Open access • Proceedings Article • DOI:10.1145/500141.500163

Comparing discriminating transformations and SVM for learning during multimedia retrieval — Source link

Xiang Sean Zhou, Thomas S. Huang

Institutions: University of Illinois at Urbana–Champaign

Published on: 01 Oct 2001 - ACM Multimedia

Topics: Semi-supervised learning, Active learning (machine learning), Instance-based learning, Computational learning theory and Online machine learning

Comparing Discriminating Transformations and SVM for Learning during Multimedia Retrieval

Xiang Sean Zhou Thomas S. Huang
Beckman Institute, University of Illinois at Urbana Champaign
405 N Mathews Ave, Urbana, IL 61801
1-217-244-2960
{xzhou2, huang}@ifp.uiuc.edu

ABSTRACT
On-line learning or "relevance feedback" techniques for multimedia information retrieval have been explored from many different points of view: from early heuristic-based feature weighting schemes to recently proposed optimal learning algorithms, probabilistic/Bayesian learning algorithms, boosting techniques, discriminant-EM algorithm, support vector machine, and other kernel-based learning machines. Based on a careful examination of the problem and a detailed analysis of the existing solutions, we propose several discriminating transforms as the learning machine during the user interaction. We argue that relevance feedback problem is best represented as a *biased classification problem*, or a (1+*x*)-*class classification problem*. *Biased Discriminant Transform* (BDT) is shown to outperform all the others. A kernel form is proposed to capture non-linearity in the class distributions.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Relevance feedback; Query formulation; Retrieval models; Search Process.*

General Terms
Algorithms, Design, Experimentation, Human Factors, Theory.

Keywords
Multimedia retrieval, relevance feedback, discriminating transform, support vector machine, kernel method.

1. INTRODUCTION
The machine-aided retrieval of multimedia information—audio [32], image[8][26], or video[28][16], etc.—is achieved based on representations in the form of *descriptors* (or *feature vectors*), i.e., a set of real numbers. Two issues arise: one is the effectiveness of the representation, i.e., to what extent can the meaningful contents of the media be represented in these vectors? The other is the selection of *similarity metric* during the retrieval process. The latter is an important issue because the similarity metric dynamically depends upon the query class, which is unknown *a priori,* and can be user dependent and time varying, thus needs to be learned on-line through user interactions. In this paper, we focus our attention on the *similarity metric* issue, i.e., the on-line learning algorithms for content-based multimedia information retrieval.

The difference between content-based multimedia retrieval and traditional textual information retrieval lies in the fact that multimedia retrieval is conducted by the machine in a continuous feature space, while text or keyword-based retrieval is primarily performed in the discrete vector space of words; as a result, multimedia retrieval is inherently a nearest neighbor or a top-*k* ranking problem, while traditional keyword-based retrieval usually makes a binary "hit-or-miss" decision based on the *occurrences* of the keyword queries, although some rule-based ranking is possible.

For the purpose of quantitative analysis, we impose the assumption that the features selected in this paper possess adequate discriminating power to support the "ground-truth" classification in the user's mind. Note that in reality this is a strong assumption since it is very difficult to find a set of adequate features to represent high-level concepts and this is still an active research area. (Imagine the query for a music or video segment that "conveys excitement", or the query for a face picture that

Figure 1 "A picture is worth a thousand words": different users at different times can be interested in either the "horse silhouette", the "sunset", or the overall artistic layout…

"looks stupid"—it is hardly imaginable that robust numerical descriptors even exist for such high-level and subtle concepts in the human minds.)

1.1 The Need for On-line Learning

Even if we assume that consensus interpretation of multimedia contents can be reached among all possible users at all times (—"universal classification assumption"), learning of similarity metrics is still desirable since different scenarios call for different similarity metrics—For example, "cars" are similar to each other more or less in terms of "shape", while "sunset" images are best discriminated from others by "color". Therefore a query for "cars" should be handled in a different way from that of the query for "sunsets", emphasizing different *discriminating subspace* of the original feature space. However, under this "universal classification assumption", which may hold for some applications such as medical image databases with specific, well-defined functionalities, off-line pre-clustering or learning may be feasible or even beneficial.

But in general, on-line learning with user in the loop is indispensable, because an inherent nature of multimedia information is its varying interpretations by different users at different times. In other words, the perceptual "similarity" depends upon the application, the user, and the context of usage. A piece of music can invoke different feelings in different people at different times; and "a picture is worth a thousand words" (Figure 1). The time-varying interpretation or classification of multimedia information can only be dealt with using real-time learning algorithms.

Early CBIR systems invited the user into the loop by asking the user to provide a feature-weighting scheme for each retrieval task. This proved to be too technical and a formidable burden on the user's side. A more natural and friendlier way of getting user in the loop is to ask the user to give feedbacks regarding the relevance of the current outputs of the system. This is referred to as "relevance feedback" techniques ([19][21][12][25][29][34], etc). Though this is an idea initiated in the text retrieval field [23], it seems to work even better in multimedia domain: it is easier to tell the relevance of an image or video segment than that of a text document—it takes time to read through a document while an image reveals its content instantly.

In CBIR systems on-line learning techniques or relevance feedback algorithms have been shown to provide dramatic performance boost [19][21][12][29][34].

1.2 Problem Statement

Since different types of multimedia information can be represented in the same form of feature vectors, media type becomes transparent to the machine. In this paper, we assume that each meaningful "unit" of information is represented by one feature vector. For images, the "unit" can be the whole image, image blocks, or segmented regions; and for videos, the "unit" can be shots, frames, or key frames, depending upon the application scenarios.

In the abstraction of the feature space, each "unit" of multimedia data becomes a point. Relevance feedback becomes a supervised classification problem, or an on-line learning problem in a batch mode, but with some unique characteristics. The uniqueness lies in at least three aspects:

First, the machine needs to learn and respond in real time. In this paper, we target the multimedia information systems in which the similarity among the data points is dynamically determined by the current user for the current task. Therefore real-time response is critical.

Second, the number of training samples is very small relative to the dimension of the feature space, and to the requirements by popular learning machines such as support vector machines (SVM)[30]. It should be underscored that the class densities, especially that of the negative examples, cannot be reliably modeled with such small sample size.

