Generalized Biased Discriminant Analysis for Content-Based Image Retrieval

Lining Zhang[1,2], Lipo Wang[1], *Senior Member, IEEE* and Weisi Lin[3], *Senior Member, IEEE*

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798
[2]Institute for Media Innovation, Nanyang Technological University, Singapore, 637553
[3]School of Computer Engineering, Nanyang Technological University, Singapore, 639798

***Abstract*—Biased Discriminant Analysis (BDA) is one of the most promising Relevance Feedback (RF) approaches to deal with the feedback samples imbalance problem for Content-Based Image Retrieval (CBIR). However, the singular problem of the positive within-class scatter and the Gaussian distribution assumption for positive samples are two main obstacles impeding the performance of the BDA RF for CBIR. To avoid both of these intrinsic problems in BDA, in this paper, we propose a novel algorithm called Generalized Biased Discriminant Analysis (GBDA) for CBIR. The GBDA algorithm avoids the singular problem by adopting the Differential Scatter Discriminant Criterion (DSDC) and handles the Gaussian distribution assumption by redesigning the between-class class scatter with a nearest neighbor approach. To alleviate the overfitting problem, GBDA integrates the locality preserving principle; therefore, a smooth and locally consistent transform can also be learned. Extensive experiments show that GBDA can substantially outperform the original BDA, its variations and related Support Vector Machine (SVM) based RF algorithms.**

***Index Terms*— Content-Based Image Retrieval, Biased Discriminant Analysis, Differential Scatter Discriminant Criterion, Relevance Feedback**

I. Introduction

Relevance Feedback (RF) [1, 2] is one of the most powerful tools to enhance the performance of a Content-Based Image Retrieval (CBIR) system [3, 4]. Most of the RF schemes involve the user into the search engine by letting the user manually label semantically relevant and irrelevant samples, which are positive and negative feedbacks respectively for a query image.

Various RF methods have been developed based on different assumptions for the positive and negative feedbacks during past few years. One-class Support Vector Machine (SVM) estimates the density of positive feedbacks but ignores the negative feedbacks [5]. Two-class SVM can identify the positive and negative feedbacks from each other but treats the two groups equally [6]. In [7], Tao et al believe that positive feedbacks are included in a set and negative feedbacks split into a small number of subset and a series of kernel marginal convex machines have been developed between one positive group and several negative subgroups. The results indicate the clustering the negative samples into several subgroups can indeed improve the overall retrieval performance. By precisely parametrizing positive feedbacks, negative feedbacks and unlabelled samples, Bian and Tao proposed an RF approach, which can find the intrinsic coordinate of image low-level visual features [8]. They also showed that the unlabelled samples are essential in finding this intrinsic coordinate. However, it is generally believed that more samples are actually required to model the exquisite geometry structure in high dimensional space. In [9], Azimi-Sadjadi et al introduced an adaptable CBIR system that incorporates kernel machines and selective sampling technique to capture the hidden user concepts and select the most informative query image during RF. However, kernel methods usually cannot exert its normal capability when the feature dimensions are much higher than the number of training samples.

For an image retrieval task, the need for RF stems from the fact that different semantic concepts lie in different subspaces and the selection of such subspaces cannot be done offline [10]. With the observation that "*all positive examples are alike; each negative example is negative in its own way*", Biased Discriminant Analysis (BDA) was introduced by Zhou and Huang as a principled way to solve the feedback samples imbalance problem and select a subset of image features to construct a suitable dissimilarity measure [10]. The BDA algorithm provides a good solution to this biased learning problem, since there is an unknown number of classes in CBIR but the user is interested in only one class.

However, the original BDA always suffers from the singular problem of the positive within-class scatter matrix because the number of positive samples is much smaller than the dimension of the representative features of images in CBIR [10, 11]. Additionally, the BDA algorithm makes a strong assumption that all positive samples form a single Gaussian distribution [10, 11], which is not true in real-world. These are the main obstacles impeding the performance of BDA for CBIR. Various research efforts have shown that high dimensional samples possibly reside on or close to a nonlinear manifold of ambient space [12-14].

Yu et al proposed a dimension reduction technique based on the hybrid analysis of principal component analysis and linear discriminant analysis which can better integrate discriminative and descriptive information for a specific data distribution [15]. But using this method, it is necessary to find the best parameter pairs setting for a special data distribution. To alleviate the singular problem and the Gaussian distribution assumption for

Manuscript received on August 19, 2010. (This work was supported by Institute for Media Innovation, Nanyang Technological University, under an IMI scholarship.)

positive samples, Direct Biased Discriminant Analysis (DBDA) and Direct Kernel Biased Discriminant Analysis (DKBDA) [16] have been proposed to enhance the performance of BDA by utilizing the direct idea [17] and the kernel trick. However, this approach still discards the null space of negative scatter with respect to the positive centroid, which contains important discriminative features, as pointed out in literatures [18, 19]. Additionally, kernel parameters tuning makes online learning infeasible. As a variant of Marginal Fisher Analysis [20], Marginal Biased Analysis (MBA) was introduced to construct an RF approach and has shown better performance than BDA [21]; however, it still suffers from the intrinsic singular problem in the original BDA.

In this paper, we propose a novel biased discriminant analysis technique, called Generalized Biased Discriminant Analysis (GBDA) for CBIR. To avoid the singular problem in BDA, the GBDA is based on the Differential Scatter Discriminant Criterion (DSDC) [11, 22-26], which defines the inter-class separability as a trace *difference* for the between-class scatter and the within-class scatter rather than a trace *ratio*. Furthermore, to avoid the Gaussian assumption for positive samples, the between-class scatter is specially designed by resorting to a nearest-neighbor approach. Additionally, to reduce the over fitting problem, the locality preserving principle emerging from the manifold learning community [12-14], which measures the local smoothness of the feature transformation, is integrated to regularize the inter-class separability. Therefore, a locally smooth transform can also be learned.

The rest of the paper is organized as follows. We briefly review BDA and DSDC in Section II, thereby introducing the necessary notations. Then the GBDA algorithm is described in Section III. Experiments are reported in Section IV. Section V presents the conclusions.

II. BRIEF REVIEWS ON BDA AND DSDC

A. Biased Discriminant Analysis (BDA)

Zhou *et al.* proposed the BDA as a principled way for CBIR [10], which is actually a (1+x)–class discriminant analysis problem. This means that there is an unknown number of classes but the user is concerned with only one class semantically related to the query.

