Discriminant Subspace Analysis: An Adaptive Approach for Image Classification

Yijuan Lu, *Member, IEEE*, and Qi Tian, *Senior Member, IEEE*

***Abstract*—Linear discriminant analysis (LDA) and biased discriminant analysis (BDA) are two effective techniques for dimension reduction, which pay attention to different roles of the positive and negative samples in finding discriminating subspace. However, the drawbacks of these two methods are obvious: LDA has limited efficiency in classifying sample data from subclasses with different distributions, and BDA does not account for the underlying distribution of negative samples.**

In order to effectively exploit favorable attributes of both BDA and LDA and avoid their unfavorable ones, we propose a novel adaptive discriminant analysis (ADA) for image classification. ADA can find an optimal discriminative subspace with adaptation to different sample distributions.

In addition, three novel variants and extensions of ADA are further proposed:

1) **Integrated Boosting (*i*.Boosting), which enhances and combines a set of ADA classifiers into a more powerful one. *i*.Boosting integrates feature re-weighting, relevance feedback, and AdaBoost into one framework. With affordable computational cost, *i*.Boosting can provide a unified and stable solution to ADA prediction result.**
2) **Fast adaptive discriminant analysis (FADA). Instead of searching parameters, FADA can directly find a close-to-optimal projection very fast based on different sample distributions.**
3) **Two-dimensional adaptive discriminant analysis (2DADA). As opposed to ADA, 2DADA is based on 2-D image matrix representation rather than 1-D vector. So it is simpler, more straightforward, and has lower time complexity to use for image feature extraction.**

Extensive experiments on synthetic data, UCI benchmark data sets, hand-digit data set, four facial image data sets, and COREL color image data sets show the superior performance of our proposed approaches.

***Index Terms*—AdaBoost, adaptive discriminant analysis, feature re-weighting, image classification, multiple classifiers, relevance feedback.**

I. Introduction

RECENT years have witnessed an explosion of digital images generated from different areas such as commerce, academia, and medical institutes. The dramatic increase of images demands efficient indexing and retrieval methods, especially for a large image database. In image retrieval, an image is represented by its feature vector as a data point in a high-dimensional space. Its dimension ranges from tens to hundreds. However, traditional statistical approaches have difficulties in modeling data directly in such a high-dimensional space. Hence, dimension reduction techniques play a critical role in alleviating the high dimensionality problem.

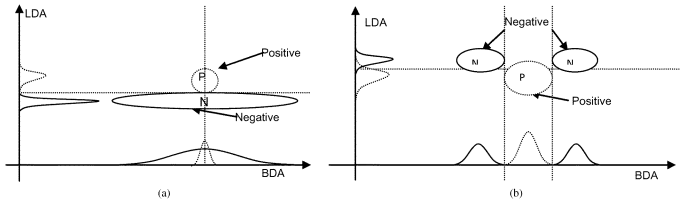


Fig. 1. LDA outperforms BDA in (a), and BDA outperforms LDA in (b).

Linear discriminant analysis (LDA) [1] and biased discriminant analysis (BDA) [2] are both effective techniques for feature dimension reduction. LDA assumes that positive and negative samples are from the same distributions, respectively, and makes the equivalent (unbiased) effort to cluster negative and positive samples. Compared to LDA, BDA assumes that positive samples must be similar while negative samples may be from different distributions. Hence, it tries to find an optimal mapping that all positive examples are clustered and all negative examples are scattered away from the centroid of the positive examples. Studies have shown that BDA works very well in content-based image retrieval (CBIR), especially when the size of the training sample set is small [2].

Obviously, LDA and BDA have their pros and cons. When all negative samples are from the same distribution and clustered together, LDA outperforms BDA [Fig. 1(a)]. However, BDA outperforms LDA when negative samples are from different classes and scattered [Fig. 1(b)]. In addition, many applications do not fit exactly into either of the two assumptions, which means neither LDA nor BDA can find an optimal projection (as shown in Fig. 2).

Hence, we propose a novel adaptive discriminant analysis (ADA) [3], which merges LDA and BDA in a unified framework and offers more flexibility and a richer set of alternatives to LDA and BDA in the parametric space. ADA can find a good projection with adaptation to different sample distributions and perform the classification in the subspace with naïve Bayes classifier.