Third, the desired output is not necessarily a binary decision on each point, but rather a rank-ordered top-k returns. This is a less demanding task since the user actually does not care the rank or configuration of the negative points as long as they are far beyond the top-k returns. In fact algorithms targeting at binary classification is ill-fitted to this problem and performs poorly. This will be illustrated in details in subsequent sections.

In this paper, we use the phrase "relevance feedback" to denote the on-line learning process during multimedia information retrieval, based on the relevance judgments fed-back by the user. The procedure is as follows:

- Machine provides initial retrieval results, through query-by-keyword, or example, etc.;

Then, iteratively:

- User provides judgment on the current results as to whether, and to what degree, they are relevant to her/his request;
- The machine learns and tries again.

In this paper, we designate the learning task as *the learning of a discriminating subspace* from the limited number of examples provided by the user in an interactive fashion.

2. STATE OF THE ART

Among different media types, on-line learning during *image retrieval* is the most active in recent years. We give a brief review of the state-of-the-art in relevance feedback techniques in the context of image retrieval. Again, many of these techniques are directly applicable for the retrieval of other media types.

2.1 Variants

Before we get into the details of various techniques, the reader should note that under the same notion of "relevance feedback", different methods might have been developed under different assumptions or problem settings thus not comparable. The following lists some of the conceptual dimensions along which some methods greatly differ from others:

a. What is the user looking for? Some assumes the user is looking for "a particular target item" [6], while many others assume the user is looking for "similar" item to the query at hand [11][12][21][22].

b. What to feedback? Some algorithm assumes the user will give a binary feedback for positive and negative examples [29]; some only takes positive examples [12]; some takes positive and negative examples with "degree of (ir)relevance" for each [21]; some assumes the feedback is only a comparative judgment, i.e., the positive examples are not necessarily "relevant" to the target, but "more like the target than the

negative ones" [6]. The latter can be related to "query refinement" techniques in others [14].

*c. **Feature representation*** While most assume one feature vector per image/region, some extract features from image blocks [36] and use mixture models as the representation [31]. A Bayesian framework is then applicable for relevance feedback. Image local matching is possible given that meaningful local features can be differentiated in the mixture [31].

*d. **Class distribution*** Another issue is what assumption to be imposed on the target class(es). Gaussian assumption is the most common and convenient one [12]. However, recent kernel based algorithms can deal with non-linearity in an elegant way [4].

*e. **Data organization*** If a hierarchical tree structure is adopted in the database for more efficient access [1], the learning becomes more difficult since the tree-structure needs to be updated in real time. The trade-off offered by [1] between the speed and accuracy in searching becomes crucial.

*f. **What to learn and how?*** A majority of the work proposes to learn a new query and the relative importance of different features[18][21][24], with some tries to learn a linear transformation in the feature space either with or without considering correlations among feature components [12][22] [21]. While others treat it either as a learning[31][34], classification [29][38], or a density estimation [4][14] problem. In the following section we discuss some of the major developments in relevance feedback techniques.

2.2 Developments

In its short history, relevance feedback developed along the path from heuristic based techniques to optimal learning algorithms, with early work inspired by term-weighting and relevance feedback techniques in document retrieval[23]. These methods proposed heuristic formulation with empirical parameter adjustment, mainly along the line of independent axis weighting in the feature space[19][21][18][20][24]. The intuition is to emphasize more on the feature(s) that best clusters the positive examples and separates the positive and the negative.

Early works [19][21] have clear birthmarks from document retrieval field. For example, In [21], learning based on "term frequency" and "inverse document frequency" in text domain is transformed into learning based on the ranks of the positive and negative images along each feature axis in the continuous feature space. [19] quantizes the features and then groups the images or regions into hierarchical trees whose nodes are constructed through single-link clustering. Then weighting on groupings is based on "set operations".

Some use Kohonen's Learning Vector Quantization (LVQ) algorithm [33] or Self-organizing Map (SOM) [13] for dynamic data clustering during relevance feedback. Laaksonen et al. [13] uses TS(Tree-Structured)-SOMs to index the images along different features. Positive and negative examples are mapped to positive and negative impulses on the maps and a low-pass operation on the maps is argued to implicitly reveal the relative importance of different features because a "good" map will keep positive examples cluster while negative examples scatter away. This is based on similar intuition as that of [18], where a probabilistic method is used to capture feature relevance.

Aside from their lack of optimality claim, the assumption of feature independence imposed in these methods is also artificial, unless independent components can be effectively extracted.

Later on researchers begin to look at this problem from a more systematic point of view by formulating it into an optimization, learning, classification, or density estimation problem. In [12] and [22], based on the minimization of total distances of positive examples from the new query, the optimal solutions turn out to be the weighted average as the new query and a whitening transform of the feature space (equivalent to principle component analysis (PCA) or the use of Mahalanobis distance). Additionally, Rui and Huang [22] adopts a two-level weighting scheme to better cope with singularity issue due to the small number of training samples. To take into account the negative examples, Schettini et al. [25] updates the feature weights along each feature axis by comparing the variance of positive examples to the variance of the union of positive and negative examples.

Assuming that the user is searching for a particular target, and the feedback is in the form of "relative judgment", Cox et al. [6] proposes the stochastic comparison search as its relevance feedback algorithm.

While most CBIR systems use well-established image features such as color histogram/moments, texture, shape, and structure features, there are alternatives. Tieu and Viola [29] used more than 45,000 "highly selective features", and a boosting technique to learn a classification function in this feature space. The features were demonstrated to be sparse with high kurtosis, and were argued to be expressive for high-level semantic concepts. Weak 2-class classifiers were formulated based on Gaussian assumption for both the positive and negative (randomly chosen) examples along each feature component, independently. The strong classifier is a weighted sum of the weak classifiers as in AdaBoost [9].

In [31], Gaussian mixture model on DCT coefficients is used as image representation. Then Bayesian inference is applied for image regional matching and learning over time.

Recently there are also attempts to incorporate support vector machine (SVM) into relevance feedback process [4][11]. However, SVM as a two-class classifier is not directly suitable for relevance feedback, because the training examples are far too few to be representatives of the true distributions [4]. However a kernel based one-class SVM as density estimator for positive examples has been shown to outperform the whitening transform based linear method [4].