As a variant of two-class Fisher Linear Discriminant Analysis (FLDA), the BDA aims to find a subspace to discriminate the positive samples from the negative samples. It is spanned by a set of vectors α maximizing the ratio between the biased matrix S_n and the positive covariance scatter matrix S_p, i.e.,

$$\alpha^* = \arg\max_{\alpha} \frac{trace(\alpha^T S_n \alpha)}{trace(\alpha^T S_p \alpha)} \tag{1}$$

In experiments, there are N^p positive and N^n negative samples in the training sets. Then S_n and S_p can be defined as follows:

$$\begin{cases} S_n = \sum_{i=1}^{N^n} (x_i^n - \bar{x}^p)(x_i^n - \bar{x}^p)^T \\ S_p = \sum_{i=1}^{N^p} (x_i^p - \bar{x}^p)(x_i^p - \bar{x}^p)^T \end{cases} \tag{2}$$

where x_i^p denotes the positive samples, x_i^n denotes the negative samples, and $\bar{x}^p = \frac{1}{N^p}\sum_{i=1}^{N^p} x_i^p$ is the mean vector of the positive samples. Usually, the projection matrix α can be computed from the eigenvectors of $S_p^{-1}S_n$, corresponding to the largest eigenvalues. Because the number of feedback samples is usually much smaller than the dimension of features: this will lead to a degenerated S_p, i.e., it is the so called small sample size problem or the singular problem of the positive within-class scatter [10, 11]. In the past decade, a lot of approaches have been proposed to alleviate the singular problem in FLDA [18, 19, 27, 28], which have shown good performance for face recognition.

B. Differential Scatter Discriminant Criterion (DSDC)

Basically, in order to describe the class separability, we should convert the separability measure to a number, which should increase when the between-class scatter increases or the within-class scatter decreases [11, pp.446-447]. DSDC [11, 22-26] defines the separability measure as a trace difference for the between-class scatter and the within-class scatter, rather than a trace ratio, i.e.,

$$\alpha^* = \arg\max_{\alpha^T\alpha=I} [trace(\alpha^T S_b \alpha) - \beta * trace(\alpha^T S_w \alpha)] \tag{3}$$

In Equation (3), β is a nonnegative tuning parameter, and $\alpha \in R^{H*L}, L \ll H, \alpha^T\alpha = I$, is the projection matrix. The matrix S_b is the between-class scatter matrix, which describes the inter-class dispersion, while S_w is the within-class scatter matrix, which describes the intra-class compactness. Both matrices in DSDC have similar meaning as in Fisher Discriminant Criterion (FDC). It is easy to verify that the solution of Equation (3) is equivalent to solving the maximum of the Lagrange function

$$L(\alpha, \lambda) = \sum_{k=1}^{L} \alpha_k^T (S_b - \beta S_w)\alpha_k - \lambda_k(\alpha_k^T\alpha_k - 1) \tag{4}$$

with multipliers λ_k. Let $(\partial L(\alpha_k, \lambda_k)/\partial\alpha_k) = 0, k = 1,..L$, we can have

$$(S_b - \beta S_w)\alpha_k = \lambda_k \alpha_k \quad k = 1,\ldots L \tag{5}$$

Thus, the problem is translated into finding the leading eigenvectors of $(S_b - \beta S_w)$, and hence we need not calculate the inverse of S_w, which allows us to avoid the singular problem of the positive within-class scatter.

Strictly speaking, the solution of Equation (3) is equivalent to the FDC, only if the parameter β in DSDC is calculated as $trace(\alpha_{opt}^T S_b \alpha_{opt})/trace(\alpha_{opt}^T S_w \alpha_{opt})$ [11, 22]. Therefore, the optimal β can only be obtained by the alternating projection method [22]. However, for CBIR, because the distribution of the testing set diverges from that of the training set, a manually chosen value of β always achieves better prediction results than the calculated value. As demonstrated in [23, 25], when the within-class scatter S_w is singular, the discriminant vectors of the DSDC are approaching the discriminant vectors of the null space method of the FDC at $\beta \to \infty$ [27]. When β is set properly, DSDC can show much better performance than the existing methods, which deal with the singular problem in FDC.

III. GENERALIZED BIASED DISCRIMINANT ANALYSIS (GBDA)

Let us denote the high-dimensional space as R^H and the low-dimensional intrinsic space as R^L. For convenience, we define $\mathrm{X}^{\mathrm{p}} = \{x_i^p\}_{i=1}^{N^p} \in R^{H*N^p}$ as the positive samples, $\mathrm{X}^{\mathrm{n}} = \{x_i^n\}_{i=1}^{N^n} \in R^{H*N^n}$ as the negative samples in R^H. Then we use $\mathrm{X}=\{x_i\}_{\mathrm{i}=1}^{(N^p+N^n)} = [X^p, X^n] \in R^{H*(N^p+N^n)}$ to denote all the feedback samples. We denote the embedding transform for all the feedback samples by: $f: \mathrm{X}^{\mathrm{p}} \to \mathrm{Y}^{\mathrm{p}}, \mathrm{X}^{\mathrm{n}} \to \mathrm{Y}^{\mathrm{n}}$. Therefore, after the embedding transform, in R^L, the feedback samples matrix can be represented as: $\mathrm{Y} = \{y_i\}_{i=1}^{N^p+N^n} = \{\mathrm{Y}^{\mathrm{p}}, \mathrm{Y}^{\mathrm{n}}\} \in R^{L*(N^p+N^n)}$. For simplicity, we restrict the embedding transform to be linear, which can be defined by a projection matrix $\alpha \in R^{H*L} (L \ll H)$. Then the low dimensional representation of the samples can be given as $y_i = \alpha^T x_i \in R^L$.

The separability part of GBDA is based on the DSDC, i.e.,

$$J_s(\alpha) = \arg\max_{\alpha^T\alpha=I} [J_1(\alpha) - \beta_1 * J_2(\alpha)] \tag{6}$$

where $J_1(\alpha)$ is the between-class scatter and describes the inter-class dispersion, $J_2(\alpha)$ is the within-class scatter and describes the intra-class compactness and β_1 is a tuning parameter, which reflects the trade-off between the two goals.

