Generalized Mean for Feature Extraction in One-class Classification Problems

Jiyong Oh[a], Nojun Kwak[a], Minsik Lee[b], Chong-Ho Choi[b]

[a]*Graduate School of Convergence Science and Technology (e-mail:yong97@snu.ac.kr, nojunk@snu.ac.kr)*
[b]*School of Electrical Engineering and Computer Science, ASRI, Seoul National University (e-mail:cutybug@csl.snu.ac.kr, chchoi@snu.ac.kr)*

Abstract

Biased discriminant analysis (BDA), which extracts discriminative features for one-class classification problems, is sensitive to outliers in negative samples. This study focuses on the drawback of BDA attributed to the objective function based on the arithmetic mean in one-class classification problems, and proposes an objective function based on a generalized mean. A novel method is also presented to effectively maximize the objective function. The experimental results show that the proposed method provides better discriminative features than the BDA and its variants.

Keywords: Generalized Mean, Biased Discriminant Analysis, Feature Extraction, Dimensionality Reduction, One-class Classification.

1. Introduction

Dimensionality reduction, which reduces the number of input variables without degrading the performance of classification systems, is a major topic in pattern recognition and machine learning. A large number of methods have been proposed for reducing the input dimensionality, which can be divided into two categories, i.e., *feature selection* and *feature extraction*. The feature selection is to select a subset of m good features from a set of n ($n > m$) input variables, while the feature extraction is to produce m good features from input variables [1]. Two of the most popular feature extraction methods are principal component analysis (PCA) and linear discriminant analysis (LDA). It is well known that LDA is more suitable for classification problems than PCA [2], [3], [4]. LDA finds a subspace where the samples within the same class are placed close together, whereas the samples belonging to different classes are located far apart. It is assumed that the distributions of samples in each class are normal and homoscedastic, i.e., their covariance matrices are identical. Thus, it may be difficult to find a good feature space if this assumption is violated, which has been addresses by many researchers [5], [6], [7], [8]. Furthermore, LDA may fail not only in heteroscedastic cases and sometimes even in homoscedastic cases [9]. In [10] and [11], Tao et al. addressed the problem of conventional LDA by showing examples where classes located close to each other in the input space overlap in the feature space generated by LDA, and this problem is referred to as the *class separation problem*. They first showed that under certain assumptions LDA maximizes the arithmetic mean of the Kullback-Leibler divergences [12] between different classes, and then proposed to replace the arithmetic mean with the geometric mean. In [13], the harmonic mean was also used to replace the arithmetic mean instead of the geometric mean. Experiments showed that the methods based on the geometric or harmonic mean gave better per-

formance than the original LDA, especially when the number of extracted features was strictly smaller than the number of classes.

Nonetheless, LDA and its extensions generally fail to deliver good performance in one-class classification problems such as image retrieval [14] and face or eye detection [15], [16]. This is because they try to find a subspace where each of the positive and negative samples is well clustered. In order to overcome this limitation of LDA, Zhou and Huang [17] introduced *biased discriminant analysis* (BDA) by modifying LDA for one-class classification problems. Unlike LDA, BDA tries to cluster only the positive samples and place the negative samples far from the mean of the positive samples, as much as possible. Unfortunately, it was shown in [18] that this could degrade the discriminatory power of BDA because outliers in the negative samples might be exaggerated. It was noted that the negative effect of outliers was caused by the objective function of BDA, which was based on the squared Euclidean distances, and two alternatives were proposed, i.e., *SBDA* and *L1-BDA* [18]. SBDA reduces the negative effect due to the squared distances by putting a limit on the distances of negative samples from the mean of positive samples by a specific value, while L1-BDA reduces the negative effect by replacing the Euclidean distances with the l_1 distances. It was shown that SBDA and L1-BDA gave similar performance, which was better than the conventional BDA.

In this paper, we propose a novel method to reduce the effect of outlier. First, we show that conventional BDA gives a subspace that maximizes the arithmetic mean of the squared distances between negative samples and the mean of positive samples. To find a subspace where positive and negative samples are well separated from each other, it is reasonable to try to place the negative samples, which are close to the mean of the positive samples, far from the mean in the feature space. However, maximizing the arithmetic mean only maximizes the sum of the squared distances and it does not prevent some of negative samples from being projected

close to the mean of positive samples. Note that this characteristic of BDA is similar to the class separation problem in LDA. Inspired by the alternative solutions to the problem presented in [11] and [13], we incorporate the *generalized mean* into the objective function of BDA instead of the arithmetic mean. The generalized mean of a set of positive numbers can vary from the minimum to the maximum value of the set depending on its intrinsic parameter. We can obtain the arithmetic, the geometric, and the harmonic means by tuning the parameter. By maximizing the generalized mean with an appropriate value of the parameter, the proposed method can find a lower dimensional feature space where the negative samples close to the mean of the positive samples are located far away. Therefore, the proposed method can find more discriminative features than BDA, SBDA and L1-BDA. One of the difficulties in maximizing the objective function based on the generalized mean is that the objective function is non-convex and there is no closed form solution. Gradient-based methods can be applied to find a local maximum, but they are slow in convergence. To expedite the process of maximizing the objective function incorporating the generalized mean, we provide a novel method exploiting the fact that the generalized mean of positive numbers can be represented as a non-negative linear combination of the numbers.

There have been many researches on object detection problems, such as Viola-Jones object detector [19] and human detection based on histogram of oriented gradients [20]. These object detection problems can also be categorized as a one-class classification problem. However, in this paper, we focus on linear feature extraction methods for one-class classification problems.

The rest of this paper is organized as follows. The next section provides a brief overview of BDA, SBDA and L1-BDA. The generalized mean is also introduced in the section. In Subsection 3.1 we show that the second step of BDA is equivalent to maximizing the arithmetic mean of the squared distances of negative samples

from the mean of positive samples, and consequently the solution can be dominated by some outliers of the negative samples. The proposed method is described in Subsections 3.2 and 3.3. Section 4 demonstrates that the proposed method effectively overcomes the limitations of BDA and gives a better performance than BDA, SBDA, and L1-BDA. Finally, the last section concludes this paper.

2. Background

2.1. BDA, SBDA, and L1-LDA

LDA is a simple and powerful feature extraction method for classification problems [2], [3], but it is not suitable for one-class classification problems because it assumes that the samples belonging to the same class are clustered together. This assumption is not reasonable for negative samples in one-class classification problems where the distribution of negative samples has no regular form. BDA was proposed to overcome this limitation of LDA for one-class classification problems. Let us consider N_x positive samples denoted as $\mathbf{x} \in \Re^n$ and N_y negative samples denoted as $\mathbf{y} \in \Re^n$. BDA requires two scatter matrices as follows:

$$\begin{aligned} \mathbf{S}_y &= \frac{1}{N_y}\sum_{i=1}^{N_y}(\mathbf{y}_i - \mathbf{m}_x)(\mathbf{y}_i - \mathbf{m}_x)^T, \\ \mathbf{S}_x &= \frac{1}{N_x}\sum_{i=1}^{N_x}(\mathbf{x}_i - \mathbf{m}_x)(\mathbf{x}_i - \mathbf{m}_x)^T, \end{aligned} \tag{1}$$

where $\mathbf{m}_x = \frac{1}{N_x}\sum_{i=1}^{N_x}\mathbf{x}_i$. Using these matrices, BDA can be formulated as

$$\mathbf{W}_{BDA} = \arg\max_{\mathbf{W}} tr\left((\mathbf{W}^T\mathbf{S}_x\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_y\mathbf{W})\right), \tag{2}$$

where $tr(\cdot)$ is the matrix trace operator. The objective function is large when positive samples are aggregated around their mean and negative samples are located far away from the mean of the positive samples in the transformed space, so the

solution of the problem $\mathbf{W}_{BDA} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$ can be constructed using the generalized eigenvectors with respect to the largest m generalized eigenvalues that satisfy

$$\mathbf{S}_y \mathbf{w}_k = \lambda_k \mathbf{S}_x \mathbf{w}_k, \qquad k = 1, \ldots, m.$$

If $\mathbf{S}_x$ is non-singular, $\mathbf{W}_{BDA}$ is obtained by the simultaneous diagonalization [2] of $\mathbf{S}_x$ and $\mathbf{S}_y$ as

$$\begin{aligned} \mathbf{W}_{BDA} &= \mathbf{W}_1 \mathbf{W}_2, \\ \mathbf{W}_1^T \mathbf{S}_x \mathbf{W}_1 &= \mathbf{I}_n, \\ \mathbf{W}_2^T \widehat{\mathbf{S}}_y \mathbf{W}_2 &= \mathbf{\Lambda}, \\ \mathbf{W}_2^T \mathbf{W}_2 &= \mathbf{I}_m, \end{aligned} \tag{3}$$

where $\widehat{\mathbf{S}}_y = \mathbf{W}_1^T \mathbf{S}_y \mathbf{W}_1$, $\mathbf{\Lambda}$ is an $m \times m$ diagonal matrix, and $\mathbf{I}_n$ is an $n \times n$ identity matrix, respectively. Equation (3) shows that the simultaneous diagonalization consists of two steps. The first step is to whiten $\mathbf{S}_x$, i.e., it finds $\mathbf{W}_1$ such that $\mathbf{W}_1^T \mathbf{S}_x \mathbf{W}_1 = \mathbf{I}_n$, and the next is to perform the eigenvalue decomposition of $\widehat{\mathbf{S}}_y$.