Formulated in the transductive learning framework, D-EM algorithm[35] uses examples from the user feedback (labeled data) as well as other data points (unlabeled data). It performs discriminant analysis inside the EM iterations to select a subspace of features, such that the two-class assumption on the data distributions has better support. However, the computation induced by the D-EM iterations is expensive, which can make real-time implementation difficult based on the current hardware capabilities.

3. TRADITIONAL DISCRIMINANT ANALYSIS AND TRANSFORMATIONS

To effectively compare the nature and merits of the algorithms presented in Section 2, it is desirable to analyze them from the

feature space transformation point of view: indeed, the feature-weighting scheme ([13][18][25], etc.) is the simplified diagonal form of a linear transformation in the original feature space, assuming feature independence. While the Mahalanobis distance or the generalized Euclidean distance using the inverse of the covariance matrix of the positive examples [12][22] is a whitening transformation based on the configuration of the positive examples, assuming Gaussian distribution.

From pattern classification point of view, when only positive examples are to be considered and with Gaussian assumption, the whitening transformation is the optimal choice [7]. When both positive and negative examples are to be considered, instead of the aforementioned various, seemingly plausible heuristics for feature-weighting[13][18][25], two optimal linear transformations based on the traditional discriminant analysis are worth investigating. Of course "optimality" depends on the choice of the objective function; in this sense, it becomes a problem of formulating the best objective function:

3.1 Two-class Assumption

One is the two-class fisher discriminant analysis (FDA). The goal is to find a lower dimensional space in which the ratio of between-class scatter over within-class scatter is maximized.

$$W = \arg\max_{W} \frac{\left|W^T S_b W\right|}{\left|W^T S_w W\right|} \tag{1}$$

where

$$S_b = (m_x - m)(m_x - m)^T + (m_y - m)(m_y - m)^T \tag{2}$$

$$S_w = \sum_{i=1}^{N_x} (x_i - m_x)(x_i - m_x)^T + \sum_{i=1}^{N_y} (y_i - m_y)(y_i - m_y)^T \tag{3}$$

And, we use $\{x_i,\ i = 1, \ldots, N_x\}$ to denote the positive examples, and $\{y_i,\ i = 1, \ldots, N_y\}$ to denote the negative examples. m_x, m_y, and m are the mean vectors of the sets $\{x_i\}$, $\{y_i\}$, and $\{x_i\} \cup \{y_i\}$, respectively. (See [7] for details.)

For two-class discriminant analysis, it is part of the objective that negative examples shall cluster in the discriminating subspace. This is an unnecessary and potentially damaging requirement since the relatively small training sample cannot be representative for the overall population, *especially for the negative examples*. In fact, very likely the negative examples will belong to more than one class. Therefore the effort of rounding up all the negative examples can mislead the resulting discriminating subspace into the wrong direction.

3.2 Multi-class Assumption

Another choice is the multiple discriminant analysis (MDA) [7], where each negative example is treated as from a different class. It becomes a $(N_y + 1)$-class discriminant analysis problem. The reason for the crude assumption on the number of negative classes is because the class labels within the negative examples are not available. One may suggest for the user to provide this information. However, from a user interface design point of view, it is reasonable for the user to click to indicate items as relevant versus non-relevant (say, "horses" and "non-horses"), but troublesome and unnatural for the user to further identify for the machine what the negative items really are ("these are tigers, those are zebras, and that is a table, …").

For MDA the objective function has the same format as in Equation (1). The difference is in the definitions of the scatter matrices:

$$S_b = (m_x - m)(m_x - m)^T + \sum_{i=1}^{N_y} (y_i - m)(y_i - m)^T \tag{4}$$

$$S_w = \sum_{i=1}^{N_x} (x_i - m_x)(x_i - m_x)^T \tag{5}$$

In this setting, it is part of the objective that all negative examples shall be apart from one another in the discriminating subspace. This is again an unnecessary and potentially damaging requirement since several negative examples can come from the same class. The effort of splitting them up can mislead the resulting discriminating subspace into the wrong direction.

3.3 Unsupervised Clustering

Without more detailed labels on the negative examples except for the label of "negative", these two are the only sensible solutions available from the tradition discriminant analysis framework. One may argue that unsupervised clustering techniques (EM, or mean shift [5]) can be applied to find out the number of clusters automatically. However, a meaningful clustering of a set of points actually depends on the subspace selection—an image of a "red table" is not necessarily closer to a "white table" than a "red horse" unless a proper discriminating subspace can be specified in the first place—which is exactly what the system is trying to learn. Iin addition, these iterative algorithms are usually too time-consuming to achieve real time responses.

3.4 Discriminating Transformation

For both FDA and MDA, the columns of the optimal W are the generalized eigenvector(s) V associated with the largest eigenvalue(s) Λ, i.e.,

$$S_b V = \Lambda S_w V \tag{6}$$

A *discriminating transformation matrix* is defined as

$$A = V\Lambda^{1/2} \tag{7}$$

In the new space $x_{new} = A^T x_{old}$, the following "actions" are employed to ensure the optimal ratio in Equation (1)—for FDA: the positive centroid is "pushed" apart from the negative centroid, while examples of the same label are "pulled" closer to one another; for MDA: the positive centroid and every negative examples are "pushed" apart from one another, while positive examples are "pulled" closer to one another.

It is important to point out that the effective dimension of the new space is independent of the original dimensionality. For FDA, since the rank of S_b is only one, the discriminating subspace has dimension one, i.e., the transformation is always a projection onto a line. This will severely limit its ability in informative modeling, even in the kernel form. For MDA, there can be multiple non-zero eigenvalues, and the number of effective subspace dimensions is at most $\min\{N_x, N_y\}$.

4. BIASED DISCRIMINANT ANALYSIS

Instead of confining ourselves to the traditional settings of the discriminant analysis, we propose a new form of discriminant analysis, namely, biased discriminant analysis (BDA).

4.1 (1+x)-class Assumption

We first define the *(1+x)-class classification problem* or *biased classification problem* as the learning problem in which there are an unknown number of classes but the user is only interested in one class, i.e., the user is biased toward one class. And the training samples are labeled by the user as only "positive" or "negative" as to whether they belong to the target class or not. Thus the negative examples can come from an uncertain number of classes.