In BDA [10], the between-class scatter matrix is defined as the negative scatter with respect to the positive centroid matrices in the feature space. It is not reasonable to describe the separability of the negative class and the positive class except that when all the positive samples are drawn from a single Gaussian distribution which is always not the case for CBIR. Recently, to take the nonlinearity of the sample distribution into account, nonparametric models [11, pp.467-468] have been developed for discriminant analysis and achieved improvement. By reformulating the within-class scatter and the between-class scatter matrix defined in FLDA, Sugiyama proposed the local FLDA [29]. In [20], Yan et al introduced a Marginal Fisher Analysis (MFA), which characterizes the inter-class dispersion and the intra-class compactness by the sum of the distances between k-nearest inter-class neighbors and k-nearest intra-class neighbors respectively. Inspired by these nonparametric techniques [20, 29], we implement the inter-class dispersion by only selecting the samples pairs near the boundary, i.e.,

$$\begin{aligned} J_1(\alpha) &= \arg\max_{\alpha} \left(\tfrac{1}{N_1} \sum_{i=1}^{N^p} \sum_{j\in N(i)} \| y_i^p - y_j^n \|^2 + \tfrac{1}{N_1} \sum_{i=N^p+1}^{N} \sum_{j\in N(i)} \| y_i^n - y_j^p \|^2 \right) \\ &= \arg\max_{\alpha} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ij} \| y_i - y_j \|^2 = \arg\max_{\alpha} \{2 * trace[\mathrm{Y}(\mathrm{D_U}\text{-U})\mathrm{Y}^{\mathrm{T}}]\} \\ &= \arg\max_{\alpha} trace[\alpha^T \mathrm{X}(\mathrm{D_U}\text{-U})\mathrm{X}^{\mathrm{T}}\alpha] \end{aligned} \tag{7}$$

where N_1 is the total number of k nearest inter-class sample pairs between the positive class and the negative class, and $N(i)$ is a set of indices of samples that are k nearest sample pairs among different classes for each sample x_i; u_{ij} is a weighting coefficient, which is defined as follows:

$$u_{ij} = \begin{cases} \frac{1}{N_1}, if\ 1 \le i \le N^p\ and\ N^p+1 \le j \le N, j \in N(i) or\ i \in N(j) \\ \frac{1}{N_1}, if\ N^p+1 \le i \le N,\ and\ 1 \le j \le N^p, j \in N(i) or\ i \in N(j) \\ 0,\ else \end{cases} \tag{8}$$

where $U = \{u_{ij}\} \in R^{N*N}$ is a symmetric matrix, and its entry is the weighting coefficient u_{ij}; $D_U \in R^{N*N}$ is a diagonal matrix and its *i-th* entry is $\sum_{j=1}^{N} u_{ij}$. Based on the definition of u_{ij}, we can see that the weighting coefficient encodes both the sample label information and the neighborhood relationship in the high dimensional space. All of these selected inter-class sample pairs are used to capture the discriminative information between different classes.

To implement the intra-class compactness, similar to the original BDA, the definition of the intra-class compactness should also only bias towards positive samples in the positive class. Therefore, we only preserve the positive class compactness for the intra-class compactness, i.e.,

$$\begin{aligned} J_2(\alpha) &= \arg\min_{\alpha} \tfrac{1}{N_2} \sum_{i=1}^{N^p} \sum_{j=1}^{N^p} \| y_i^p - y_j^p \|^2 = \arg\min_{\alpha} \sum_{i=1}^{N^p} \sum_{j=1}^{N^p} v_{ij} \| y_i^p - y_j^p \|^2 \\ &= \arg\min_{\alpha} \{2 * trace[\mathrm{Y}^{\mathrm{p}}(\mathrm{D_v}\text{-V})\mathrm{Y}^{\mathrm{pT}}]\} \\ &= \arg\min_{\alpha} trace[\alpha^T \mathrm{X}^{\mathrm{p}}(\mathrm{D_v}\text{-V})\mathrm{X}^{\mathrm{pT}}\alpha] \end{aligned} \tag{9}$$

where N_2 is the total number of pairs of samples in the positive class, and v_{ij} is the weighting coefficient, which is defined as follows:

$$v_{ij} = \begin{cases} \frac{1}{N_2}, if\ 1 \le i, j \le N^p, i \ne j \\ 0,\ if\ 1 \le i, j \le N^p, i = j \end{cases} \tag{10}$$

where $V = \{v_{ij}\} \in R^{N^p*N^p}$ is a symmetric matrix; $\mathrm{D_V}$ is a diagonal matrix and its *i-th* entry is $\sum_{j=1}^{N} v_{ij}$.

To reduce the risk of overfitting, we introduce the notion of local consistency into Equation (6) to regularize the objective of separability which was emerging from manifold learning community [12-14, 30, 31]. Recently, large numbers of nonlinear or linear techniques have been proposed to discover the intrinsic manifold structure of the samples in high dimensional space. For example, the Laplacian Eigenmaps algorithm [30] preserves the similarities among neighboring samples. These approaches yield impressive results both on benchmark artificial data sets, as well as real world data sets. Therefore, it is reasonable to expect that integrating the essential manifold structure of the positive samples will further improve the performance of the biased discriminant analysis RF for CBIR.

We choose the Locality Preserving Projection (LPP) [31], which is the direct linearization of Laplacian Eigenmaps [30], to regularize the separability between different classes. The LPP implements the local consistency principle by preserving the similarity among the neighboring samples and is widely used in face recognition [32] and image retrieval [33]. The local consistency for the positive samples can be defined as follows:

$$\begin{aligned} J_3(\alpha) &= \arg\min_{\alpha} \tfrac{1}{N_3} \sum_{i=1}^{N^p} \sum_{j\in S(i)} \| y_i^p - y_j^p \|^2 \omega_{ij} = \arg\min_{\alpha} \sum_{i=1}^{N^p} \sum_{j\in S(i)} \| y_i^p - y_j^p \|^2 w_{ij} \\ &= \arg\min_{\alpha} \{2 * trace[\mathrm{Y}^{\mathrm{p}}(\mathrm{D_w}\text{-W})\mathrm{Y}^{\mathrm{pT}}]\} \\ &= \arg\min_{\alpha} trace[\mathrm{Y}^{\mathrm{p}}(\mathrm{D_w}\text{-W})\mathrm{Y}^{\mathrm{pT}}] \\ &= \arg\min_{\alpha} trace[\alpha^T \mathrm{X}^{\mathrm{p}}(\mathrm{D_w}\text{-W})\mathrm{X}^{\mathrm{pT}}\alpha] \end{aligned} \tag{11}$$

where $\omega_{ij} = \exp(-\| x_i^p - x_j^p \|^2 / \delta^2)$ is the heat kernel according to Laplacian Eigenmaps [30] and Locality Preserving Projection [31], which reflects the affinity of the sample pairs; $S(i)$ is the set of indices of the neighboring samples in the positive class for the positive sample x_i; N_3 is the total number of k nearest positive samples for all the positive samples; the weighting coefficient w_{ij} can be defined as follows:

$$w_{ij} = \begin{cases} \frac{\omega_{ij}}{N_3}, if\ 1 \le i, j \le N^p, j \in S(i)\ or\ i \in S(j) \\ 0, else \end{cases} \quad (12)$$

where $W = \{w_{ij}\} \in R^{N^p * N^p}$ is a symmetric matrix, which reflects the local geometry of the positive samples in high dimensional space; and D_w is a diagonal matrix and its *i-th* entry is $\sum_{j=1}^{N} w_{ij}$.