If the number of positive samples is smaller than the dimensionality of the samples, i.e., $N_x < n$, the simultaneous diagonalization can not be performed because $\mathbf{S}_x$ is singular. This is the *small sample size* (SSS) problem. Many approaches have been presented to overcome the SSS problem [21], [22], [23], [24], but the regularization method [25] is simple and effective. It only requires $\mathbf{S}_x$ in (1) to be replaced by $\widehat{\mathbf{S}}_x = \mathbf{S}_x + \mu \mathbf{I}$ where μ is a small positive constant. It also provides a more robust solution when the number of samples is not sufficiently larger than the dimensionality of the samples. After obtaining $\mathbf{W}_{BDA}$, an arbitrary sample $\mathbf{z}$ can be transformed in the feature space as

$$\mathbf{z}_f = \mathbf{W}_{BDA}^T (\mathbf{z} - \mathbf{m}_x). \tag{4}$$

Using $\mathbf{z} - \mathbf{m}_x$ instead of $\mathbf{z}$, the mean of positive samples is mapped to the origin in the transformed space. The sample can be classified as positive if the distance between its projection and the mean of the positive samples, i.e., $\|\mathbf{z}_f\|_2$, is smaller than a predetermined threshold, otherwise it is classified as negative.

The conventional BDA tries to find a feature space where each of the negative samples is as far away from the mean of the positive samples as possible. However, the subspace produced by BDA tends to be sensitive to the outliers of negative samples. SBDA was proposed to alleviate the negative effect of the outliers in [18]. When $\mathbf{S}_y$ is calculated in SBDA, the distance of a negative sample from the mean of the positive samples is set to a positive number γ if the distance is greater than γ. This can be implemented by modifying $\mathbf{S}_y$ in (1) as

$$\begin{aligned} \mathbf{S}'_y &= \frac{1}{N_y}\sum_{i=1}^{N_y}(\mathbf{y}'_i - \mathbf{m}_x)(\mathbf{y}'_i - \mathbf{m}_x)^T, \\ \mathbf{y}'_i &= \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|_2}\min(\|\mathbf{y}_i\|_2, \gamma). \end{aligned} \tag{5}$$

where $\|\mathbf{x}\|_2$ is the l_2 norm of a vector $\mathbf{x}$. This is equivalent to constructing a hypersphere with a radius γ and projecting the negative samples outside the hypersphere onto the surface of the hypersphere. The transformation matrix $\mathbf{W}_{SBDA}$ can be obtained by the simultaneous diagonalization of $\mathbf{S}'_y$ and $\mathbf{S}_x$, and test samples can be classified using their projections transformed by $\mathbf{W}_{SBDA}$.

SBDA limits the distances of negative samples to γ to reduce the effects of the outliers, whereas L1-BDA utilizes the l_1 norm instead of the Euclidean norm in placing negative samples far from the mean of positive samples. After whitening $\mathbf{S}_x$ in (1), as in BDA and SBDA, its optimization problem is formulated as the

following:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{1}{N_y} \sum_{i=1}^{N_y} |\mathbf{w}^T(\mathbf{y}'_i - \mathbf{m}'_x)|,$$
$$\textit{s.t. } \mathbf{w}^T\mathbf{w} = 1.$$

Here, $\mathbf{y}'_i = \mathbf{W}_1^T\mathbf{y}_i$ and $\mathbf{m}'_x = \mathbf{W}_1^T\mathbf{m}_x$ such that $\mathbf{W}_1^T\mathbf{S}_x\mathbf{W}_1 = \mathbf{I}$. A novel method, which was originally presented as an alternative to PCA in [26], was developed to solve the above problem. Note that the solution $\mathbf{w}^*$ is a vector and not a matrix. Therefore, the projection vectors are obtained one by one under the constraint that they are mutually orthogonal. Further details can be found in [18].

2.2. Generalized Mean

For a non-zero p, the generalized mean or power mean [27] $\mathcal{M}_p$ of N positive numbers $a_1, a_2, \ldots, a_N$ is defined as

$$\mathcal{M}_p\{a_1, \ldots, a_N\} = \left(\frac{1}{N}\sum_{i=1}^{N} a_i^p\right)^{1/p}.$$

This equation is very similar to the l_p-norm of a vector $\mathbf{a} = [a_1, a_2, \ldots, a_N]^T$, which is a generalization of the length of the vector [28]. However, the generalized mean quite differs from the l_p-norm. Figure 1 shows that $\mathcal{M}_p\{1, 2, \ldots, 10\}$ monotonically increases as p changes from -10 to 10. For some positive numbers, the arithmetic mean, the geometric mean, and the harmonic mean are special cases of the generalized mean. Moreover, the maximum and minimum values of the numbers are also obtained from the generalized mean by making $p \to \infty$ and $p \to -\infty$, respectively. For example, the generalized mean of the two numbers a_1

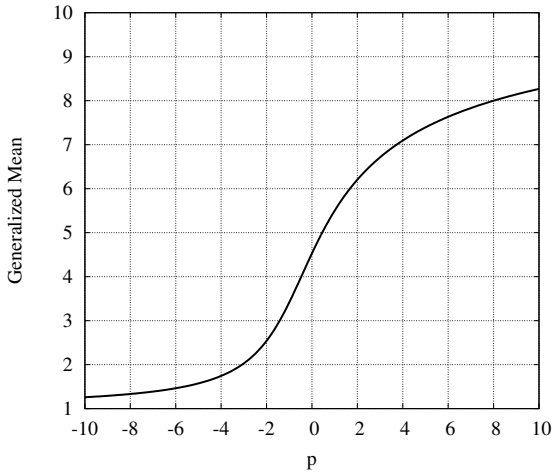
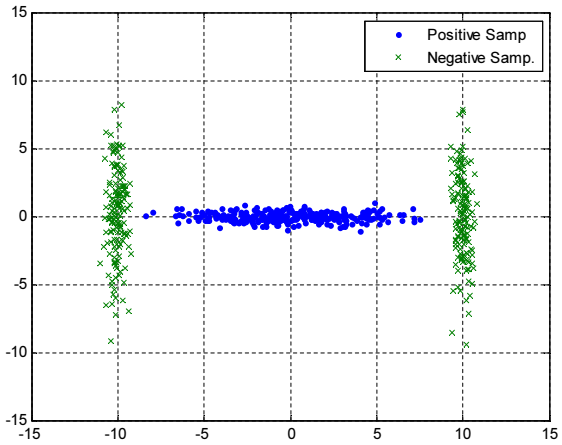


Figure 1: The generalized mean of $\{1, \ldots, 10\}$ for various values of p.

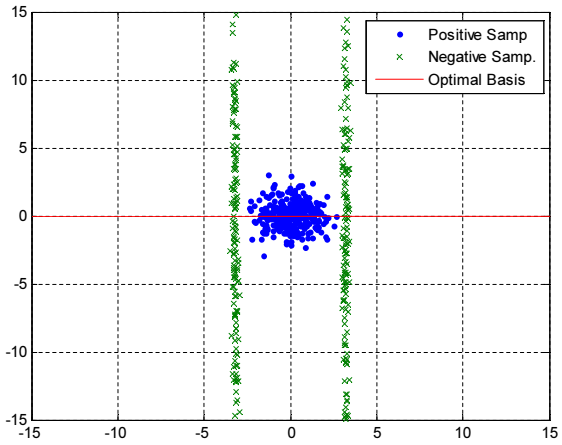
and a_2 $(0 < a_1 \leq a_2)$ can have various values for different p values as

$$\mathcal{M}_p\{a_1, a_2\} = \begin{cases} a_1 & (\min(a_1, a_2)) & \text{if } p \to -\infty, \\ \frac{2a_1a_2}{a_1+a_2} & (\text{harmonic mean}) & \text{if } p = -1, \\ \sqrt{a_1a_2} & (\text{geometric mean}) & \text{if } p \to 0, \\ \frac{a_1+a_2}{2} & (\text{arithmetic mean}) & \text{if } p = 1, \\ a_2 & (\max(a_1, a_2)) & \text{if } p \to \infty. \end{cases}$$

Note that as p decreases (increases), the generalized mean is more affected by the smaller (larger) number than the larger (smaller) number, i.e., controlling p makes it possible to adjust the contribution of each a_i to the generalized mean. In the next section, this property of the generalized mean will be utilized to extract more discriminative features for one-class classification problems.



(a)



(b)

Figure 2: Data distribution before and after whitening $\mathbf{S}_x$.