Much research has addressed this problem simply as a two-class classification problem with symmetric treatment on positive and negative examples, such as FDA. However the intuition is like "all happy families are alike, each unhappy family is unhappy in its own fashion"(Leo Tolstoy's *Anna Karenina*); or we say, "all positive examples are alike in a way, each negative example is negative in its own way". Therefore it is necessary to distinguish a real two-class problem from a (1+x)-class problem. When the negative examples are far from representative for their true distributions—which is certainly true in our case—this distinction becomes critical. (Tieu and Viola [29] used a random sampling strategy to increase the number of negative examples thus their representative power. This is somewhat dangerous since unlabeled positive examples can be included in these "negative" samples.)

4.2 Biased Discriminant Analysis (BDA)

For a biased classification problem, we ask the following question instead: what is the optimal discriminating subspace in which the positive examples are "pulled" closer to one another while the negative examples are "pushed" away from the positive ones?

Or mathematically, what is the optimal transformation such that the ratio of "the negative scatter with respect to positive centroid" over "the positive with-in class scatter" is maximized? We call this biased discriminant analysis (BDA) due to the biased treatment toward the positive examples. We define the biased criterion function

$$W = \arg\max_{W} \frac{\left|W^T S_y W\right|}{\left|W^T S_x W\right|} \tag{8}$$

where

$$S_y = \sum_{i=1}^{N_y} (y_i - m_x)(y_i - m_x)^T \tag{9}$$

$$S_x = \sum_{i=1}^{N_x} (x_i - m_x)(x_i - m_x)^T \tag{10}$$

The optimal solution and transformations are of the same formats as those of FDA or MDA, subject to the differences defined by Equation (9) and (10).

Note that the discriminating subspace of BDA, obtained through a transformation of the form similar to the one in Equation (7), has effective dimension of $\min\{N_x, N_y\}$, the same as MDA and higher than that of FDA. Even though the weighting of the eigenvectors by the square roots of their corresponding eigenvalues does not affect the value of the objective function in Equation (8), the resulting *biased discriminating transform* (BDT) matrix has the form of a "*generalized* whitening transform", or a "*discriminative* whitening transform". BDT can be regarded as the *informative modeling of positive examples incorporating discriminative information from negative examples.* Whitening transform is the special case when only positive examples are considered.

4.3 Regularization and Discounting Factors

It is well known that the sample-based plug-in estimates of the scatter matrices based on Equations (2)(3)(4)(5)(9)(10) will be severely biased for small number of training examples, i.e., the largest eigenvalue becomes larger, while the small ones smaller. A compensation or regularization can be done by adding small quantities to the diagonal of the scatter matrices[10]. The regularized version of S_x, with n being the dimension of the original space and I being the identity matrix, is:

$$S_x^r = (1-\mu)S_x + \frac{\mu}{n} tr[S_x] I \tag{11}$$

The parameter μ control shrinkage toward a multiple of the identity matrix. And *tr*[] denotes the trace operation for a matrix.

The influence of the negative examples can be tuned down by a discounting factor γ, and the discounted version of S_y is:

$$S_y^d = (1-\gamma)S_y + \frac{\gamma}{n} tr[S_y] I \tag{12}$$

With different combinations of the (μ, γ) values, the regularized and/or discounted BDA provides a rich set of alternatives: ($\mu = 0$, $\gamma = 1$) gives a subspace that is mainly defined by minimizing the scatters among the positive examples, resembling the effect of a whitening transform; ($\mu = 1$, $\gamma = 0$) gives a subspace that mainly separates the negative from the positive centroid, with minimal effort on clustering the positive examples; ($\mu = 0$, $\gamma = 0$) is the full BDA and ($\mu = 1$, $\gamma = 1$) represents the extreme of discounting all configurations of the training examples and keep the original feature space unchanged.

BDA captures the essential nature of the problem with minimal assumption. In fact, even the Gaussian assumption on the positive examples can be further relaxed by incorporating kernels.

5. KERNEL-BASED BIASED DISCRIMINANT ANALYSIS (KBDA)

To take into account non-linearity in the data, we propose a kernel-based approach.

The original BDA algorithm is applied in a "*feature space*"[1], which is related to the original space by a non-linear mapping $\phi : x \to \phi(x)$. Since in general the number of components in $\phi(x)$ can be very large or even infinite, this mapping is too expensive and will not be carried out explicitly, but through the evaluation of a kernel K, with elements $k_{ij} = \phi^T(x_i)\ \phi(x_j)$. This is the same idea adopted by the support vector machine[30], kernel PCA, and kernel discriminant analysis [1][15]. The trick is to rewrite the BDA formulae using only dot-products of the form $\phi_i^T \phi_j$, so that the reproducing kernel matrix can be substituted into the formulation and the solution, eliminate the need for direct non-linear transformations.

Using superscript $^\phi$ to denote quantities in the new space, we have the objective function in the following form:

$$w^* = \arg\max_{w} \frac{w^T S_y^\phi w}{w^T S_x^\phi w} \tag{13}$$

[1] A term used in kernel machine literatures to denote the new space after the nonlinear transform—this is not to be confused with the *feature space* concept previously used to denote the space for features/descriptors extracted from the media data.

where

$$S_y^{\phi} = \sum_{i=1}^{N_y} (\phi(y_i) - m_x^{\phi})(\phi(y_i) - m_x^{\phi})^T \quad (14)$$

And

$$S_x^{\phi} = \sum_{i=1}^{N_x} (\phi(x_i) - m_x^{\phi})(\phi(x_i) - m_x^{\phi})^T \quad (15)$$

Since w^* is the eigenvector(s), it can be expressed as a weighted sum of input vectors:

$$w = \sum_{i=1}^{N_x} \alpha_i \phi(x_i) + \sum_{j=1}^{N_y} \alpha_{j+N_x} \phi(y_j) = \Phi\alpha \quad (16)$$

It can be shown that the numerator of (13) can be rewritten as:

$$\begin{aligned} w^T S_y^{\phi} w &= \alpha^T \Phi^T \sum_{j=1}^{N_y} (\phi(y_j) - m_x^{\phi})(\phi(y_j) - m_x^{\phi})^T \Phi\alpha \\ &= \alpha^T \sum_{j=1}^{N_y} (K_{y_j} - K_{mx})(K_{y_j} - K_{mx})^T \alpha \\ &= \alpha^T (K_y - K_x I^y_{N_x})(K_y - K_x I^y_{N_x})^T \alpha \end{aligned} \quad (17)$$

Where Φ is defined in (16), and

$$K_{y_j} = \Phi^T \phi(y_j), \qquad K_{mx} = \Phi^T m_x^{\phi}, \qquad (K_y)_{:,j} = K_{y_j}$$

and $I^y_{N_x}$ is an N_x by N_y matrix of all elements being $1/N_x$.