To improve the performance of ADA and save the computational cost, three novel variants and extensions of ADA are proposed in this paper:

Manuscript received November 12, 2008; revised April 23, 2009. First published August 18, 2009; current version published October 16, 2009. This work was supported in part by Research Enhancement Program (REP) and start-up funding from Texas State University and Army Research Office (ARO) grant under W911NF-05-1-0404, and in part by the Department of Homeland Security (DHS). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Z. Jane Wang.

Y. Lu is with the Department of Computer Science, Texas State University, San Marcos, TX 78666 USA (e-mail: yl12@txstate.edu).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2030632

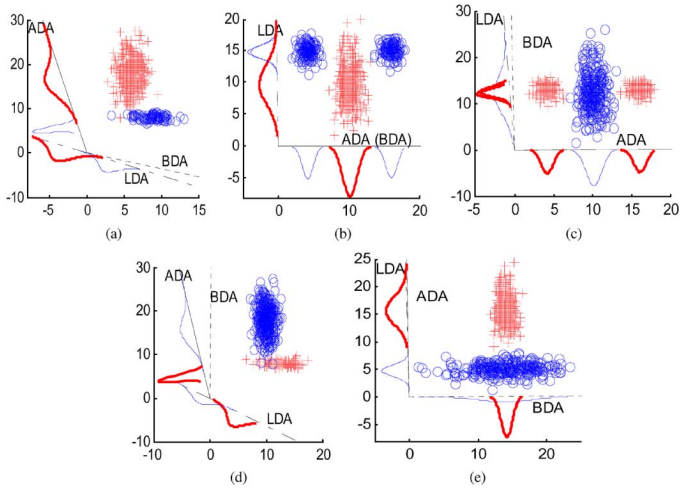


Fig. 2. Comparison of BDA, LDA, and ADA on 2-D synthetic data. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5.

1) *i*.Boosting, which is proposed to improve the performance by incorporating feature re-weighting and relevance feedback in boosting algorithm.
   ADA is a parametric method. Parameter optimization and selection are important but difficult. In ADA, it needs searching the whole parameter space to find the optimal one. The computational cost could be expensive. In addition, excessive searching also causes overfitting problem. In order to solve these two problems, a solution is to boost a set of ADA classifiers (an ADA classifier is denoted as ADA projection and a base classifier in the projected space). Then combine these boosted classifiers using some fusion scheme in the projected space.
   Here, boosting algorithms are designed to construct a "strong" classifier from a "weak" learning algorithm and present the superior result given by a thresholded linear combination of the weak classifiers. AdaBoost [4] is often regarded as the generic boosting algorithm. The basic idea of AdaBoost is to iteratively re-weight the training samples based on the outputs of some weak learners. Misclassified samples will receive higher weights in the next iteration. This forces the classifier to focus more on the incorrectly classified examples.
   However, in traditional AdaBoost, only weights of samples are updated. It does not update any feature element weight, which is important and very useful especially for image databases using high-dimensional image features [5]. In this paper, we incorporate feature re-weighting into boosting. To further improve the understanding of visual content and user interest and alleviate the overfitting problem, we integrate relevance feedback into boosting scheme and propose a novel integrated boosting framework (*i*.Boosting). *i*.Boosting not only weights the samples but also weights the feature elements iteratively. Besides, in *i*.Boosting, relevance feedback provides boosting with more training information. Better than simple relevance feedback, *i*.Boosting forces classifiers to pay more attention to wrongfully predicted samples through user feedback.
2) FADA, which stands for fast adaptive discriminant analysis. The major difference between FADA and ADA lies in the adaptation method. FADA does not need searching parameters like ADA. It can directly calculate the close-to-optimal prediction in a fast way according to sample distributions. Extensive experiments show that FADA has distinctly lower costs in time than ADA and achieves classification accuracy that is comparable to ADA.
3) 2DADA. In contrast to vector representation used in ADA, 2DADA works on the matrix representation of images directly. As a result, 2DADA has two advantages. First, it is easier to evaluate the scatter matrix as the spatiality and locality information are better preserved. Second, less time is required to determine the corresponding eigenvectors since the size of scatter matrix is smaller.

The rest of the paper is organized as follows. In Section II, we illustrate the ADA in detail. In Sections III– V, *i*.Boosting, FADA, and 2DADA are introduced and discussed, respectively. Extensive experiments are also conducted to evaluate all these proposed methods on UCI benchmark data sets, hand-digit data set, four facial image data sets, and COREL color image data sets. Finally, conclusions and future work are discussed in Section VI.