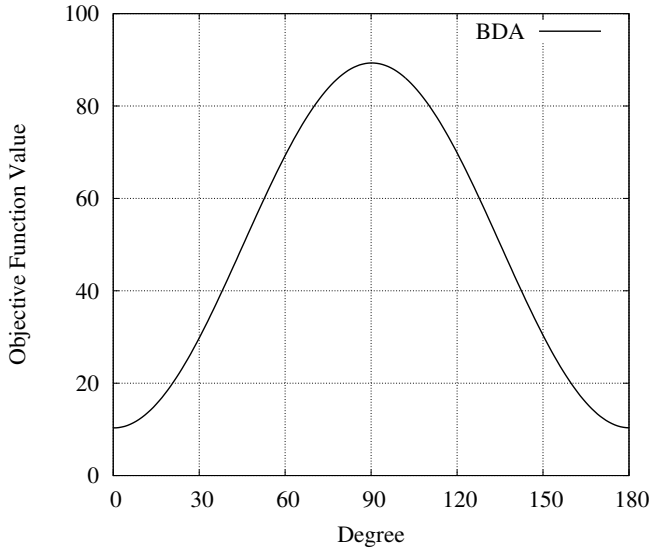
3. Biased Discriminant Analysis Using the Generalized Mean

3.1. A limitation of BDA

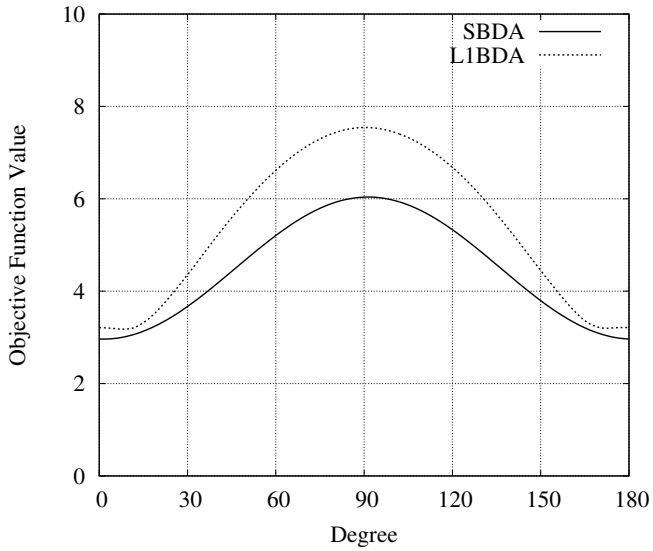
As mentioned above, the simultaneous diagonalization to solve BDA can be divided into two steps. Figure 2 shows a two-dimensional example of 300 positive samples and 300 negative samples before and after whitening $\mathbf{S}_x$, which is the first step of simultaneous diagonalization. Note that the positive samples in Fig. 2(b) are normally distributed. The next step is to find an orthogonal matrix $\mathbf{W}_2$ that places the negative samples far away from the mean of the positive samples in the transformed space, which is formulated as

$$\begin{aligned} \mathbf{W}_2 = &\arg\max_{\mathbf{W}} tr(\mathbf{W}^T \widehat{\mathbf{S}}_y \mathbf{W}), \\ &\textit{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}. \end{aligned} \tag{6}$$

where $\widehat{\mathbf{S}}_y = \mathbf{W}_1^T \mathbf{S}_y \mathbf{W}_1$. However, it may fail to find a subspace where positive and negative samples are well separated. For the negative samples in Fig. 2(b), Fig. 3(a) shows the value of the objective function in (6). In the figure, *Degree* is the angle by which the projection vector is counterclockwise rotated from the horizontal axis, $[1 \quad 0]^T$. Even though the optimal projection vector is 0 degrees ($[1 \quad 0]^T$), the objective function is maximized at 90 degrees, i.e., $\mathbf{W}_2 = [0 \quad 1]^T$ as shown in Fig. 3(a). This limitation of the conventional BDA is due to the fact that (6) maximizes the arithmetic mean of the squared Euclidean distances of negative samples from the mean of positive samples. Let us denote the negative sample after whitening $\mathbf{S}_x$ as $\mathbf{y}$. It is also assumed that the negative samples are transformed into a lower dimensional subspace by (4) so that the mean of positive samples is placed at the origin, i.e., $\mathbf{W}^T\mathbf{m}_x = 0$. Thus, the Euclidean distance between the transformed negative sample $\mathbf{W}^T\mathbf{y}_i$ and the mean of the positive samples is represented



(a)



(b)

Figure 3: Values of the objective functions for (a) BDA and (b) SBDA and L1-BDA.

as $\sqrt{(\mathbf{W}^T\mathbf{y}_i)^T(\mathbf{W}^T\mathbf{y}_i)}$ and the arithmetic mean of their squared values is

$$\begin{aligned}
&\frac{1}{N_y}\sum_{i=1}^{N_y}\mathbf{y}_i^T\mathbf{W}\mathbf{W}^T\mathbf{y}_i \\
&= \frac{1}{N_y}\sum_{i=1}^{N_y} tr(\mathbf{W}^T\mathbf{y}_i\mathbf{y}_i^T\mathbf{W}) \\
&= tr\left(\mathbf{W}^T\left[\frac{1}{N_y}\sum_{i=1}^{N_y}(\mathbf{y}_i-\mathbf{m}_x)(\mathbf{y}_i-\mathbf{m}_x)^T\right]\mathbf{W}\right) \\
&= tr\left(\mathbf{W}^T\mathbf{S}_y\mathbf{W}\right).
\end{aligned}$$

This shows that the solution given in (6) maximizes the arithmetic mean of the squared Euclidean distances of negative samples in the feature space. It can be expected that the difficulty encountered in finding the optimal projection by BDA, demonstrated in Fig. 2, is caused by the squared distances in the objective function. This point of view was considered in [18] where the conventional BDA can be affected too much by some negative samples referred to as outliers, which results in the degradation of classification performance. Two alternatives, SBDA and L1-BDA, were suggested in [18] and it was shown that they could alleviate the negative effects of outliers to some extent. However, both methods also fail to find the optimal solution to the problem in Fig. 2, as shown in Fig. 3(b).

Another reason why the conventional BDA fails for the problem shown in Fig. 2 is that it focuses on the average distance of negative samples in the feature space. Since the positive samples are normally distributed around their mean after whitening $\mathbf{S}_x$, it is desirable to place negative samples with small distances from the mean of positive samples ($\mathbf{m}_x$) as far away as possible in the lower dimensional feature space. However, maximizing the arithmetic mean is guaranteed to maximize the average squared distances of negative samples from $\mathbf{m}_x$ in the feature space, regardless of the distribution of the negative samples close to $\mathbf{m}_x$. This problem can

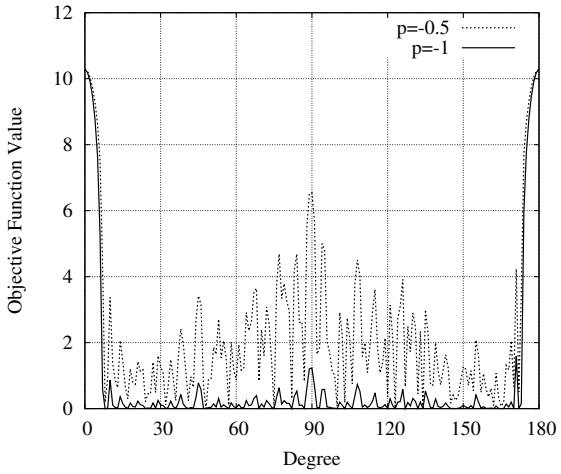


Figure 4: The objective functions for $p = -0.5$ and $p = -1$.

be overcome by substituting the arithmetic mean with the generalized mean. As mentioned in Subsection 2.2, the generalized mean is dominated by small values of given positive numbers as p decreases below zero whereas the arithmetic mean depends equally on every number. This property of the generalized mean makes it possible to construct a more discriminative feature space. Indeed, for the negative samples in the problem shown in Fig. 2, a modified objective function with a generalized mean of $p = -0.5$ or $p = -1$ is maximized at 0 and 180 degrees, as shown in Fig. 4, both of which gives the optimal projection for the problem. This indicates that the objective function using the generalized mean with an appropriate p value can provide a more discriminative feature space than SBDA or L1-BDA. In the following subsections, we show how to find the solution for the objective function using the generalized mean.

3.2. Maximizing the generalized mean: 1st-order approximation

Under the assumptions that $\mathbf{y}_i$ denotes the i-th negative sample in the space where $\mathbf{S}_x$ has been whitened and that all of the samples are transformed using (4),

the squared distance between $\mathbf{y}_i$ and the mean of positive samples in the feature space can be represented as

$$\begin{aligned} d_i(\mathbf{W}) =& \mathbf{y}_i^T \mathbf{W}\mathbf{W}^T \mathbf{y}_i \\ =& tr\left(\mathbf{W}^T \mathbf{S}_i \mathbf{W}\right), \end{aligned}$$

where $\mathbf{S}_i = \mathbf{y}_i \mathbf{y}_i^T$. If the arithmetic mean is replaced by the generalized mean, the optimization problem in (6) becomes

$$\begin{aligned} \mathbf{W}_2 &= \arg\max_{\mathbf{W}} J(\mathbf{W}) \\ &= \arg\max_{\mathbf{W}} \left(\frac{1}{N_y} \sum_{i=1}^{N_y} d_i(\mathbf{W})^p \right)^{1/p}, \\ &\textit{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}. \end{aligned} \tag{7}$$

Note that this problem is equivalent to the problem given in (6) if $p = 1$. The necessary condition for $\mathbf{W}$ to be a local maximum is that the gradient of the objective function with respect to $\mathbf{W}$ is zero, and the gradient of (7) is calculated as

$$\begin{aligned} \nabla_{\mathbf{W}} =& \frac{\partial}{\partial \mathbf{W}} \left[\left(\frac{1}{N_y} \sum_{i=1}^{N_y} d_i(\mathbf{W})^p \right)^{1/p} \right] \\ =& \frac{2}{N_y} \left(\frac{1}{N_y} \sum_{i=1}^{N_y} d_i(\mathbf{W})^p \right)^{\frac{1-p}{p}} \left(\sum_{i=1}^{N_y} d_i(\mathbf{W})^{p-1} \mathbf{S}_i \right) \mathbf{W}. \end{aligned} \tag{8}$$