Similarly, rewrite the denominator of (13),

$$\begin{aligned} w^T S_x^{\phi} w &= \alpha^T (K_x - K_x I^x_{N_x})(K_x - K_x I^x_{N_x})^T \alpha \\ &= \alpha^T K_x (I - I^x_{N_x})(I - I^x_{N_x})^T K_x^T \alpha \\ &= \alpha^T K_x (I - I^x_{N_x}) K_x^T \alpha \end{aligned} \quad (18)$$

and $I^x_{N_x}$ is an N_x by N_x matrix of all elements being $1/N_x$.

Now we can solve for α, which is the eigenvector(s) associated with the largest eigenvalue(s) for the generalized eigenanalysis problem defined by Equation (8), (17), and (18).

With optimal α's, the projection of a new pattern z onto w is given by:

$$w^T \phi(z) = \sum_{i=1}^{N_x} \alpha_i k(x_i, z) + \sum_{j=1}^{N_y} \alpha_{j+N_x} k(y_j, z) \quad (19)$$

In this nonlinearly transformed new space with w's (weighted by the square-rooted eigenvalues) as the axes, the nearest neighbors of the positive centroid are returned as the outputs of the learning process. If not satisfied, the user can give further judgments on these new outputs to enter another round of relevance feedback interaction. The machine can combine the new feedbacks with all the previous feedbacks together in the next round of learning.

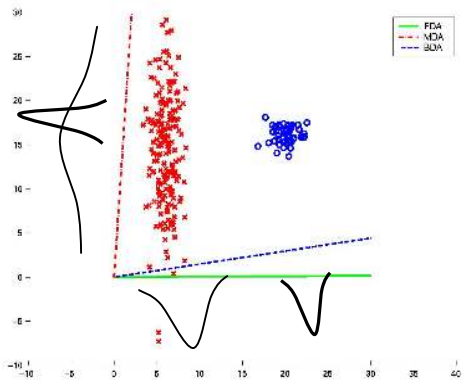
6. COMPARISONS AND ANALYSIS

Using image retrieval as the application, we compare the three proposed discriminating transforms to the optimal two level whitening transforms [22], and compare the kernel versions with SVM, on both synthetic data and real world image databases. The scenario is "query by example" followed by several rounds of relevance feedback by the user. For each round, the machine learns an optimal transform, linear or non-linear, from the training examples fed-back by the user; then all training and testing points are transformed into the new space, where the new query is the mean of the transformed positive examples, and its Euclidean nearest neighbors are returned for further feedbacks from the user.

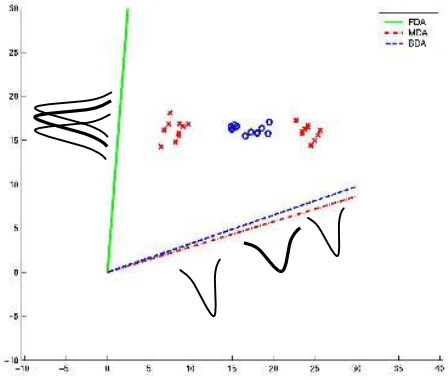
6.1 Linear/Quadratic Case

For the non-kernel versions of FDA, MDA, and BDA, all the transform matrices are linear, and the decision boundaries are either linear or quadratic.

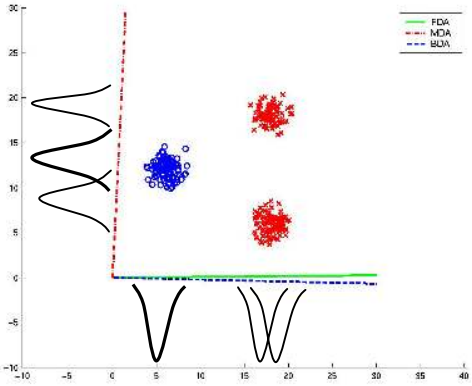
6.1.1 Toy Problems



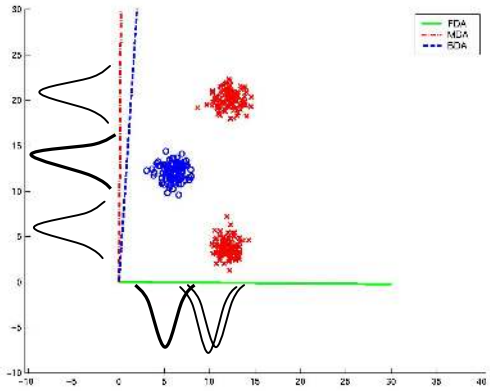
(a)



(b)



(c)



(d)

Figure 2. Comparing FDA, MDA, and BDA for dimensionality reduction from 2-D to 1-D. (a) FDA and BDA yield projection with nice class separation, MDA failed; (b) MDA and BDA yield projection with nice class separation, FDA failed; From (c) to (d), Notice the two modes of the negative examples moved apart from each other and toward the positive examples, and BDA is able to adapt to the change and gives better class separation in both cases. MDA fails in (c), and FDA fails in (d).

To illustrate the advantages of BDA over FDA or MDA, we use some toy problems as depicted in Figure 2. Original data are in 2-D feature space, and positive examples are "o"s and negetive examples are "x"s in the figure. FDA, MDA, and BDA are applied to find the best projection direction by their own criterion functions for each case, and the resulting (generalized) eigenvector corresponding to the maximum eigenvalue is drawn in solid green, dash-dotted red, and dashed blue straight lines, respectively. The would-be projections of the data points onto these eigenvectors are also draw as bell-shaped curves to the side of the corresponding eigenvectors, assuming Gaussian distribution for each mode. The thicker curves represent the projections of the positive modes.

