-L1 performs best compared to other approaches.

7. Conclusions and Future Work

In this paper, we have proposed some methods of LDA based on L1-norm optimization, which better characterize the between-class separability and within-class compactness. The proposed methods are to find projections that maximize the L1-norm in the projected space instead of the conventional L2-norm. 2DLDA-L1 and BLDA-L1 are not only robust to outliers, but also treat an image as a matrix to make good use of the spatial structure information. The proposed L1-norm optimization technique avoids to compute the eigen-composition problem and is easy to implement. In more specification, it first suppresses the negative effects of outliers, second it has no 3S problem and third it is invariant to rotations. Experimental results have demonstrated the effectiveness of the proposed methods compared to the existing approaches.

In our future work, some more variances, such as tensor based and kernel based LDA and some more applications based on the proposed methods will be evaluated. For BLDA-L1, in principle, the blocks do not need to have same

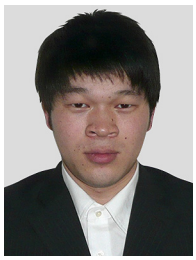
shape or cover the whole image, for example, in our face application, some key blocks around eyes, nose and mouth can be selected. How to choose these blocks is also our further research.

Acknowledgment

We would like to thank all the people providing their data for test and all the observers for giving their contributions to this study. This work was supported in part by a grant of Knowledge Cluster Initiative 2nd stage by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

[1] J. Landon, B. Jeffs, and K. Warnick, "Model-based subspace projection beamforming for deep interference nulling," IEEE Trans. Signal Process., vol.60, no.3, pp.1215–1228, March 2012.
[2] Y. Geng, C. Shan, and P. Hao, "Square loss based regularized lda for face recognition using image sets," CVPRW, pp.99–106, June 2009.
[3] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley-Interscience, 2005.
[4] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," 3rd International Conference on Automatic Face and Gesture Recognition, 1998.
[5] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.7, pp.711–720, July 1997.
[6] F. Chen, H.Y. 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 Recognit., vol.33, no.10, pp.1713–1726, 2000.
[7] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," Pattern Recognit., vol.34, no.12, pp.2067–2070, 2001.
[8] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," CVPR, vol.2, pp.564–569, 2004.
[9] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.2, pp.230–245, Feb. 2005.
[10] N. Kwak, "Principal component analysis based on l1-norm maximization," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.9, pp.1672–1680, Sept. 2008.
[11] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2DPCA," IEEE Trans. Syst. Man Cybern., B, Cybern., vol.40, no.4, pp.1170–1175, Aug. 2009.
[12] X. Li, W. Hu, H. Wang, and Z. Zhang, "Linear discriminant analysis using rotational invariant l1 norm," Neurocomputing, pp.2571–2579, 2010.
[13] Available at http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[14] A. Martånez and R. Benavente, "The ar-face database," CVC Technical Report 24, June 1998.
[15] P.J. Phillips, H. Moon, P.J. Rauss, and S. Rizvi, "The feret evaluation methodology for face recognition algorithms," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.10, pp.1090–1104, 2000.
[16] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.12, pp.2037–2041, Dec. 2006.
[17] W. Zhou, A. Ahrary, and S. Kamata, "Image description with local patterns: An application to face recognition," IEICE Trans. Inf. & Syst., vol.E95-D, no.5, pp.1494–1505, May 2012.



Wei Zhou received the B.E. degree in software engineering from Nanjing University, Nanjing, China, in July 2007, and the M.E. degree in information engineering from Waseda University, Japan in Sep. 2009. He is now pursuing the Ph.D. degree at the Graduate School of Information, Production and Systems, Waseda University, Fukuoka, Japan. His current research is focus on face understanding and object classification.

Sei-ichiro Kamata received the M.S. degree in computer science from Kyushu University, Fukuoka, Japan, in 1985, and the doctor of Engineering degree from the Department of Computer Science, Kyushu Institute of Technology, Kitakyushu, Japan, in 1995. From 1985 to 1988, he was in NEC, Ltd., Kawasaki, Japan. In 1988, he joined the faculty at Kyushu Institute of Technology. From 1996 to 2001, he was an Associate Professor in the Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University. Since 2003, he has been a professor in Graduate School of Information, Production and Systems, Waseda University. In 1990 and 1994, he was a Visiting Researcher at the University of Maine, Orono. His research interests include image processing, pattern recognition, image compression, remotely sensed image analysis, Image database and space-filling curve and fractals. Prof. Kamata is a member of the IEEE, and the ITE in Japan.