Unlike the optimization problem in (6) it is difficult to find a closed form solution that makes (8) equal to zero. Furthermore, the objective function is not convex if $p < 0$, as shown in Fig. 4. In this case, $\mathbf{W}_2$ can be obtained iteratively using the gradient method [29]. Its update rule at the t-th iteration is represented as

$$\mathbf{W}' = \mathbf{W}^{(t)} + \eta \nabla_{\mathbf{W}}, \tag{9}$$

where η is the learning rate. In order to satisfy the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, the following process is included in each iteration:

$$\mathbf{W}^{(t+1)} = \text{Orth}(\mathbf{W}^{'}). \tag{10}$$

where Orth($\mathbf{A}$) is the orthonormalization of a matrix $\mathbf{A}$, which can be performed by QR decomposition [30]. Using a randomly initialized $\mathbf{W}$ that satisfies the orthonormality condition, (9) and (10) are repeated until the difference in the objective function values during an iteration is less than a small positive number ϵ, i.e.,

$$|J(\mathbf{W}^{(t+1)}) - J(\mathbf{W}^{(t)})| < \epsilon. \tag{11}$$

It is well known that the solutions obtained from gradient based optimization methods such as (9) are sensitive to initial points. One approach to mitigate this problem is to solve the optimization problem many (N_r) times with different initial values, and $\mathbf{W}_2$ is selected that yields the maximum convergent value. However, this approach is unlikely to find a satisfactory local maximum because the objective function has a large number of local maxima as shown in Fig. 4. This occurs when maximizing the generalized mean with $p < 0$ because $d_i(\mathbf{W})^p$ becomes extremely large when $d_i(\mathbf{W})$ is close to zero. In order to eliminate large fluctuations in the objective function, we modify the objective function by adding a small positive number ρ to each $d_i(\mathbf{W})$, i.e.,

$$\widehat{d_i}(\mathbf{W}) = d_i(\mathbf{W}) + \rho. \tag{12}$$

Although this modification distorts the original objective function, it effectively prevents $d_i(\mathbf{W})^p$ from being very large when $p < 0$. Figure 5 shows the original and the modified objective functions with $p = -1$ in (7). Note that as ρ increases, the objective function becomes smoother and the number of local maximum points decreases. This figure also shows the recommended range for ρ, i.e., it is acceptable

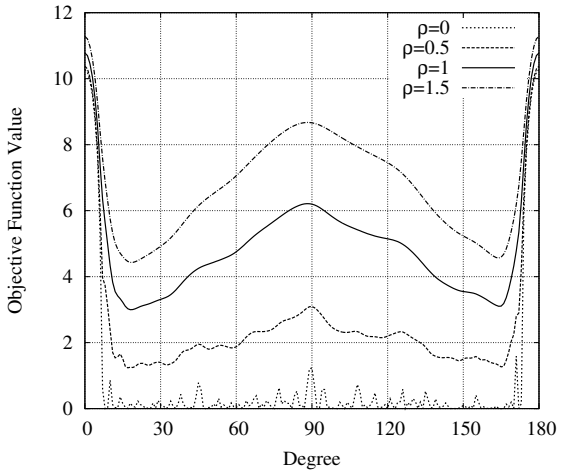


Figure 5: The original and smoothed objective functions with $p = -1$.

to set ρ to a number between 0.5 and 1.5 after $\mathbf{S}_x$ is whitened. This procedure is summarized in Algorithm 1. However, this gradient-based method still has some drawbacks. In particular, it converges very slowly. In the next subsection, we will present a novel method that finds the solution more efficiently.

3.3. Maximizing the generalized mean: 2nd order approximation

The slow convergence of gradient-based methods is due to the first order approximation of the objective function. In order to improve the convergence speed, we propose to approximate the objective function in (7) as a second order function of $\mathbf{W}$ by the following theorem.

Theorem 1. *For a fixed $p \neq 0$ and N positive numbers $\{a_1, \ldots, a_N\}$, the generalized mean $\mathcal{M}_p\{a_1, \ldots, a_N\}$ can be expressed as a nonnegative combination of $a_1, a_2, \ldots, a_N$, i.e.,*

$$\left(\frac{1}{N}\sum_{i=1}^{N} a_i^p\right)^{1/p} = c_1 a_1 + c_2 a_2 + \cdots + c_N a_N \tag{13}$$

Algorithm 1 The gradient-based method

Input: $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_y}\}$, m, p, η, N_r, and ϵ.
for $r = 1 \rightarrow N_r$ **do**
  $t \longleftarrow 0$.
  Randomly initialize an orthonormal matrix $\mathbf{W}^{(0)} \in \Re^{n \times m}$.
  Calculate $J(\mathbf{W}^{(0)})$.
  repeat
    $t \longleftarrow t + 1$.
    Calculate the gradient using (8).
    Update $\mathbf{W}^{(t)}$ using (9) and (10).
  until $|J(\mathbf{W}^{(t)}) - J(\mathbf{W}^{(t-1)})| < \epsilon$
  $\mathbf{W}_r \longleftarrow \mathbf{W}^{(t)}$.
end for
Output: $\mathbf{W}_2 = \arg\max_{\mathbf{W}_r} J(\mathbf{W}_r)$.

where

$$c_i = \frac{1}{N}\left(\frac{1}{N}\sum_i a_i^p\right)^{\frac{1}{p}-1} a_i^{p-1} > 0. \tag{14}$$

Proof. By replacing c_i in (13) with (14), the equality holds in (13). Each c_i in (14) can be obtained by differentiating (13) with respect to a_i. □

Note that each c_i has the same value $\frac{1}{N}$ if $p = 1$, where the generalized mean becomes the arithmetic mean. We can see in (14) that the coefficient c_i corresponding to a smaller a_i becomes relatively larger as p decreases below one. This is related to the property that the generalized mean is a monotonically increasing function of p as described in Subsection 2.2. For a fixed $\mathbf{W}$, the generalized mean in (7) can be represented as a linear combination of $d_i(\mathbf{W})$ by Theorem 1.

$$\begin{aligned}\left(\frac{1}{N_y}\sum_{i=1}^{N_y} d_i(\mathbf{W})^p\right)^{\frac{1}{p}} &= \alpha_1 d_1(\mathbf{W}) + \cdots + \alpha_{N_y} d_{N_y}(\mathbf{W}) \\ &= \sum_{i=1}^{N_y} \alpha_i d_i(\mathbf{W}),\end{aligned}$$

where

$$\alpha_i = \frac{1}{N_y}\left(\frac{1}{N_y}\sum_{i=1}^{N_y} d_i(\mathbf{W})^p\right)^{\frac{1-p}{p}} d_i(\mathbf{W})^{p-1}. \tag{15}$$

Therefore, the objective function near the fixed $\mathbf{W}$ can be approximated as

$$\begin{aligned}&\left(\frac{1}{N_y}\sum_{i=1}^{N_y} d_i(\mathbf{W})^p\right)^{\frac{1}{p}} \approx \sum_{i=1}^{N_y} \alpha_i tr\left(\mathbf{W}^T\mathbf{S}_i\mathbf{W}\right) \\ &= tr\left(\mathbf{W}^T\mathbf{S}_\alpha\mathbf{W}\right),\end{aligned}$$

where $\mathbf{S}_\alpha = \sum_{i=1}^{N_y} \alpha_i \mathbf{S}_i$. If we assume that $\mathbf{S}_\alpha$ is fixed, then the optimal solution of the approximated objective function, i.e.,

$$\widehat{\mathbf{W}} = \arg\max_{\mathbf{W}} tr\left(\mathbf{W}^T\mathbf{S}_\alpha\mathbf{W}\right), \tag{16}$$

can easily be obtained by the eigenvalue decomposition of $\mathbf{S}_\alpha$. However, $\widehat{\mathbf{W}}$ does not guarantee $J(\widehat{\mathbf{W}}) > J(\mathbf{W})$ because $\widehat{\mathbf{W}}$ is the maximizer of the approximated objective function near the fixed $\mathbf{W}$ rather than the original objective function. To find $\mathbf{W}^{(t+1)}$, which satisfies $J(\mathbf{W}^{(t+1)}) > J(\mathbf{W}^{(t)})$, it is necessary to include a line search step, as follows:

$$\begin{aligned} \mathbf{W}' &\longleftarrow \beta\widehat{\mathbf{W}} + (1-\beta)\mathbf{W}^{(t)}, \\ \mathbf{W}^{(t+1)} &\longleftarrow \text{Orth}(\mathbf{W}'), \end{aligned} \tag{17}$$

where β is a search parameter. We initially set β to 1, and reduce it by $\beta_{j+1} \longleftarrow \delta\beta_j$, $(0 < \delta < 1)$ until $J(\mathbf{W}^{(t+1)}) > J(\mathbf{W}^{(t)})$. This step is stopped if j is equal to or greater than a predefined value. In this paper, δ is set to 0.5. As mentioned above, $d(\mathbf{W})$ is replaced by $\widehat{d}(\mathbf{W})$ in (12) to smooth the objective function when $p < 0$. This changes $\mathbf{S}_\alpha$ in (16) to

$$\widehat{\mathbf{S}}_\alpha = \sum_{i=1}^{N_y} \alpha_i \left(\mathbf{y}_i \mathbf{y}_i^T + \frac{\rho}{n} \right),$$

where n is the dimensionality of $\mathbf{y}_i$.