Here, FDA treats positive and negative examples equally, i.e., it tries to decrease the scatter among negative examples as part of the effort. This makes it a bad choice in cases (b) or (d). Without any prior knowledge about the number of classes to which the negative examples belong, MDA can only treat each example as a separate class/mode. Since MDA has in its criterion function the tendency of increasing the scatter among all classes/modes, which includes the scatter among negative examples, this makes it a bad choice for cases (a) and (c).

In all cases, BDA yields good separation of negative examples from positive ones, as well as clustering of positive examples (it finds a balance between these two goals, which are embedded in the criterion function). Note from (c) to (d), the two negative modes move apart from each other and toward the positive ones. FDA and MDA cannot adapt to the changing configurations and will fail for one of the two cases: for (c) MDA fails and for (d) FDA fails. Whereas BDA is able to adapt to the change and gives better separation in both cases.

FDA and MDA are inadequate for biased classification or biased dimensionality reduction problems because of their forceful assumption on the number of modes. BDA avoids making this assumption by directly modeling the problem into the objective function hence gives better results.

Figure 3 Some example images from the Corel set used in the experiments.

Table 1 Comparing relevance feedback results: the first row is the averaged number of hits in top 20, and the second row shows their variances.

No feedback	*WT*	*FDA*	*MDA*	*BDA*
8.2	13.0	13.9	16.2	17.0
8.43	17.43	16.50	10.26	8.86

6.1.2 Image Database Testing

In this experiments, a COREL image set of 17695 images are tested. A feature space of 37 dimensions is used with 9 color moments, 10 wavelet moments [27], and 18 edge-based structure features [37]. Without prior knowledge on feature-class correlations, all feature components are normalized to normal distributions.

For the first round with only one positive example—the query image—the system uses Euclidean distance metric to retrieve the 20 nearest neighbors. Subsequently, a subject selects the training examples on the fly. Up to 20 rounds of feedback (or until convergence) are performed for every query under each of the four relevance feedback schemes: two-level optimal whitening transform (WT) [22], FDA, MDA, and BDA. Altogether over 1000 rounds of subject guided retrieval/relevance feedback are performed over 20 classes of images (See Figure 3 for some examples. It should be noted that some of the semantic classes in the Corel set are too difficult for content-based retrieval using the currently available low-level features. The ones shown here are the relatively "good examples" that can yield reasonable initial results for further user interactions). The numbers of hits in top 20 are recorded for different schemes. And their means and variances are compared in Table 1.

It is apparent that all the three proposed transforms outperform the WT scheme based solely on positive examples, especially the MDA and BDA-based transforms. BDA not only yields the highest average score, but also has the minimum variation, which indicates the most robust performance. FDA and MDA have larger performance variation because they are affected by the clustering patterns in negative examples, which are generally

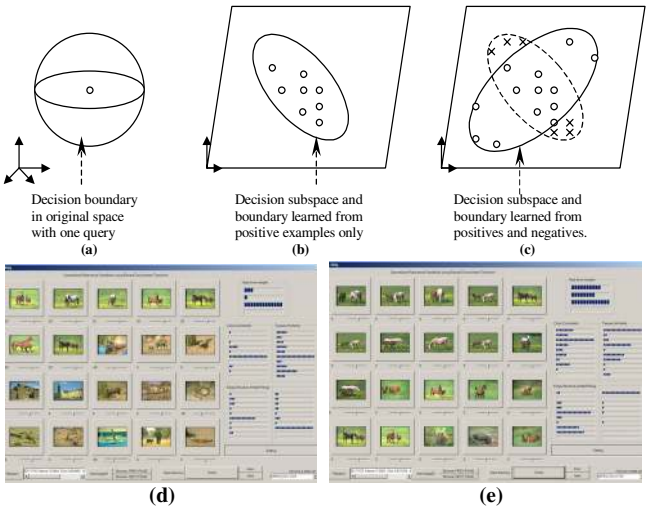


Figure 4. The open circles represent positive examples and the crosses negative. (a) the system uses Euclidean distance for one query; (b) the system uses the subspace spanned by the positive examples. It can stagnate at a local minimum; (c) adding negative examples the system finds a better transformation; (d) Top 20 returns with only positive feedback. The system stagnates at this point, repeating the same response; (e) Adding negative feedback and using BDA can pull the system out of stagnation and arrive at a much better solution.

unstable. MDA in this case is close to BDA in performance because the subject for this test tends to give small number (average around 3) of negative examples that are usually *not* from the same class, i.e., if two "tigers" appear when searching for "horses", the subject only mark one of them as negative to see whether the other one can be "pushed" out in the next round—with negative examples all coming from different classes, the problem associated with MDA can not be fully observed (See Section 3.2 for analysis.) WT has low average score and large performance variation mainly because it is prone to be trapped at local minimum, which is frequently observed in our experiments. Figure 4 illustrate this point with a hypothetic feature space configuration, as well as a real image retrieval example. It shows that using BDT the system can climb out of local minimum with the "push" from negative examples.

All the four algorithms run in real time on a Pentium III PC, with a maximum latency of less than 2 seconds during relevance feedback on the COREL set of 17695 images with a 37 dimensional feature space.

6.2 Non-linear Case

For the non-linear case, we compare the kernel BDA with BDA, and SVM, over the same RBF kernel.

6.2.1 Does Kernel Help?

To test the ability of the KBDA in dealing with non-linearly distributed positive examples, six sets of synthetic data in two-dimensional space are used (see Figure 5).

A significant boost in averaged hit rates is observed when using KBDA.