II. Adaptive Discriminant Analysis

A. Linear Discriminant Analysis

LDA is one of most widely used discriminant analysis techniques in classification and dimension reduction. LDA tries to find an optimal projection W from originally high d_1-dimensional space to a low d_2-dimensional space, which makes samples from the same class cluster and samples from different classes separate. The problem of finding the optimal W can be mathematically represented as the following maximization problem:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{1}$$

$$S_B = \sum_{j=1}^{C} N_j \cdot (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \tag{2}$$

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} \left(\mathbf{x}_i^{(j)} - \mathbf{m}_j\right)\left(\mathbf{x}_i^{(j)} - \mathbf{m}_j\right)^T. \tag{3}$$

Here, the between-class matrix S_B measures the separability of class centers and the within-class scatter matrix S_W measures the within-class variance in the low-dimensional space. $\{\boldsymbol{x}_i^{(j)}, i = 1, \ldots, N_j\}$, $j = 1, \ldots, C$ denote the feature vectors of training samples. C is the number of classes. N_j is the number of the samples of the jth class, $\boldsymbol{x}_i^{(j)}$ is the ith sample vector from the jth class, $\boldsymbol{m}_j$ is the mean vector of the jth class, and $\boldsymbol{m}$ is the grand mean of all examples.

To maximize the ratio of (1), the optimal W is composed of the generalized eigenvector(s) $\boldsymbol{w}_i$ associated with the largest eigenvalue(s). $W_{opt} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{C-1}]$ contains $C-1$ eigenvectors corresponding to $C-1$ eigenvalues, i.e., $S_B w_i =$

$\lambda_i S_W w_i$ [1]. It should be noted that W maps the original d_1-dimensional data space $\boldsymbol{X}$ to a d_2-dimensional space Δ (where $d_2 \leq C-1$).

B. Biased Discriminant Analysis

In two-class LDA, the equivalent (unbiased) effort has been made to cluster negative and positive samples. Intuition suggests that clustering the negative samples may be difficult and unnecessary because they may be from different classes [Fig. 1(b)]. Zhou and Huang [2] proposed biased discriminant analysis (BDA). The intuition behind the BDA is that "all positive examples are alike, and each negative example is negative in its own way". That means that the positive samples are visually similar and should be clustered in the projected space. On the other hand, the negative samples might be from different classes, and it is difficult to find a mapping to make them close to each other.

BDA is defined to find an optimal projection:

$$W_{opt} = \arg\max_W \frac{|W^T S_{N\to P} W|}{|W^T S_P W|} \tag{4}$$

$$S_{N->P} = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{5}$$

$$S_P = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{6}$$

where $\boldsymbol{m}_P$ is the mean vector of the positive examples. $S_{N\to P}$ is the scatter matrix between the negative examples and the centroid of the positive examples, and S_P is the scatter matrix of the positive examples. $N \to P$ indicates the asymmetric property of this approach, which means the user's biased opinion towards the positive class, thus the name of biased discriminant analysis [2].

Although the idea of BDA is simple and it is quite effective in content-based image retrieval, we find that its assumption is still inappropriate in some scenarios, which will be explained in the next section. The complex nature of image data requires a classification method that can adaptively fit the distribution of image data from different classes and discover a good classification boundary.

C. Adaptive Discriminant Analysis

Given that LDA and BDA have their own assumptions and pay attention to different roles of the positive and the negative examples in finding the optimal discriminating subspace, it is our expectation that they can be unified. In addition, there are many cases that both LDA and BDA are not applicable.

To provide a better model fitting the complex distributions for positive and negative samples, an adaptive discriminant analysis (ADA) was proposed [3], which finds an optimal projection

$$W_{opt} = \arg\max_W \frac{\left|W^T\left[\lambda\cdot S_{P\to N} + (1-\lambda)\cdot S_{N\to P}\right]W\right|}{\left|W^T\left[\eta\cdot S_P + (1-\eta)\cdot S_N\right]W\right|} \tag{7}$$

in which

$$S_{N->P} = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{8}$$

TABLE I
SPECIAL CASES OF ADA

(λ,η)	Optimal Projection	Note
(0,0)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{N->P}W^T\rvert}{\lvert WS_NW^T\rvert}$	Case 1
(0,1)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{N->P}W^T\rvert}{\lvert WS_PW^T\rvert}$	Case 2 (BDA)
(1,0)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{P->N}W^T\rvert}{\lvert WS_NW^T\rvert}$	Case 3 (Counter-BDA)
(1,1)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{P->N}W^T\rvert}{\lvert WS_PW^T\rvert}$	Case 4
(0.5,0.5)	$W_{ADA} = \arg\max_W \frac{\lvert W(S_{P->N}+S_{N->P})W^T\rvert}{\lvert W(S_P+S_N)W^T\rvert}$	Case5 (LDA-like)