These processes are summarized in Algorithm 2. As a simple test, Algorithm 2 was applied to the problem shown in Fig. 2. The samples in the figure were first transformed by $\mathbf{W}_1$ such that $\mathbf{W}_1^T \mathbf{S}_x \mathbf{W}_1 = \mathbf{I}_2$, and p, N_r, and ρ were set to -1, 1, and 1, respectively. The initial value $\mathbf{W}^{(0)}$ was set as the unit vector rotated from the x axis by $+80$ degrees, and Algorithm 2 was terminated at $t = 3$. We obtained the solution, which is the unit vector rotated by $+178.05$ degrees. Note that the optimal solution of the problem, which was 0 or 180 degrees, could not be founded by Algorithm 1 if $\mathbf{W}^{(0)}$ was set to degree 80. The positive and negative samples in Fig. 2 were well discriminated in the feature space generated by Algorithm 2, but the solution $\mathbf{W}_2$ was not exactly the same as to the optimal solution in Fig. 5. From this observation, we can see that Algorithm 2 gives a solution that is close to

Algorithm 2 The 2nd-order approximation using Theorem 1

Input: $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_y}\}$, m, p, N_r, β, δ, and ϵ.
for $r = 1 \rightarrow N_r$ **do**
  $t \longleftarrow 0$.
  Randomly initialize an orthonormal matrix $\mathbf{W}^{(0)} \in \Re^{n \times m}$.
  Calculate $J(\mathbf{W}^{(0)})$.
  repeat
    $t \longleftarrow t + 1$.
    Calculate each α_i using (15).
    Find $\widehat{\mathbf{W}}$ by solving (16).
    Obtain a new $\mathbf{W}^{(t)}$ satisfying $J(\mathbf{W}^{(t)}) > J(\mathbf{W}^{(t-1)})$ by (17).
  until $|J(\mathbf{W}^{(t)}) - J(\mathbf{W}^{(t-1)})| < \epsilon$
  $\mathbf{W}_r \longleftarrow \mathbf{W}^{(t)}$.
end for
Output: $\mathbf{W}_2 = \arg\max_{\mathbf{W}_r} J(\mathbf{W}_r)$.

a local optimum rather than the local optimum, although it converges much faster than Algorithm 1. Therefore, we perform Algorithm 2 and then Algorithm 1 by setting its $\mathbf{W}^{(0)}$ as the output of Algorithm 2. Indeed, a satisfactory solution of 179.97 degrees was found for the problem in Fig. 2 by performing Algorithm 1 successively.

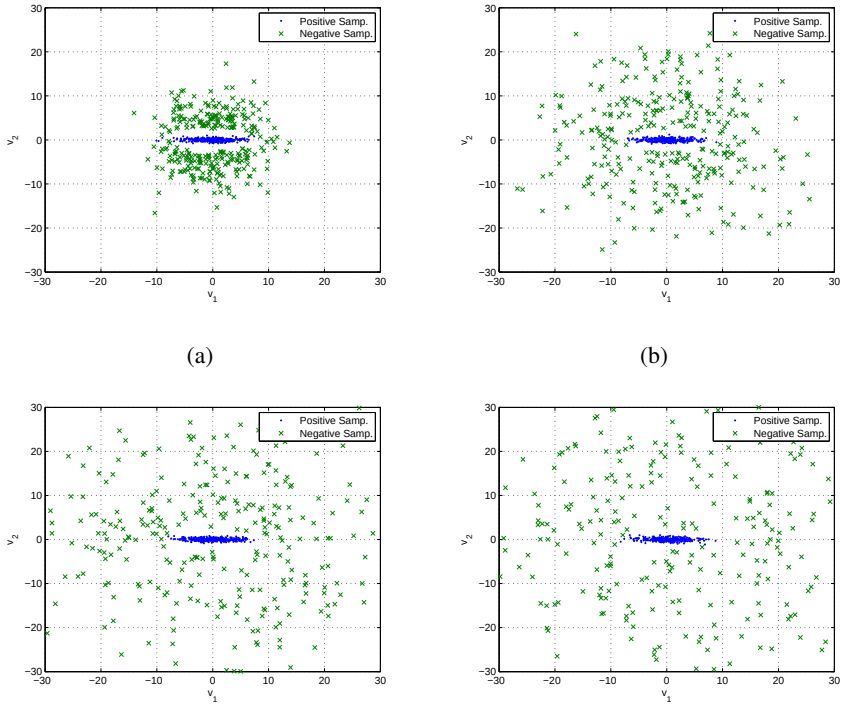
Finally, the method proposed in this paper, which is referred to as *BDAGM*, is described in Algorithm 3. The first step is to whiten $\mathbf{S}_x$ as BDA. It then finds N_r local maxima by running Algorithm 2 N_r times using random initial values. This allows us to rapidly search for intermediate solutions, among which we select the solution that yields the largest value for the objective function to set as the initial value for Algorithm 1. Then, we find the solution $\mathbf{W}_2$ by Algorithm 1, and finally obtain $\mathbf{W}_{BDAGM} = \mathbf{W}_1\mathbf{W}_2$.

Algorithm 3 BDA using the Generalized Mean (BDAGM)

1: **Input:** $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_x}\}$, $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_y}\}$, p and m.
2: Calculate $\mathbf{S}_x$ from $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_x}\}$.
3: Calculate $\mathbf{W}_1$ such that $\mathbf{W}_1^T\mathbf{S}_x\mathbf{W}_1 = \mathbf{I}_n$.
4: Project $\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_y}\}$ using (4).
5: Perform Algorithm 2 N_r times.
6: Calculate $\mathbf{W}_2$ by Algorithm 1.
7: **Output:** $\mathbf{W}_{BDAGM} = \mathbf{W}_1\mathbf{W}_2$.

4. Experiments

In this section, the effectiveness of BDAGM is demonstrated by experimental results on some artificial and real-world data sets.



(a) (b)

(c) (d)

Figure 6: The toy problems (a) $\sigma_y = 5$ (b) $\sigma_y = 10$ (c) $\sigma_y = 15$ (d) $\sigma_y = 20$.

Table 1: The average number of false positives

σ_y	BDA	SBDA	L1-BDA	BDAGM
5	9.13	9.12	9.08	8.48
10	12.23	12.19	12.22	11.10
15	10.37	10.36	10.35	8.91
20	9.14	9.08	9.11	7.09

4.1. Toy problems

BDAGM was motivated by the simple problem depicted in Fig. 2. Although the problem shows the advantage of BDAGM over other conventional methods, the distribution of the negative samples in the problem is somewhat artificial. In order to evaluate the proposed method in more realistic situations, we considered four more toy problems, each of which consists of 300 positive and 300 negative samples. For each problem, positive samples were randomly generated from a two-dimensional Gaussian distribution with the mean $\mathbf{m}_x = 0$ and covariance matrix $\mathbf{\Sigma}_x = diag\left[3, \frac{1}{3}\right]$, whereas negative samples were randomly generated from another two-dimensional Gaussian distribution with $\mathbf{m}_y = 0$ and $\mathbf{\Sigma}_y = diag\left[\sigma_y^2, \sigma_y^2\right]$, where σ_y is set to 5, 10, 15, and 20, respectively. We made 500 sets of such positive and negative samples for each problem. To make the positive and negative samples mostly separated in the input space, the following constraint was imposed on each negative sample $\mathbf{y}$:

$$\mathbf{y}^T \mathbf{\Sigma}_x^{-1} \mathbf{y} > 20.$$

Figure 6 shows the distributions of positive and negative samples corresponding to different σ_y values. For the purpose of comparison, BDA, SBDA, L1-BDA, and BDAGM were performed to obtain one-dimensional subspace. To evaluate the discriminating power of each method, the samples were projected to the subspace obtained by each method, and the distance from the mean of positive samples in the subspace was computed. Then, the number of negative samples, whose distances were smaller than the maximum distance of the projected positive samples, were counted. With this setting, this number corresponds to the number of false positives with zero miss rate. To find a suitable value of p in BDAGM, we conducted BDAGM ten times using the values $\{-0.1, -0.2, \ldots, -1\}$ and chose the value showing the best performance. Table 1 shows the average numbers of false

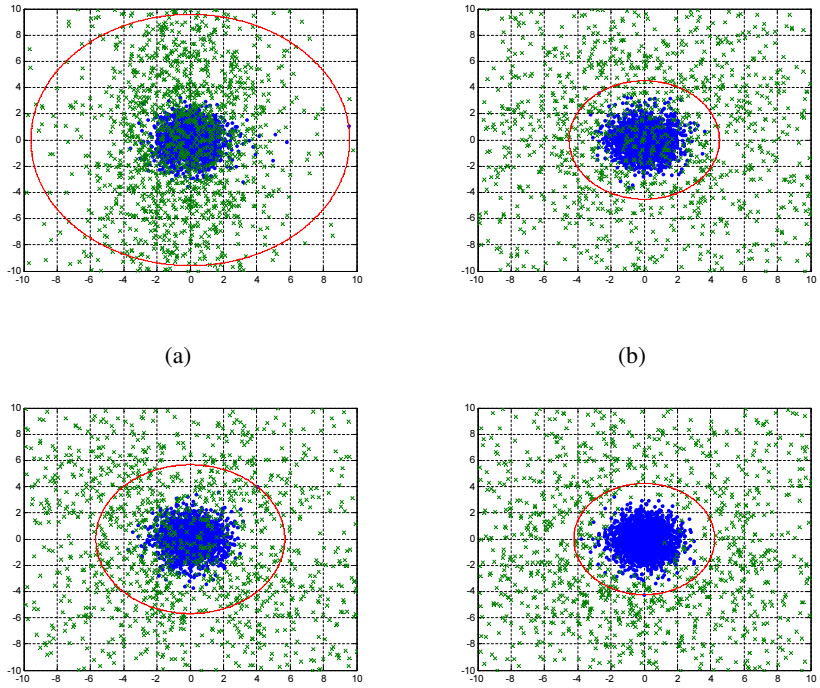
Figure 7: Examples of face and non-face images.

positives for all of the 500 sets. On average, for these problems, the performances of BDA, SBDA, and L1-BDA are almost the same, but BDAGM yields a better performance. This means that the proposed method can provide better discriminative features than the conventional methods if p is appropriately chosen.