According to [31], a definition in Equation (11) corresponds to the approximation of $\int_M \| \nabla f(x) \|^2$, the manifold on which the positive samples reside. Minimizing the objective function can encourage the consistent output for the positive samples in the high dimensional space and this will result in transforming with high local smoothness and best local preservation. Hence, a smooth transform that is expected to be less likely to over fit the training samples can be learnt by this manifold regularization.

To sum up, the GBDA algorithm can be formulated by combining the above two terms ($J_s(\alpha)$ and $J_3(\alpha)$) together, as shown in Equation (13).

$$\begin{aligned} J(\alpha) &= J_s(\alpha) - \beta_2 * J_3(\alpha) = J_1(\alpha) - \beta_1 * J_2(\alpha) - \beta_2 * J_3(\alpha) \\ &= \arg\max_{\alpha} \sum_{i=1}^{N}\sum_{i=1}^{N} u_{ij} \| y_i - y_j \|^2 - \beta_1 * \sum_{i=1}^{N^p}\sum_{i=1}^{N^p} v_{ij} \| y_i^p - y_j^p \|^2 - \beta_2 * \sum_{i=1}^{N^p}\sum_{j \in S(i)} w_{ij} \| y_i^p - y_j^p \|^2 \\ &= \arg\max_{\alpha} trace\{\alpha^T [X(D_U - U)X^T - \beta_1 * X^p (D_V - V)X^{p^T} - \beta_2 * X^p (D_W - W)X^{p^T}]\alpha\} \end{aligned} \quad (13)$$

where β_2 is the regularization coefficient controlling the trade-off between the two objectives. i.e., the separability and the local consistency. By imposing the constraint $\alpha^T \alpha = I$ on Equation (13), the optimal solution can be calculated by generalized eigenvalue decomposition and the low dimensional space is spanned by the L eigenvectors α associated with the L largest eigenvalues.

We empirically set the value of β_1 for RF in CBIR based on experiments, since the optimal β_1 may not be the best for classification. For the elements u_{ij} of the between-class scatter, we have normalized them by setting $u_{ij} = 1/N_1$; for the elements v_{ij} of the within-class scatter, we have also normalized them by setting $v_{ij} = 1/N_2$. Therefore, from the view point of normalization, the two terms in $J_s(\alpha)$ are actually balanced, i.e., we can set $\beta_1 = 1$ in RF for simplicity. The value of β_2 is used to trade off the separability and the local consistency. However, it is still an open question that how to tune the regularization coefficient and balance the two objectives. Intuitionally, a larger value of β_2 will result in a solution that can enlarge $J_s(\alpha)$ and diminish $J_3(\alpha)$, and therefore it will lead to enhance the separability and encourage the local consistency. In the following experiments, we present the sensitivity of GBDA in relation to the parameter β_2 and then select the value that shows the best performance.

IV. Experimental results

We have implemented an image retrieval system on a Corel Image Database that includes 10763 images with 80 different concepts [16, 34]. To represent images, we choose three types of low-level visual features. For color, we utilize the color histogram [35] to represent the color information. We quantized hue and saturation into 8 bins and value into 4 bins. We use Webber's Law Descriptors [36] to represent the local features of images, which result in a feature vector of 240 values. For shape, the edge directional histogram from the Y component in YCrCb space is adopted to capture the spatial distribution of edges [37]. Five categories including horizontal, 45° diagonal, vertical, 135° diagonal and isotropic directions are calculated to form shape features. All of these features are combined into a feature vector, which results in a vector with 510 values (i.e., 8*8*4+9+240+5=510). Then all feature components are normalized to distributions with zero mean and one standard deviation to represent images.

In experiments, 500 query samples are randomly selected from this image database and then RF is automatically implemented by the system. In some initial experiments, we note that the number of relevant images (i.e., images with the same concept as the query image) at each iteration may range from 0 to the number of images displayed to the user and the number of irrelevant images (i.e., images with the different concepts with the query image) may range from the number of images displayed to the user to 0. Therefore, we design the following feedback procedure: at each feedback iteration, the top 20 images resulting from the resorted results are serially examined from the top; the first 5 query relevant images are labeled as positive feedbacks and the first 5 query irrelevant images are marked as negative feedbacks unless fewer such images are found among the top 20 images, in which case the fewer number of samples found are used as the feedbacks. Note that, the images which have been selected in the previous iterations are excluded from later selections. All the labeled images in the feedback iterations are used to train an RF model.

We use average precision, standard deviation and average recall to evaluate the performance of RF algorithms. The average precision refers to the percentage of relevant images in top retrieved images and is calculated as the averaged precision values of all the queries to evaluate the effectiveness of the algorithm. The standard deviation can indicate the stability of different algorithms, and is calculated for all the query precision values to describe the robustness of the algorithm. Average recall shows the fraction of the related images that are successfully retrieved and is defined as the percentage of retrieved images among all relevant images in the data set.

A. Sensitivity in Relation to Parameters

In order to select a proper quantity of *k*-nearest inter-class sample pairs to describe the discriminative information, we first show the performance comparison of GBDA on different

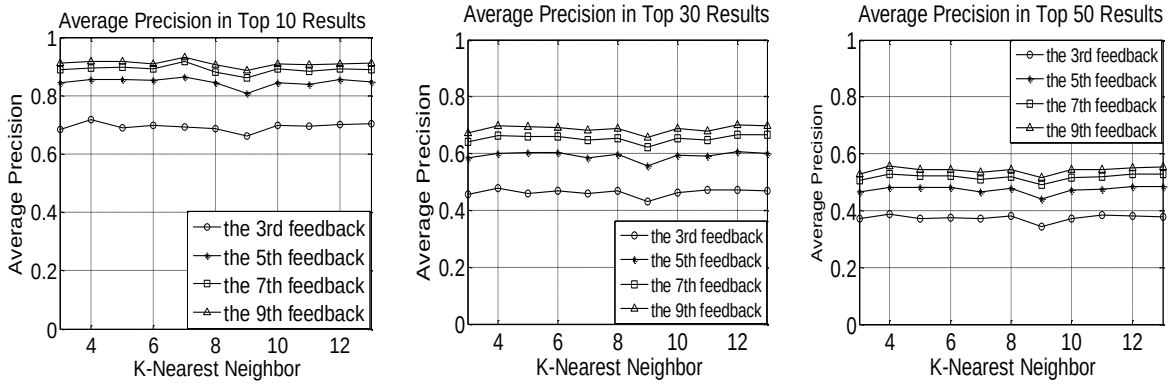
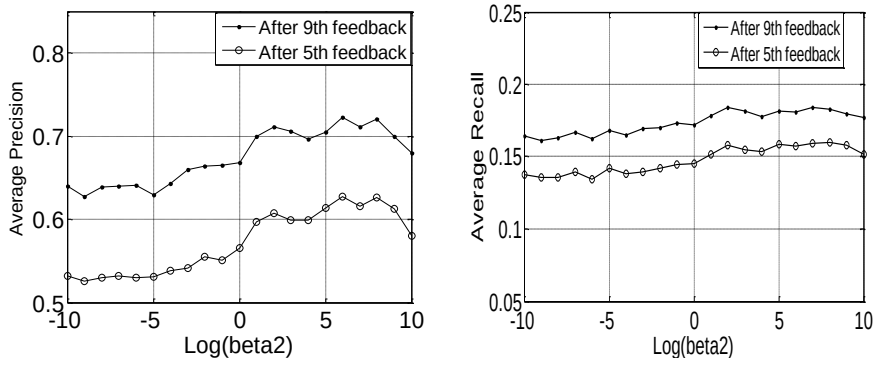