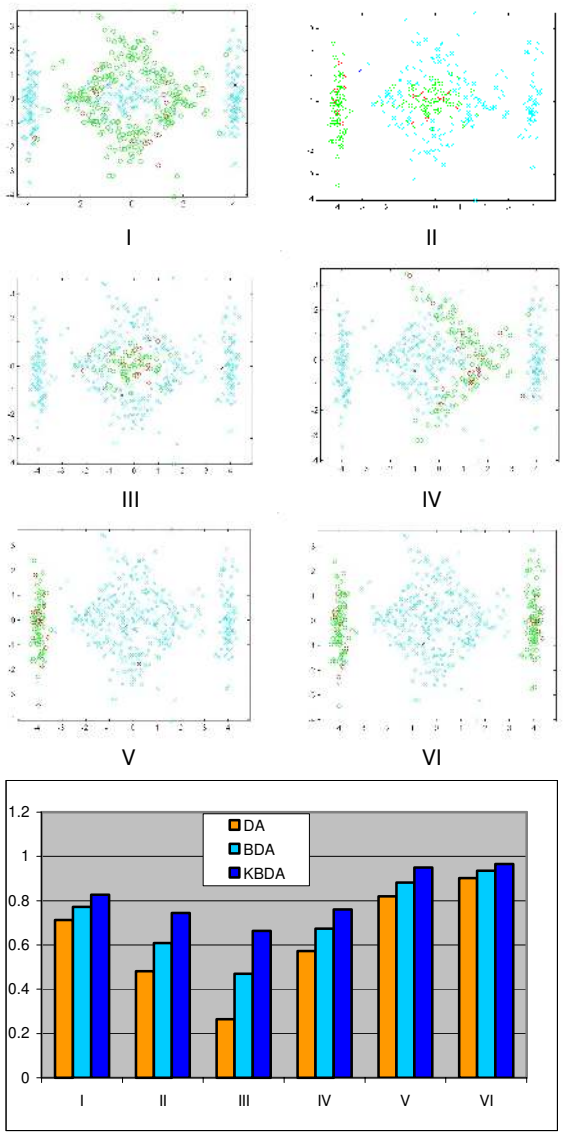


Figure 5 Test results on synthetic training data: six different configurations of non-linearity. The circles are positive examples and the crosses negative. A simulated query process is used for training sample selection, i.e., the 20 nearest neighbors of a randomly selected positive point are used as training samples. The bar diagram shows the averaged hit rate in top 20 returns.

Next we try KBDA on a real image database to see whether it helps to introduce kernel for nonlinearity in the real world applications.

A fully labeled set of 500 images from COREL is used for automated testing in the next experiment. It contains five classes, each with 100 images. Each round 10 positive and 10 negative images are randomly drawn as training samples. For each

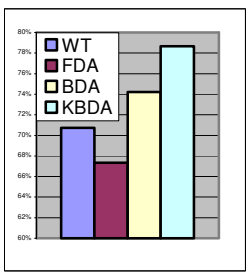


Figure 6
Averaged hit rates

round the error rate in the top 100 returns is recorded as the performance measures. 500 rounds of testing are performed on the 5 classes and the averaged hit rates are shown in Figure 6 where four schemes are compared: WT, FDA, BDA, and KBDA. One can see that KBDA outperforms BDA on average by a significant margin.

6.2.2 KBDA vs. SVM

It is also desirable to see how the proposed kernel method compares with support vector machines (SVM).

SVM assumes that negative examples are representative of the true distribution, which is far from the reality for the relevance feedback scenario. In the information retrieval application, given the usually large number of classes the unlabeled areas in the feature space are very likely to be negative. When SVM is directly implemented as a two-class learning machine during information retrieval, the result is that after the user's feedback, the machine returns a totally different set of points, with most of them likely to be negative. Of course incremental training iterations [2] can be applied to *eventually* arrive at the correct boundary, but this may require a significant number of further iterations and more training examples, from an extremely patient user.

Here we compare KBDA and SVM in the context of face and non-face classification under small number of training samples, this example shall reveal more clearly the nature of the two algorithms. Among the 1000 faces and 1000 non-face images, some examples are shown in Figure 7. All the images are 16-by-16 in size and the original pixel values are used as the features, resulting in a 256-dimensional space. We use different numbers of positive and negative examples to train a KBDA and a SVM learner. For the SVM, the distance to the hyperplane is used to rank order all the points and the percentage of face images in the top 1000 returns is used to compare with the hit rate of KBDA in its top 1000 returns.

Figure 8 illustrates the experimental results with a fixed number of positive examples—100, and a varying number of negative examples—from 1, 2, …, to 500. The vertical axis denotes the hit rate in top 1000 returns. Each point on the two curves represents the averaged rate of 100 random trials. This figure clearly shows that when the number of negative examples is small (< 200), KBDA outperforms SVM.

7. CONCLUSION

In this paper, we briefly reviewed existing relevance feedback techniques. Emphasize was put on the analysis of the unique characteristics of multimedia information retrieval problems and the corresponding on-line learning algorithms. A novel scheme was proposed with experimental results supporting its superior performance than existing schemes. The proposed KBDA algorithm can be applied not only in multimedia information retrieval problems, but also for other classification problems whenever the number of negative training samples is too small to be representative for the true distribution.

8. ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation under the grants CDA 96-24396 and EIA 99-75019. The authors would like to thank the anonymous reviews for their comments and suggestions!

Figure 7 Some examples of the face and non-face images

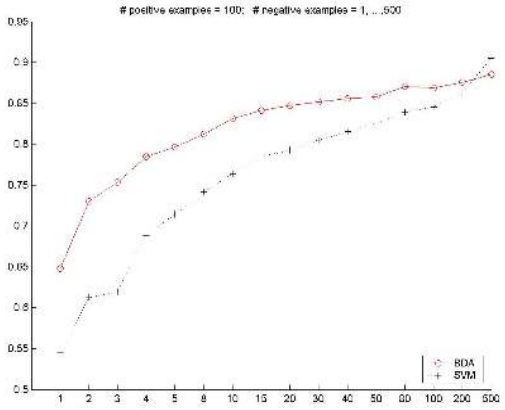


Figure 8 KBDA outperforms SVM for small number of negative examples. The number of positive examples is fixed at 100, and the horizontal axis shows the changing number of negative examples.

9. REFERENCE

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. Neural Computation, 12(10):2385--2404, 2000

[2] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning", in T. K Leen, T.G. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems 13, MIT Press, 2001

[3] J-Y. Chen, C. A. Bouman, and J. C. Dalton, "Hierarchical Browsing and Search of Large Image Databases," IEEE Trans. on Image Processing, vol. 9, no. 3, pp. 442-455, March 2000.