4.2. Discriminating faces from non-faces

We tested the proposed method on three real-world problems. The first problem is to distinguish face images from non-face images. We collected 2000 face and 2000 non-face images from the Color FERET face database [31]. The face images were cropped based on the centers of the right and left pupils and they were resized to 24×24 pixels. The non-face images were also randomly cropped and resized to have the same size. Histogram equalization was applied to the resized face and non-face images for illumination normalization. Figure 7 shows some of the normalized face and non-face images. Each of the images was converted into a 576-dimensional vector. In this experiment, the face images were considered to be positive samples whereas the non-face images were considered to be negative samples. The regularization method was applied using $\mu = 0.01$ for $\mathbf{S}_x$, and the parameter γ in SBDA was also set to 3, which is the value recommended in [18]. To select an appropriate value of p, BDAGM was performed for the values of p in $\{-0.25, -0.5, -0.75, -1, -1.25, -1.5, -1.75, -2\}$.

After performing BDA, SBDA, L1-BDA, and BDAGM using all of 2000 pos-



(a) (b)

(c) (d)

Figure 8: The positive and negative samples projected onto the feature spaces, which are constructed using (a) BDA, (b) SBDA, (c) L1-BDA, and (d) BDAGM with $p = -0.25$.

itive and 2000 negative samples, the samples were projected to the corresponding feature space. Figure 8 shows the positive and negative samples projected to the two-dimensional feature spaces. As in Figs. 2 and 6, the positive samples are denoted as $\cdot$, whereas the negative samples are denoted as $\times$. The radius of each circle in the figure is the maximum distance of the positive samples from the origin, which corresponds to the mean of the positive samples. We can see that there are much less negative samples near the origin in Fig. 8(d) than in Figs. 8(a), 8(b), and 8(c). This desirable characteristic is attributed to the use of the generalized mean with $p < 0$. The negative samples in the circles represent the false positives if the threshold for classification is set to the maximum distance of the positive samples from the origin. Note that the numbers of the negative samples within the four circles in Fig. 8 are 1333, 318, 537, and 127, respectively. This indicates that the proposed method can produce more discriminative features than BDA, SBDA, and L1-BDA.

To compare the generalization ability of the proposed method with BDA, SBDA and L1BDA, a 10-fold cross validation was performed using 2000 positive and 2000 negative samples. In this test, after obtaining feature spaces using a training set, we projected the test samples to the feature space and computed the hit rate, i.e., the ratio of the positive samples among the 200 test samples that were near the mean of the positive samples in a training set. If the hit rate is 100%, all of the positive samples in a test set can be distinguished from the negative samples in the test set by comparing their distances from $\mathbf{W}^T\mathbf{m}_x$ and a proper threshold. Also, training samples were normalized before extracting features so as to make each input variable have zero-mean and unit-variance, and test samples were normalized before projecting into feature space. This normalization was applied in each of the following experiments. Table 2 shows the average hit rates with the corresponding standard deviations. Regardless of the number of extracted features m, the best hit

Table 2: Hit rates of face/non-face problem using 10-fold cross validation (%)

m	BDA	SBDA	L1-BDA	BDAGM (p)							
				-0.25	-0.5	-0.75	-1	-1.25	-1.5	-1.75	-2
2	81.60 ±1.78	88.35 ±0.91	88.65 ±1.29	**91.10** ±1.47	90.65 ±0.75	90.15 ±1.20	90.75 ±1.46	90.75 ±0.75	90.05 ±2.71	89.35 ±2.78	87.80 ±2.06
4	91.25 ±1.38	94.35 ±0.67	93.05 ±1.42	**95.15** ±0.97	94.90 ±0.77	95.10 ±0.97	**95.15** ±0.85	**95.15** ±1.00	95.00 ±0.82	94.90 ±0.70	94.75 ±1.09
6	93.75 ±1.14	96.25 ±0.98	95.15 ±1.45	96.55 ±0.72	96.75 ±0.92	96.65 ±0.75	96.60 ±0.46	96.60 ±0.81	**96.95** ±0.69	96.45 ±0.76	96.75 ±0.86
8	94.70 ±0.95	97.05 ±0.55	96.00 ±1.18	97.55 ±0.55	97.25 ±0.59	97.30 ±0.54	97.40 ±0.39	**97.85** ±0.58	97.65 ±0.85	97.25 ±0.95	97.30 ±0.48
10	95.40 ±0.70	97.30 ±0.54	96.40 ±0.94	97.60 ±0.61	97.60 ±0.52	97.70 ±0.63	97.75 ±0.68	97.95 ±0.60	97.90 ±0.74	**98.05** ±0.83	97.70 ±0.59
12	95.90 ±0.84	97.75 ±0.92	96.70 ±0.95	97.70 ±0.67	98.00 ±0.71	97.90 ±0.77	98.00 ±0.78	**98.10** ±0.77	97.80 ±0.86	**98.10** ±0.81	98.05 ±0.69
14	95.90 ±0.99	97.70 ±0.89	96.95 ±1.01	97.90 ±0.77	97.75 ±0.86	97.95 ±1.04	98.00 ±0.82	98.05 ±0.64	**98.10** ±0.94	97.95 ±0.69	97.90 ±0.52
16	95.90 ±1.33	97.60 ±0.61	97.25 ±0.92	97.75 ±1.01	97.70 ±0.95	97.95 ±0.83	97.95 ±0.93	97.75 ±0.90	**98.10** ±0.74	97.75 ±0.90	98.00 ±1.03
18	96.10 ±1.22	97.55 ±0.93	97.40 ±0.91	97.70 ±0.67	97.70 ±1.03	97.70 ±0.98	97.70 ±0.82	98.15 ±0.91	98.05 ±0.69	**98.20** ±0.92	98.10 ±0.81
20	96.15 ±1.06	97.55 ±0.96	97.30 ±0.98	97.60 ±0.94	97.80 ±0.82	97.90 ±0.99	97.90 ±0.88	98.05 ±0.83	98.00 ±0.82	98.00 ±0.88	**98.20** ±0.79

rate of BDAGM was higher than those of the other conventional methods. Especially, when $m = 2$, the BDAGM with $p = -0.25$ resulted in 9.5% higher average hit rate than BDA, and 2.75% and 2.45% higher than SBDA and L1-BDA. When $m = 20$, the average hit rate of BDAGM with $p = -2$ was 98.2%, which was 2.05%, 0.65%, and 0.9% better than BDA, SBDA, and L1-BDA, respectively. It is worth noting that except for the cases of $m = 2$ and $m = 12$, the lowest hit rate of BDAGM is greater than the highest hit rate of BDA, SBDA, and L1-BDA.

In order to interpret the statistical significance of the results in Table 2, the one tailed Welch's t-test [32] was performed. The null hypothesis H_0 and the alternative hypothesis H_A for the test were defined as follows:

- H_0: For the same dimensionality m, the maximum performance of BDA, SBDA, L1-BDA is equal to the maximum performance of BDAGM.

Table 3: Welch's t-test for the results in Table 2

m	2	4	6	8	10	12	14	16	18	20
T-value	3.961	2.337	1.847	3.164	2.395	0.922	0.977	1.648	1.571	1.653
DOF	18	17	16	18	15	17	18	17	18	17
$T_{95\%}$	1.734	1.740	1.746	1.734	1.753	1.740	1.734	1.740	1.734	1.740
$T_{90\%}$	1.330	1.333	1.337	1.330	1.341	1.333	1.330	1.333	1.330	1.333
Accepted (95%)	H_A	H_A	H_A	H_A	H_A	H_0	H_0	H_0	H_0	H_0
Accepted (90%)	H_A	H_A	H_A	H_A	H_A	H_0	H_0	H_0	H_A	H_A

Figure 9: Examples of car and non-car images.

- H_A: For the same dimensionality m, the maximum performance of BDAGM is higher than the maximum performance of BDA, SBDA, L1-BDA.

Table 3 shows the computed T-values, degree of freedom (DOF), and the target T-values $T_{95\%}$ and $T_{90\%}$. The test indicates that H_0 is rejected with 95% (90%) of confidence if T-value is greater than $T_{95\%}$ ($T_{90\%}$), thus H_A is accepted. Actually, H_A was accepted with 95% of confidence when m is 2, 4, 6, 8, and 10. Note that for the cases of $m = 2$ and $m = 8$, T-values are greater than 3, which makes the alternative hypotheses accepted with 99% of confidence. When $m = 18$ and $m = 20$, the alternative hypotheses can be accepted if the confidence level decreases to 90%. These results demonstrate that BDAGM outperforms BDA, SBDA, and L1-BDA in the problem of discriminating face images from non-face images.