Fig.1 The average precision for different value of *k* inter-class samples



(a) (b)

Fig.2 Performance evaluation at the top 30 reranked images for different value of β_2 (a) Average precision. (b) Average recall

quantity of *k*-nearest inter-class samples pairs. Fig. 1 shows the top 10, 30, and 50 retrieved results of *3-th*, *5-th*, *7-th*, *9-th* feedback iterations with different *k* values from 3 to 13 based on 400 independent experiments.

As we can see from Fig. 3, the precision curves change slightly for different *k* values. Even a small number of inter-class samples pairs can capture the discriminative information well and obtain good performance with regard to the average precision. Therefore, in the following section we select *k*=4 in all the following experiments.

Then we show the sensitivity of GBDA with regard to different values of the parameter β_2 and empirically set the parameter β_2 as a value in a sequence, i.e., $\{2^i, i = -10, -9, \ldots, 9, 10\}$. Fig.2 shows the average precision and average recall curves in top 30 results of the 5-th and the 9-th feedback iterations with different β_2 values based on 500 independent experiments respectively. In experiments, we find that the parameter β_2 significantly affects the results. As shown in Fig. 4, we can see that when β_2 is small enough, that is, the local consistency contributes little to the formulation, the performance degrades significantly. When β_2 become larger, the GBDA algorithm shows much better performance regarding to the precision and recall. However, if β_2 is too large, the performance may degenerate. This is mainly due to the over-smoothing. From the results, we can see that for this problem, the algorithm achieves best performance when β_2 is set as 2^6. Therefore, in the following experiments, we empirically set the tradeoff parameter $\beta_2 = 2^6$. It is convinced that the parameter β_2 can be further tuned to achieve better performance. This analysis above also indicates the important role of local consistency for improving the generalization ability.

B. Experimental Results

In this subsection, we focus on the comparison of the proposed GBDA with the original BDA [10] and some of its variants, namely, the enhanced DBDA [16], Null-space BDA (NBDA) [27], and MBA [21], all of which are linear embedding algorithms and obtain much better performance comparing the original BDA. Simultaneously, SVM based algorithms including SVM [6] and CSVM [38] are also compared to evaluate the performance of GBDA. BDA, MBA, DBDA and NBDA are all based on FDC.

For MBA, we empirically set the within-class compactness parameter k_1 according to the LPP and between-class separability parameter $k_2 = 4$. Due to the high-dimensional features, both the original BDA and MBA will encounter the singular problem, and hence in the experiments that follow, regularization method is used to solve the singular problem. The enhanced version of the original BDA, DBDA[11] are solved by the direct method, which first removes the null space of the negative scatter with respect to the positive centroid matrix and then the eigenvectors of the positive within class matrix corresponding to the smallest eigenvalues are extracted as the most discriminative directions. We choose the Gaussian kernel for SVM and CSVM because it achieves the best

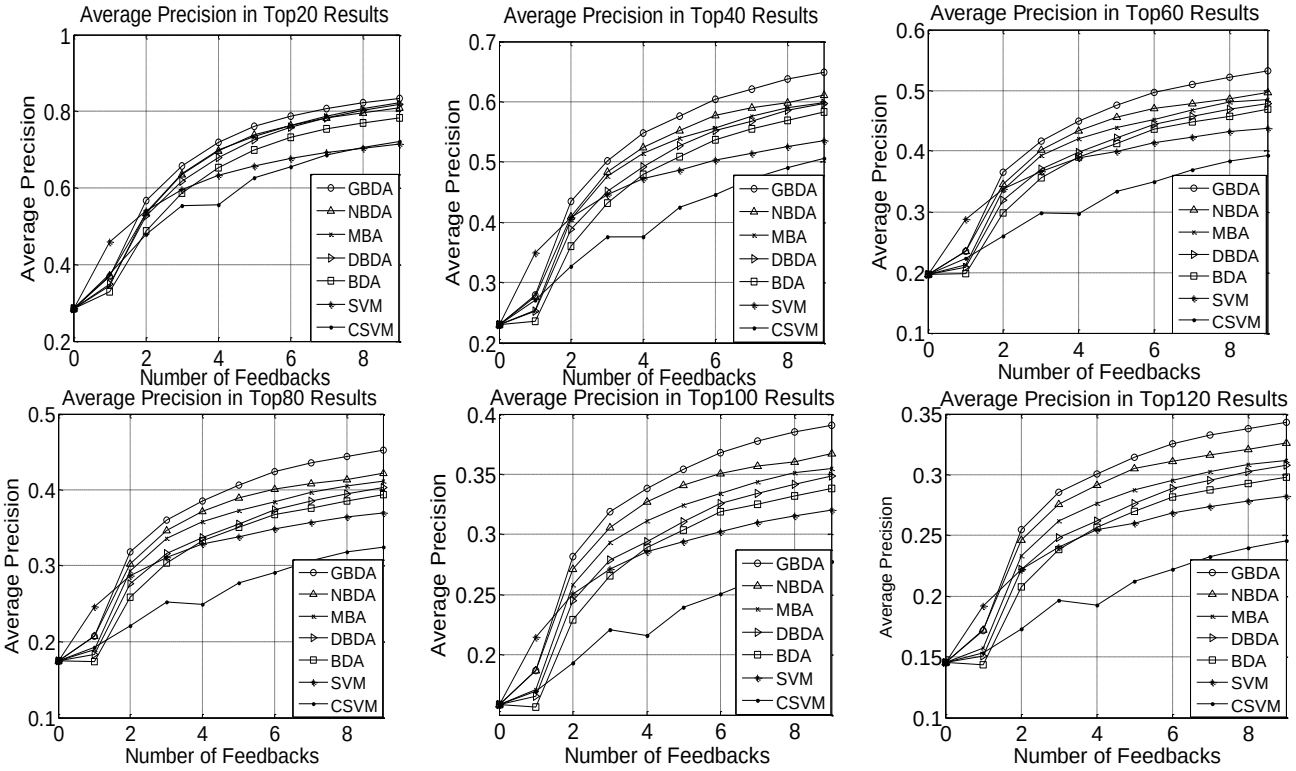


Fig. 3 Average precision of GBDA, BDA, MBA, DBDA, NBDA, SVM and CSVM for 500 independent runs