[4] Y. Chen, X. S. Zhou, T. S. Huang, "One-class SVM for Learning in Image Retrieval", submitted to IEEE Int'l Conf. on Image Proc. (ICIP'2001), Thessaloniki, Greece, October 7-10, 2001

[5] D. Comaniciu, P. Meer, "Distribution free decomposition of multivariate data," Pattern Analysis and Applications, 2, 22-30, 1999

[6] I. J. Cox, M. Miller, T. Minka, P. Yianilos, "An Optimized Interaction Strategy for Bayesian Relevance Feedback", IEEE Conf. Computer Vision and Pattern Recognition, Santa Barbara, CA. June 1998

[7] R.O. Duda and P.E. Hart. Pattern Classification and Scene Analysis. John Wiley & Sons, Inc., New York, 1973

[8] M. Flickner, et al., “Query by image and video content: The qbic system”, IEEE Computers. 1995

[9] Y. Freund and R. E. Schapire. “A short introduction to boosting”, Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, September, 1999

[10] J. Friedman, “Regularized Discriminant Analysis,” Journal of American Statistical Association, vol 84, no. 405, pp. 165-175, 1989

[11] P. Hong, Q. Tian, T. S. Huang, "Incorporate Support Vector Machines to Content-Based Image Retrieval with Relevance Feedback", IEEE Int’l Conf. on Image Proc. (ICIP'2000), Vancouver, Canada, Sep 10-13, 2000.

[12] Y. Ishikawa, R. Subramanya, and C. Faloutsos, “MindReader: Query databases through multiple examples”, in Proc. Of the 24th VLDB Conf. (New York), 1998

[13] J. Laaksonen, M. Koskela, and E. Oja. “PicSOM: Self-Organizing Maps for Content-Based Image Retrieval,” Proc. of IJCNN'99. Washington, DC. July 1999

[14] C. Meilhac and C. Nastar. “Relevance feedback and category search in image databases,” In IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, June 1999

[15] S. Mika, G. Ratsch, and K.-R. Muller. “A mathematical programming approach to the Kernel Fisher algorithm,” In Advances in Neural Information Processing Systems 13, 2001(accepted)

[16] M. Naphades, R. Wang, T. Huang, “Multimodal pattern matching for audio/visual query and retrieval,” SPIE Photonics West, Storage and Retrieval for Media Databases, San Jose, CA. Jan 2001.

[17] C. Nastar, M. Mitschke and C. Meilhac “Efficient Query Refinement for Image Retrieval”, IEEE Conf. Computer Vision and Pattern Recognition CVPR'98, Santa Barbara, CA, June 1998.

[18] J. Peng, B. Bhanu, and S.Qing, “Probabilistic feature relevance learning for content-based image retrieval”, Computer Vision and Image Understanding, 75:150-164, 1999

[19] R. W. Picard, T. P. Minka, and M. Szummer, “Modeling User Subjectivity in Image Libraries”, IEEE Int’l Conf. Image Processing (ICIP’96), Lausanne, Sept. 1996.

[20] K. Porkaew, S. Mehrotra, and M. Ortega, “Query Reformulation for Content Based Multimedia Retrieval in MARS”, IEEE Int’l Conf. Multimedia Computing and Systems (ICMCS’99), June, 1999

[21] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval”, IEEE Tran on Circuits and Systems for Video Technology, Vol 8, No. 5: 644-655, Sept., 1998,

[22] Y. Rui, T. S. Huang, “Optimizing learning in image retrieval”, Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, June, 2000

[23] G. Salton, Automatic text processing, Addison-Wesley, Reading, Mass., 1989

[24] S. Santini, and Jain, R., "Integrated Browsing and Querying for Image Database," IEEE Multimedia, Vol. 7, No.3, 2000, page 26-39

[25] R. Schettini, G. Ciocca, and I. Gagliardi, “Content-based Color Image Retrieval with Relevance Feedback”, IEEE Int’l Conf. Image Processing (ICIP’99), Kobe, 1999.

[26] J.R. Smith and S. F. Chang. "VisualSEEk: a Fully Automated Content-Based Image Query System" In Proc. ACM Intern. Conf. Multimedia, pages 87--98, 1996

[27] J. R. Smith and S. F. Chang, Transform features for texture classification and discrimination in large image databases, Proc. IEEE Int’l Conf. Image Processing, Austin, Texas, Oct. 1994

[28] S. W. Smoliar and H. Zhang. Content-based video indexing and retrieval. IEEE Multimedia, 1(2):62--75, 1994

[29] K. Tieu and P. Viola, “Boosting Image Retrieval”, Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, SC, June, 2000.

[30] V. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995

[31] N. Vasconcelos, A. Lippman, “Bayesian relevance feedback for content-based image retrieval”, in Proc.IEEE Workshop on Content-based Access of Image and Video Libraries, CVPR’00, Hilton Head Island, SC, June, 2000

[32] E. Wold, T. Blum, D. Keislar, and J. Wheaton. "Content-based Classification Search and Retrieval of Audio". IEEE Multimedia Magazine, Fall 1996

[33] M. E. J. Wood, N. W. Campbell, and B. T. Thomas. “Iterative Refinement by Relevance Feedback in Content-Based Digital Image Retrieval,” ACM Multimedia 98, pages 13--20, Bristol, UK, September 1998

[34] M. Worring, A. Smeulders, and S. Santini "Interaction in Content-based Image Retrieval: a state-of-the-art review" Int’l Conf. on Visual Information Systems, Visual 2000, Lyon, France, August 2000

[35] Y. Wu, Q. Tian, T. S. Huang, "Discriminant EM Algorithm with Application to Image Retrieval", Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, SC, June, 2000.

[36] X. S. Zhou, T. S. Huang, “Image retrieval: Feature primitives, feature representation, and relevance feedback”, IEEE Workshop on Content-based Access of Image and Video Libraries, CVPR-2000, Hilton Head, SC. June, 2000.

[37] X. S. Zhou, T. S. Huang, “Edge-based structural feature for content-based image retrieval”, Pattern Recognition Letters, Special issue on Image and Video Indexing, 2001, accepted.

[38] X. S. Zhou, T. S. Hunag, "A generalized relevance feedback scheme for image retrieval", Proceedings of SPIE Vol. 4210: Internet Multimedia Management Systems, Boston, MA. November 6-7, 200