Table 4: Hit rates of vehicle/non-vehicle problem using 10-fold cross validation (%)

m	BDA	SBDA	L1-BDA	BDAGM (p)							
				-0.25	-0.5	-0.75	-1	-1.25	-1.5	-1.75	-2
2	89.05	89.40	90.25	93.30	**93.45**	92.60	92.75	90.45	90.00	87.55	89.05
	±1.54	±1.61	±1.55	±1.77	±1.76	±1.13	±1.18	±1.80	±2.29	±3.27	±3.10
4	91.40	93.30	93.15	94.55	95.05	95.35	95.45	**95.60**	95.40	95.25	95.30
	±1.45	±1.34	±1.45	±1.23	±1.19	±0.97	±1.04	±1.13	±1.52	±0.86	±1.23
6	91.65	95.00	93.35	95.10	95.30	95.45	95.45	95.50	95.70	95.60	**95.85**
	±1.45	±1.11	±1.31	±1.29	±1.30	±1.40	±1.14	±1.15	±0.75	±0.99	±0.91
8	91.60	94.15	93.60	94.70	95.20	95.30	95.35	95.60	95.65	95.65	**95.75**
	±1.20	±1.27	±1.52	±1.21	±1.40	±1.21	±1.16	±0.94	±1.20	±0.94	±1.21
10	92.20	94.30	93.45	94.30	94.50	94.75	94.75	94.80	95.10	95.10	**95.20**
	±1.06	±0.98	±1.34	±1.30	±1.35	±1.21	±1.00	±1.09	±0.99	±0.91	±0.79
12	92.20	94.00	93.80	94.00	94.15	94.45	94.35	94.50	**94.65**	94.50	**94.65**
	±1.09	±1.33	±1.30	±1.29	±1.06	±0.80	±1.06	±1.03	±0.91	±0.97	±0.85
14	92.60	93.65	93.30	93.70	94.10	94.15	94.05	93.95	94.40	**94.50**	94.45
	±1.54	±1.20	±1.18	±1.14	±0.91	±0.97	±1.01	±0.98	±0.99	±0.88	±1.07
16	92.05	93.70	92.95	93.75	94.15	94.10	**94.20**	93.95	93.95	93.95	94.00
	±1.35	±1.06	±1.21	±1.14	±0.94	±0.91	±0.79	±1.17	±1.14	±1.32	±1.31
18	92.05	93.35	92.95	93.70	**93.85**	93.70	**93.85**	93.75	93.80	93.60	93.60
	±1.23	±1.08	±1.21	±0.98	±1.03	±1.11	±1.13	±1.06	±1.21	±1.24	±1.05
20	91.65	92.90	92.65	93.30	**93.70**	93.55	93.50	93.50	93.25	93.35	93.35
	±1.18	±1.02	±1.29	±1.09	±0.86	±1.07	±1.13	±1.05	±1.18	±1.13	±1.29

Table 5: Welch's t-test for the results in Table 4

m	2	4	6	8	10	12	14	16	18	20
T-value	4.315	4.419	1.873	2.884	2.261	1.302	1.806	1.196	1.059	1.896
DOF	18	18	17	18	17	15	17	17	18	18
$T_{95\%}$	1.734	1.734	1.740	1.734	1.740	1.753	1.740	1.740	1.734	1.734
$T_{90\%}$	1.330	1.330	1.333	1.330	1.333	1.341	1.333	1.333	1.330	1.330
Accepted (95%)	H_A	H_A	H_A	H_A	H_A	H_N	H_A	H_N	H_N	H_A
Accepted (90%)	H_A	H_A	H_A	H_A	H_A	H_N	H_A	H_N	H_N	H_A

4.3. Discriminating vehicles from non-vehicles

The next experiment aimed to distinguish vehicle images from non-vehicle images. Some samples of the vehicle and non-vehicle images are shown in Fig. 9. In total, 2000 rear images of vehicles and 4000 random non-vehicle images were collected from video sequences, which were captured using a camera mounted on a moving vehicle. In this experiment, each vehicle image was considered to be a positive sample, whereas non-vehicle images were regarded as negative samples. Each positive sample was scaled and aligned to an image of 24×24 pixels while the negative samples were resized to 24×24 pixels. After this scale normalization, histogram equalization was performed for all of the samples to normalize the illumination variation. The 10-fold cross validation was also conducted to compare the generalization performance of BDAGM with those of BDA, SBDA, L1-BDA. Note that the numbers of positive and negative samples were not balanced. These uneven class priors are very common in one-class classification problems such as face detection [33]. The same candidates used in the previous experiment were considered for the parameter p of BDAGM. Table 4 shows the average hit rates for the 200 nearest test samples from $\mathbf{W}^T\mathbf{m}_x$ in each feature space. This table demonstrates that BDAGM produced a feature space where the vehicle images could be more readily distinguished from the non-vehicle images compared to BDA and its variants. When $m = 2$ in particular, the BDAGM with $p = -0.5$ delivered 4.4%, 4.05%, and 3.2% higher average hit rates than BDA, SBDA, and L1-BDA, respectively. Table 4 also shows a similar result of the previous experiments, i.e., except for $m = 2$, the minimum hit rates of BDAGM are higher than or equal to the maximum hit rates of BDA, SBDA, and L1BDA.

The Welch's t-test was also performed based on the results in Table 4 as in the previous experiment. Table 5 shows the necessary information for the test and the accepted hypotheses. With 95% of confidence H_A was adopted for all of m values

Table 6: Hit rates of each category in ETH-80 database using 10-fold cross validation ($m = 2$)

Category	BDA	SBDA	L1-BDA	BDAGM (p)							
				-0.25	-0.5	-0.75	-1	-1.25	-1.5	-1.75	-2
Apple	19.61 ±4.95	44.49 ±3.52	31.41 ±4.02	53.95 ±4.67	57.07 ±4.39	59.76 ±4.54	59.80 ±5.53	61.46 ±3.93	**62.15** ±4.69	61.90 ±3.54	60.39 ±5.28
Car	29.90 ±8.46	49.27 ±6.62	49.27 ±8.65	52.59 ±7.53	53.61 ±6.94	54.73 ±6.47	**54.88** ±5.70	54.68 ±6.44	54.05 ±5.14	54.15 ±5.45	54.39 ±5.85
Cow	7.76 ±9.25	37.07 ±3.58	39.02 ±5.14	47.71 ±6.32	48.73 ±6.51	**49.17** ±6.44	48.98 ±6.25	48.05 ±6.66	48.34 ±7.09	48.05 ±6.80	46.49 ±6.58
Cup	29.46 ±18.42	48.78 ±20.07	46.15 ±19.70	52.24 ±19.32	52.73 ±19.07	52.88 ±18.59	53.07 ±19.21	**53.71** ±18.29	51.27 ±19.20	44.44 ±22.12	49.51 ±16.84
Dog	28.10 ±9.15	39.61 ±5.82	40.15 ±6.30	43.27 ±7.72	43.80 ±7.29	43.71 ±6.78	**43.90** ±7.94	42.88 ±7.86	43.41 ±7.86	43.85 ±8.07	42.05 ±7.29
Horse	31.46 ±3.91	34.88 ±4.40	32.88 ±5.24	37.22 ±5.30	38.44 ±6.14	39.71 ±4.81	39.46 ±5.61	40.68 ±6.87	39.27 ±5.65	40.10 ±6.19	**40.73** ±5.98
Pear	52.44 ±5.40	58.73 ±5.71	64.68 ±4.82	75.27 ±3.98	76.39 ±4.59	**78.54** ±3.36	78.20 ±4.35	77.41 ±4.04	77.46 ±4.56	77.71 ±3.98	75.85 ±3.02
Tomato	71.80 ±8.09	74.88 ±6.76	74.15 ±7.20	84.54 ±3.50	85.66 ±3.62	86.20 ±3.24	**86.78** ±3.52	86.15 ±3.44	85.32 ±4.24	84.44 ±4.15	84.15 ±4.05

except for $m = 12$, $m = 16$, and $m = 18$. Note that T-values for $m = 2$ and $m = 4$ are 4.315 and 4.419, which are large enough for H_A to be accepted with 99.9% of confidence because $T_{99.9\%}$ is 3.610 when DOF is 18. This means that the proposed method can extract better features to distinguish vehicle images from non-vehicle images than BDA as well as SBDA and L1-BDA.

4.4. ETH-80 database

The ETH-80 database [34] contains the images of eighty objects. Each object belongs to one of eight categories, i.e., apple, car, cow, cup, dog, horse, pear, and tomato. Each category contains ten different objects and each object in a category includes 41 images taken from different viewpoints so a total of 3280 images are contained in the database. Since there are eight categories in the database and each category can be considered as a target class for a one-class classification problem, we formulated eight one-class classification problems using the database. In each problem, the training set consisted of images that corresponded to five randomly

Table 7: Hit rates of each category in ETH-80 database using 10-fold cross validation ($m = 10$)

Category	BDA	SBDA	L1-BDA	BDAGM (p)							
				-0.25	-0.5	-0.75	-1	-1.25	-1.5	-1.75	-2
Apple	23.37 ±2.05	57.22 ±7.45	45.80 ±3.65	58.54 ±5.74	58.78 ±6.10	59.02 ±5.59	59.37 ±5.58	60.54 ±4.68	60.78 ±4.34	61.71 ±4.36	**62.05** ±4.33
Car	57.17 ±5.42	61.37 ±7.46	59.90 ±5.84	61.46 ±6.79	62.00 ±7.52	62.24 ±7.15	62.54 ±7.29	**62.88** ±7.32	62.83 ±7.18	62.83 ±7.32	62.78 ±7.47
Cow	43.51 ±6.61	49.85 ±6.77	47.22 ±6.35	51.90 ±6.10	53.27 ±6.46	54.73 ±6.94	55.02 ±6.91	55.76 ±6.80	55.80 ±7.29	**56.15** ±7.20	**56.15** ±7.40
Cup	62.20 ±11.44	65.37 ±10.27	64.39 ±10.69	65.71 ±10.24	65.56 ±10.28	65.61 ±10.14	65.51 ±10.33	65.66 ±10.44	65.80 ±10.32	**65.90** ±10.34	65.66 ±10.30
Dog	42.29 ±6.06	47.12 ±9.46	46.10 ±7.84	47.46 ±9.98	47.76 ±10.65	**48.29** ±10.91	47.90 ±11.63	47.95 ±11.91	47.66 ±11.89	47.22 ±11.83	47.46 ±11.47
Horse	34.05 ±4.10	51.12 ±4.97	45.61 ±4.86	52.68 ±5.03	52.88 ±4.54	53.76 ±5.11	54.24 ±4.97	54.54 ±5.03	54.78 ±5.14	**55.07** ±4.97	54.54 ±5.37
Pear	68.39 ±5.83	80.54 ±2.99	64.83 ±3.59	82.24 ±3.64	83.22 ±3.29	82.98 ±3.21	83.46 ±3.51	**83.56** ±3.61	83.32 ±3.81	83.41 ±3.34	83.46 ±3.55
Tomato	74.98 ±5.05	86.49 ±2.74	80.20 ±3.43	86.68 ±2.46	86.39 ±2.44	86.68 ±2.44	86.78 ±2.47	86.78 ±2.37	86.93 ±2.20	87.02 ±2.07	**87.32** ±2.53