performance for all kernel-based algorithms with different parameters. For all SVM based approaches, we use the OSU-SVM [39] to implement the classification. All of the parameters are set identically as the description in the corresponding papers [10, 16, 21].

As can be seen in Fig.3, the proposed GBDA algorithm consistently outperforms the BDA, MBA, DBDA, NBDA, SVM and CSVM for RF. The figures in Fig.3 show the average precision of the 500 experiments for top 20, 40, 60, 80, 100 and 120 results. We can see that the GBDA can achieve better performance comparing with the original BDA. The unreasonable Gaussian distribution assumption for the positive samples in the original BDA and the regularization method used are the main reasons which usually result in poor performance of RF for CBIR. The DBDA algorithm solves the singular problem by the direct method and can achieve better performance than the original BDA. However, much discriminative information contained in the null space of S_b is discarded. The MBA algorithm effectively extracts the discriminative information from the marginal samples; however, it still suffers from the singular problem, which causes serious stability problems for MBA.

The proposed GBDA can extract the most discriminative information from the k nearest neighborhood inter-class samples, but never encounters the singular problem. Basically, the GBDA is an effective approach that can work in the whole input space rather than only in the principal space of S_b [17] or in the null space of S_w [27]. Therefore, GBDA can keep more discriminative information. By introducing the manifold regularization, a locally smooth and consistent transform can be learned which is expected to be less vulnerable to over fit the training samples as shown in Subsection A.

Table 1 Average precision of top ranked results for different algorithms after the 9-th feedback (average precision %)

Algorithm	top20	top40	top60	top80	top100	top120	top140
GBDA	**83.35**	**64.91**	**53.18**	**45.20**	**39.09**	**34.34**	**30.73**
NBDA	80.98	61.09	49.65	42.22	36.70	32.61	29.39
MBA	82.23	59.88	48.50	41.13	35.46	31.19	27.83
DBDA	81.86	59.73	47.84	40.27	34.87	30.08	27.64
BDA	78.33	58.26	46.87	39.33	33.79	29.77	26.74
SVM	71.50	53.48	43.79	36.98	32.01	28.19	25.17
CSVM	72.08	50.53	39.33	32.51	27.77	24.54	22.03

Table 2 Average recall of top ranked results for different algorithms after the 9-th feedback (average recall %)

Algorithm	top20	top40	top60	top80	top100	top120	top140
GBDA	14.18	**21.88**	**26.77**	**30.21**	**32.47**	**33.93**	**35.27**
NBDA	13.74	20.46	24.73	27.84	30.02	31.81	33.17
MBA	**14.19**	20.69	25.23	28.25	30.44	31.86	33.07
DBDA	13.48	19.44	23.14	25.78	27.68	29.18	30.43
BDA	13.38	18.54	22.78	26.25	28.11	29.60	30.81
SVM	12.42	18.54	22.78	25.44	27.32	28.69	29.79
CSVM	12.64	17.43	20.16	22.19	23.50	24.73	25.79

The standard deviation corresponding to the six approaches in top 20, 40, 60, 80, 100 and 120 for the 500 experiments are shown in Fig. 4. For top 20 results, the GBDA algorithm is much more stable and effective than other approaches. Then for other top results, the standard deviation of GBDA is similar to the other discriminant analysis based algorithms.

We give the average precision-scope curves and the average recall-scope curves after the 9-th feedback in Fig.5. The horizontal axis is the number of top ranked results used in evaluation and the vertical axis is the average precision and average recall measured at the top ranked images. As we can see from Fig.5, it is evident that GBDA substantially outperform the original BDA algorithm and other approaches. The detailed results of average precision and average recall of