Table 8: Welch's t-test for the results in Tables 6 and 7

m	Category	Apple	Car	Cow	Cup	Dog	Horse	Pear	Tomato
2	T-value	9.524	2.031	3.895	0.574	1.170	2.492	7.460	4.938
	DOF	17	18	17	18	17	17	16	14
	$T_{95\%}$	1.740	1.734	1.740	1.734	1.740	1.740	1.746	1.761
	$T_{90\%}$	1.333	1.330	1.333	1.330	1.333	1.333	1.337	1.345
	Accepted (95%)	H_A	H_A	H_A	H_N	H_A	H_A	H_A	H_A
	Accepted (90%)	H_A	H_A	H_A	H_N	H_A	H_A	H_A	H_A
10	T-value	1.773	0.457	2.016	0.115	0.256	1.777	2.037	0.704
	DOF	14	18	18	18	18	18	17	18
	$T_{95\%}$	1.761	1.734	1.735	1.734	1.734	1.734	1.740	1.734
	$T_{90\%}$	1.345	1.330	1.330	1.330	1.330	1.330	1.333	1.330
	Accepted (95%)	H_A	H_A	H_A	H_N	H_N	H_A	H_A	H_N
	Accepted (90%)	H_A	H_A	H_A	H_N	H_N	H_A	H_A	H_N

selected objects from each category, and the images of the non-selected objects in each category were used for test. Thus, the number of images in the training and test sets for each problem was 1640, and each set comprised 205 positive samples and 1435 negative samples. As in the previous experiments, after applying BDA, SBDA, L1-BDA and BDAGM to the same training set, test samples were projected to the feature spaces to compute the hit rates. This procedure was repeated ten times using the randomly separated training and test sets. The pixel intensity values of an image were used as the input variables in the previous experiments, whereas the derivative-of-Gaussian filters were applied to generate the input variables in this experiment as in [35]. Each image was convolved with the first derivatives with respect to the x-and y-axes of three Gaussian filters with $\sigma = \{1, 2, 4\}$, and the responses of the six filters were represented as a histogram with 32 bins so that each element in the histogram was used as an input variable.

Tables 6 and 7 show the average hit rates for the 205 nearest test samples in the eight problems where $m = 2$ and $m = 10$, respectively. The tables show that the samples corresponding to tomato were better distinguished than the samples in the other seven categories, and the samples obtained from the pear images were also classified well. Although it was difficult to distinguish images of cows, dogs, or horses from the other categories, BDAGM provided higher hit rates than BDA, SBDA, and L1BDA both for $m = 2$ and $m = 10$. Note that BDAGM with a suitable value of p yielded better discriminative features than BDA, SBDA, and L1-BDA, especially when $m = 2$. For the apple-classification problem, the average hit rate of BDAGM with $p = -1.5$ was 42.54%, 17.66%, and 30.74% higher than those of BDA, SBDA and L1-BDA, respectively. The performance of the tomato-classification problem was improved from 75% to 86.78% when using the BDAGM with $p = -1$ instead of BDA, or its variants. When $m = 10$, BDAGM still produced a better performance than BDA, SBDA, and L1-BDA although the

differences were not significant for some problems. Note that, on average, the minimum performances of BDAGM are higher than those of BDA, SBDA, L1-BDA for most values of m.

As in the previous experiments, the statistical significance of the results in Tables 6 and 7 was also checked by performing the Welch's t-test. Table 8 shows the computed values for the test and the accepted hypotheses. In the case of $m = 2$, H_A was accepted with 95% of confidence for all the problems except the cup-classification problem. It is notable that, for $m = 2$, the T-values associated with apple-, cow-, pear-, and tomato-classification problems are 9.524, 3.895, 7.460, and 4.938. The corresponding alternative hypotheses were accepted with 99.9% of confidence. Also, in the case of $m = 10$, H_A was accepted with 95% of confidence for the apple-, car-, cow-, horse-, and pear-classification problems.

5. Conclusions

This study proposed a novel method to extract discriminative features for one-class classification problems. BDA was developed for one-class classification problems and was considered to be an alternative to LDA in these problems. But, it is sensitive to outliers in negative samples. This drawback arises when maximizing the arithmetic mean of the squared Euclidean distances during the second step of BDA. It is better to place negative samples from the mean of positive samples in the feature space as far away as possible, especially for the negative samples located close to the mean. However, this can not be possible by maximizing the arithmetic mean of the squared distances between negative samples and the mean of positive samples. Thus, we have proposed to use the generalized mean instead of the arithmetic mean in the objective function. SBDA and L1-BDA, which were proposed to address the same problem with BDA, prevent the outliers from being

over-weighted by using other distance schemes instead of the Euclidean distance. On the other hand, different from SBDA and L1-BDA, our proposed method employed an extended concept of the arithmetic mean that can assign different weights to different numbers. However, incorporating the generalized mean makes the objective function become non-convex. Gradient-based iterative methods can be applied to such a problem, but they generally require a long time to find a local optimum of the optimization problem. To develop an efficient method in finding a solution, we exploited the fact that the generalized mean of positive numbers can be represented as a non-negative linear combination of the numbers, and we have finally proposed a novel iterative method based on the fact. We conducted four experiments to demonstrate the usefulness of the proposed method and the results have shown that BDAGM can effectively alleviate the negative effects of outlier to yield better discriminative features for one-class classification problems than BDA, and its state-of-the-art variants SBDA and L1BDA, especially when the number of extracted features is small.

[1] A. Jain, R. Duin, M. Jianchang, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd Edition, Academic Press, New York, 1990.

[3] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley-Interscience, 2001.

[4] A. R. Webb, Statistical Pattern Recognition, 2nd Edition, John Wiley and Sons, 2002.

[5] R. Lotlikar, R. Kothari, Fractional-Step Dimensionality Reduction, IEEE

Transactions on Pattern Analysis and Machine Intelligence 22 (6) (2000) 623–627.

[6] M. Loog, R. Duin, R. Haeb-Umbach, Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (7) (2001) 762–766.

[7] R. Duin, M. Loog, Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: the Chernoff Criterion, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6) (2004) 732–739.

[8] M. Zhu, A. Martinez, Subclass Discriminant Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1274–1286.

[9] A. Martinez, M. Zhu, Where Are Linear Feature Extraction Methods Applicable, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1934–1944.

[10] D. Tao, X. Li, X. Wu, S. Maybank, General Averaged Divergence Analysis, in: Proceedings of IEEE International Conference on Data Mining, 2007.

[11] D. Tao, X. Li, X. Wu, S. Maybank, Geometric Mean for Subspace Selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 260–274.

[12] T. M. Cover, J. A. Thomas, Elementary of Information Theory, John Wiley and Sons, New York, 1991.

[13] W. Bian, D. Tao, Harmonic Mean for Subspace Selection, in: Proceedings of International Conference on Pattern Recognition, 2008.

[14] D. Tao, X. Tang, X. Li, Y. Rui, Direct Kernel Biased Discriminant Analysis: a New Content-Based Image Retrieval Relevance Feedback Algorithm, IEEE Transactions on Multimedia 8 (4) (2006) 716–727.

[15] M.-H. Yang, D. Kriegman, N. Ahuja, Detecting Faces in Images: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.

[16] C. Kim, S.-I. Choi, M. Turk, C.-H. Choi, A New Biased Discriminant Analysis Using Composite Vectors for Eye Detection, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42 (4) (2012) 1095–1106.

[17] X. S. Zhou, T. S. Huang, Small Sample Learning during Multimedia Retrieval using BiasMap, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[18] N. Kwak, J. Oh, Feature extraction for one-class classification problem: enhancements to biased discriminant analysis, Pattern Recognition 42 (1) (2009) 17–26.

[19] P. Viola, M. J. Jones, Robust Real-Time Face Detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[20] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, Vol. 1, pp. 886–893.

[21] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[22] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

[23] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative Common Vectors for Face Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1) (2005) 4–13.

[24] S. Zhang, T. Sim, Discriminant Subspace Analysis: A Fukunaga-Koontz Approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1732–1745.

[25] J. H. Friedman, Regularized Discriminant Analysis, Journal of the American Statistical Association 84 (405) (1989) 165–175.

[26] N. Kwak, Principal Component Analysis Based on L1-Norm Maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (9) (2008) 1672–1680.

[27] P. Bullen, Handbood of Means and Their Inequalities, 2nd Edition, Kluwer Academic Pubulisher, 2003.

[28] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[29] D. G. Luenberger, Y. Ye, Linear and Nonlinear Programming, 3rd Edition, Springer, 2008.

[30] G. Strang, Linear Algebra and Its Applications, 4th Edition, Thomson Brooks/Cole, 2006.

[31] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET Evaluation Methodology for Face-Recognition Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1090–1104.

[32] B. L. Welch, The generalization of 'Student's' problem when several different population variances are involved, Biometrika 34 (1947) 28–35.

[33] J. Wu, S. Brubaker, M. Mullin, J. Rehg, Fast Asymmetric Learning for Cascade Face Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008) 369–382.

[34] B. Leibe, B. Schiele, Analyzing Appearance and Contour Based Methods for Object Categorization, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.

[35] O. Hamsici, A. Martinez, Bayes Optimality in Linear Discriminant Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 647–657.