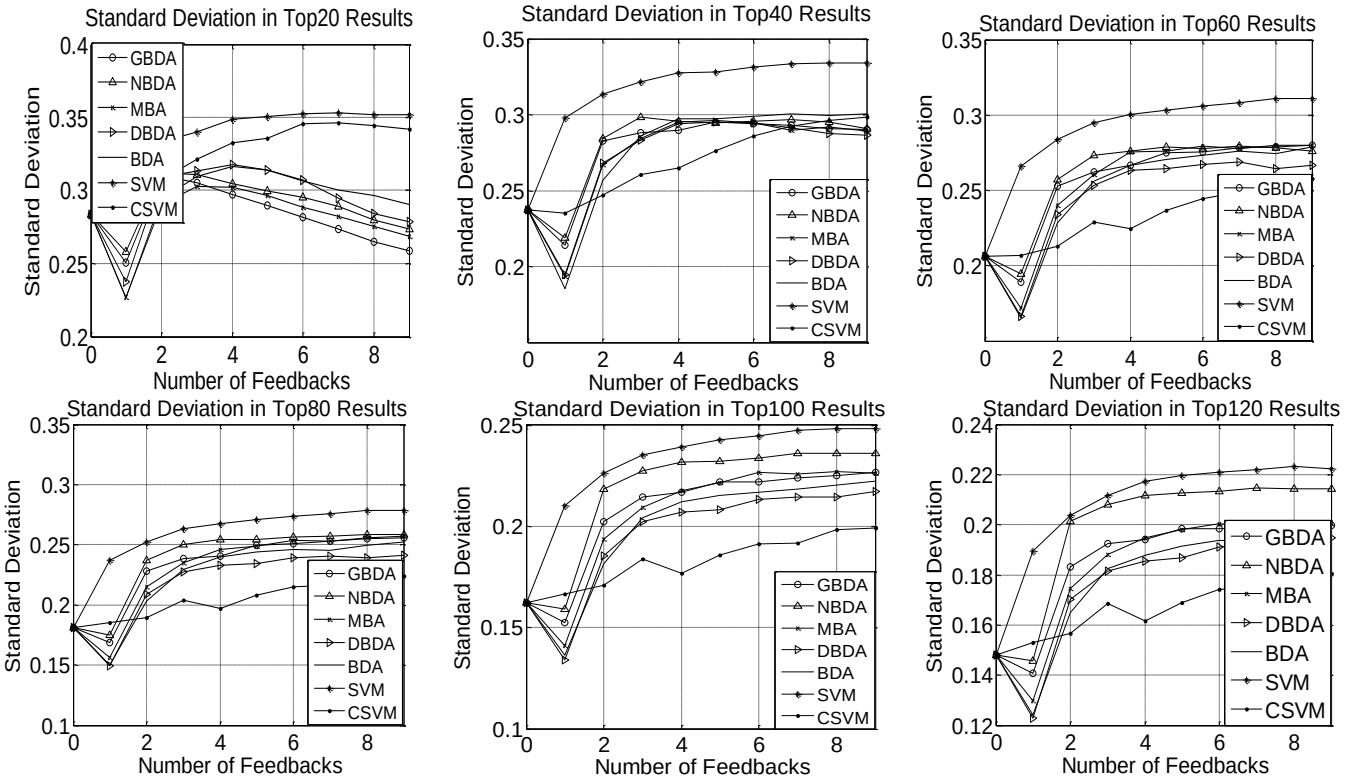
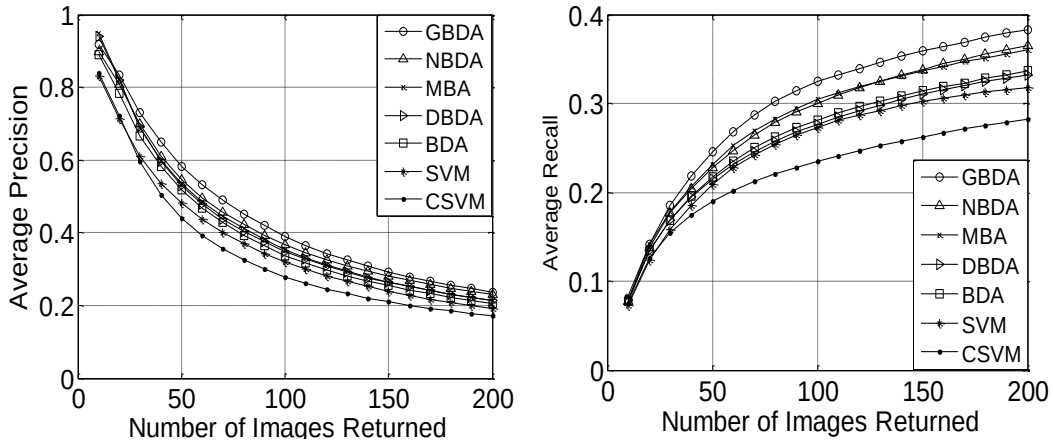


Fig.4 Standard deviation of GBDA, BDA, MBA, DBDA, NBDA, SVM and CSVM for 500 independent runs



(a) (b)

Fig.5 Performance evaluation for different algorithms after the 9-th feedback. (a) average precision (b) average recall

top ranked results for different algorithms after the 9-th feedback in Table 1 and Table 2. According to the two tables, we can observe that GBDA is more tolerant and achieves more promising results comparing with other approaches when the retrieval process is stopped. Specifically, the GBDA enjoys almost all the best performances except the average recall in top 20 results. Based on the above observation, we can empirically conclude that GBDA is more effective than the other compared algorithms in experiments.

C. Discussions and future work

In general, for CBIR, biased discriminant analysis algorithms perform much better than the traditional discriminant analysis algorithms, e.g., FLDA. This is mainly because when the user provides a query image, he/she would like to get more conceptually related images which all share a common concept, but never cares about the irrelevant images, all of which differ in various concepts.

Basically, devising a reasonable similarity metric, e.g., Mahalanobis Distance and Neighborhood Counting Measure [40], plays an important role for an image retrieval task. Different from previous work [7, 8, 9, 10], in this study, we have shown that by alleviating the unstable numerical computation problem (i.e., the singular problem) and unreasonable model (i.e., the Gaussian distribution) in a metric learning algorithm (i.e., BDA), the performance of system can be significantly improved for an image retrieval task.

Recently, the contextual and semantic information (e.g., user feedback log data [41] and image tags [42]) has shown its

potential in improving the performance of a CBIR system. A promising approach is to combine the low-level visual features and the high-level semantics to improve both speed and precision of the CBIR system.

Several aspects can be improved regarding to the RF algorithm and image representations. For example, the kernel machines can be incorporated to enhance the performance of GBDA as in the previous work [9, 10, 16, 21]; newly proposed features may outperform the traditional ones, e.g., a sparse coding representation [43].

V. CONCLUSION

To avoid the intrinsic problems (i.e., the singular problem of the positive within class scatter and the Gaussian distribution assumption for the positive samples) in the original Biased Discriminant Analysis (BDA) [10], this paper introduces a Generalized Biased Discriminant Analysis (GBDA) approach for CBIR, which is mainly based on the Differential Scatter Discriminant Criterion (DSDC) [11, 22]. The GBDA algorithm defines the separation of different classes as a trace *difference* rather than a trace *ratio*, which can avoid the singular problem of the positive within-class scatter in the original BDA. To avoid the Gaussian assumption for the positive samples, the GBDA defines the between-class scatter by resorting to inter-class nearest neighborhood samples, thereby extracting the most discriminative information. By integrating the manifold regularization, a smooth and locally consistent transform can also be learnt for CBIR RF to effectively reduce the risk of over fitting. Extensive experiments on a large Corel Image Database of 10,763 images with 80 semantic concepts have shown that the proposed GBDA significantly outperforms the original BDA, its enhanced versions (namely, DBDA, NBDA and MBA), as well as SVM and CSVM.

ACKNOWLEDGMENT

The authors would like to thank Prof. James Z. Wang (with the College of Information Sciences and Technology in Pennsylvania State University) for his kindly providing the Corel Image Gallery. The authors would like to thank the handling Associate Editor and three anonymous reviewers for their constructive comments on this manuscript in the three rounds of review.

REFERENCES

[1] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool in interactive content-based image retrieval", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 8, No. 5, pp. 644-655, Sep., 1998.

[2] X. Zhou and T. Huang, "Relevance feedback for image retrieval: a comprehensive review", *ACM Multimedia Systems J.*, Vol. 8, pp. 536-544, 2003.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 1, pp. 1349–1380, Jan.2000.

[4] R. Datta, D. Joshi, Jia Li, J.Z. Wang, "Image retrieval: ideas, influences, and trends of the new age". *ACM Computing Surveys*, Vol. 40, No.2, pp.1–60, Apr., 2008.

[5] Y. Chen, X.-S. Zhou, and T.-S. Huang, One-class SVM for Learning in Image Retrieval, *In Proceedings of IEEE Int. Conf. Image Processing (ICIP'01)*, ,pp. 815–818, 2001

[6] P. Hong, Q. Tian, and T. S. Huang. Incorporate support vector machines to content-based image retrieval with relevant feedback. In *Proc. IEEE International Conference on Image Processing (ICIP'00)*, Vancouver,BC, Canada, 2000.

[7] D. Tao, X. Li, and S. Maybank, "Negative samples analysis in relevance feedback," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 4, pp. 568–580,Apr. 2007.

[8] W. Bian and D. Tao, "Biased Discriminant Euclidean Embedding for Content-Based Image Retrieval", *IEEE Trans. On Image Processing* Vol.19, No,2, pp.545-554, Feb. 2010.

[9] M.R. Azimi-Sadjadi, J. Salazar and S. Srinivasan, "An Adaptable Image Retrieval System with Relevance Feedback Using Kernel Machines and Selective Sampling", *IEEE Trans. On Image Processing* Vol.18, No,7, pp.1645-1658, Jul. 2009.

[10] X. Zhou and T. Huang, "Small sample learning during multimedia retrieval using biasmap", *In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp.11-17,2001.

[11] K. Fukunnaga, "Introduction to statistical pattern recognition", *Academic Press*, second edition, 1991.

[12] Sam T. Roweis, and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, Vol.290, No.22, pp.2323-2326, Dec. 2000.

[13] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, Vol. 290, No. 22, pp. 2319-2323, Dec. 2000.

[14] H. Sebastian Seung and Daniel D. Lee, "The manifold ways of perception", *Science*, Vol. 290, No. 22 Dec. 2000.

[15] J. Yu, Q. Tian, T. Rui, and T.S. Huang, "Integrating discriminant and descriptive information for dimension reduction and classification", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 17, No.3, pp. 372-377, Mar. 2007.

[16] D. Tao, X. Tang, X. Li, and Y. Rui, "Kernel direct biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm", *IEEE Trans. Multimedia*, Vol.8, No.4, pp.716-724, Aug. 2006.

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

[18] X. Wang and X. Tang, "Unified subspace analysis for face recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 26, No. 9, pp. 1222-1227, Jan.2004.

[19] H. Gao, J. W. Davis, "Why direct LDA is not equivalent to LDA", *Pattern Recognition*, Vol. 39, No. 5, 1002-1006, May.2006.

[20] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 1, pp.40–51, Jan. 2007.

[21] D. Xu, S. Yan, D. Tao, S. Lin, H. Zhang, "Marginal fisher analysis and Its variants for human gait recognition and content-based image retrieval", *IEEE Trans. On Image Processing* Vol.16, No,11, pp.2811-2821, Nov. 2007.

[22] D. Tao, X. Li, X. Wu, S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 10, pp.1700-1714, Oct. 2007.

[23] F. Song, D. Zhang, D. Mei, Z. Guo, "A multiple maximum scatter difference discriminant criterion for facial feature extraction", *IEEE Trans. Syst., Man, Cybern- B, Cybern.*, Vol.37,No.6, pp.1599-1606, Dec. 2007.

[24] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion", *IEEE Trans. Neural Network.*, Vol. 17, No. 1, pp. 157–165, Jan. 2006.

[25] Q. Liu, X. Tang, H. Lu and S. Ma, "Face recognition using kernel scatter- difference based discriminant analysis", *IEEE Trans. Neural Network*, Vol. 17, No.4, pp. 1081-1085, July, 2006.

[26] X. Li, S. Lin, S. Yan, D. Xu, "Discriminant locally linear embedding With high-order tensor data". *IEEE Trans. Syst., Man Cybern.- B, Cybern.*, Vol.38, No.2, Apr. 2008.

[27] L. 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 Recognition*, Vol. 33, No.10, pp. 1713-1726, 2000.

[28] P. N. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 19, No. 7, pp. 711-720, Jul. 1997.

[29] M. Sugiyama, "Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis," J. Machine Learning Research, vol. 8, pp. 1027-1061, 2007.
[30] M. Belkin and P. Niyogi. "Laplacian eigenmaps and spectral techniques for embedding and clustering". *In Advances in Neural Information Processing Systems*, Vol.14, 2001.
[31] X. He and P. Niyogi, "Locality preserving projections", *In Advances in Neural Information Processing Systems*, Vol. 16, 2004.
[32] X. He, S. Yan, Y.Hu, P. Niyogi and H. Zhang, "Face recognition using laplacian faces", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 27, No. 3, pp.1624–1637, Dec. 2005.
[33] K. Lu and X. He, "Image retrieval based on incremental subspace learning", *Pattern Recognition*, Vol.38, No.11, pp.2047-2054, 2005.
[34] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 23, No. 9, pp. 947-963, Sep. 2001.
[35] M. J. Swain and D. H. Ballard, "Color indexing", *International. Journal of Compututer. Vision.*, Vol.7, No. 1, pp. 11–32, 1991.
[36] J. Chen, S. Shan, G. Zhao, X. Chen, W. Gao, Matti Pietikainen, "WLD: a robust descriptor based on weber's law", *IEEE Trans. Pattern Anal. Mach. Intell.*,2009.
[37] Y. Rubner, J. Puzicha, C. Tomasi, J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture", *Computer Vision and Image Understanding*, Vol.84, pp.25-43. 2001.
[38] G. Guo, A. Jain, W. Ma, and H. Zhang, "Learning similarity measure for natural image retrieval with relevance feedback", *IEEE Trans. Neural Network.*, Vol. 12, No. 4, pp. 811–820, Jul. 2002.
[39] http://www.ece.osu.edu/~maj/osu_svm/.
[40] H. Wang, "Nearest Neighbors by Neighborhood Counting", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 28, No. 6, pp. 942-953, Jun. 2006.
[41] C. Hoi, W. Liu, M. Lyu, and W. Ma. "Learning distance metrics with contextual constraints for image retrieval." *In Proc. Computer Vision and Pattern Recognition*, 2006.
[42] Y. Liu, D. Xu, I.W. Tsang, J. Luo, "Textual query of personal photos facilitated by large-scale web data", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 33, No. 5, pp.1022-1036, May. 2011.
[43] J. Yang, K. Yu, F. Lv, T. S. Huang, "Supervised translation-invariant sparse coding". *In Proc. Computer Vision and Pattern Recognition* pp. 3514-3524, 2010.