$$S_{P->N} = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{9}$$

$$S_P = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{10}$$

$$S_N = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T. \tag{11}$$

$\boldsymbol{m}_P$, $\boldsymbol{m}_N$, S_P, S_N, $S_{N\to P}$, $S_{P\to N}$ are defined the same or in a similar way as before. The two parameters λ and η control the bias between the positive and negative samples and range from (0,0) to (1,1).

Table I summarizes five special cases of ADA. From Table I, we can find that the ADA reduces to BDA when λ and η are set to be 0 and 0 (Case 1). Case 5 corresponds to a LDA-like projection with λ and η set to 0.5 and 0.5. Case 4 finds a projection that is on the opposite side of BDA, which is called Counter-BDA. Case 2 and Case 3 is a couple of contrary distribution scenarios, which assume that the negative (positive) samples are similar and positive (negative) samples might be from different classes. All these five cases fit certain image distributions and have correspondence with some scenarios as illustrated in Fig. 2.

In order to show the advantages of ADA, we use synthetic data to simulate different sample distributions as shown in Fig. 2. Original data are simulated in 2-D space, and the positive examples are marked with $+$' s and the negative examples are marked with o's as shown in the figure. In each case, we apply BDA, LDA, and ADA to find the best projection direction by their own criterion functions. ADA searches 36 parameter combinations (λ, η) sampled from 0 to 1 with step size of 0.2 to find the best one. The resulting projection lines are drawn in dotted, dash-dotted, and solid lines, respectively. In addition, the distributions of the positive and negative samples along these projections are also drawn like bell-shaped thicker and thinner curves along projection line, assuming Gaussian distribution for each class.

From Fig. 2, we can see these five cases could represent several data distribution scenarios. Case 1 best fits the distribution where all positive samples are alike while negative ones may be irrelevant to each other and from different distributions [Fig. 2(a)]. Case 4 is on the opposite side of Case 1, in which negative samples share strong correlations while positive samples may be quite different [Fig. 2(d)]. Cases 2 and 3 represent the imbalanced data set, where the size of positive (negative)

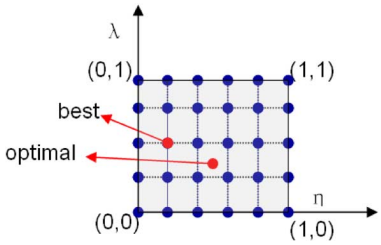


Fig. 3. Best ADA classifier found in the parameter space may not be the one with optimal setting.

samples is much larger than that of negative (positive) samples. Case 5 is the scenario where the major descriptive directions of positive samples and negative samples are upright.

From projection results, we can see LDA treats positive and negative samples equally. This makes it a bad choice in Cases 1 and 4. Similarly, since BDA assumes all positive samples are projected together, it fails in Cases 4 and 5. In Cases 2 and 3, BDA and LDA are found not applicable for imbalanced data sets. The reason for this is that LDA or BDA tends to severely bias to the dominating samples.

In all five cases, ADA yields good projection with positive samples and negative samples well separated and outperforms BDA and LDA. Note in Case 4, both BDA and LDA totally fail while ADA still produces a good projection. It clearly demonstrates that no matter whether it is an imbalanced data set or samples are from different subclass clusters, ADA can adaptively fit into different distributions of samples and find a balance between clustering and separating, which are embedded in the criterion function. Here, only five special cases of ADA are shown. More accurate data model fitting could be achieved by fine parameter tuning.

III. *i*.BOOSTING

In previous experiments on synthetic data, ADA analysis has shown good performance. However, the optimal classifier often lies not only between but also beyond BDA and LDA in the parametric space of (λ, η) (Fig. 3). To find the best parameter setting for ADA on a particular data set, exhaustive searching in the square region of parameters from (0,0) to (1,1) is needed. But it is hard to decide a trade-off between computational cost and accuracy. When the searching-step size is large, it will miss the global optimal value, and when the step size is small, it often causes overfitting problem. It is also true that the best pair we found for one particular data set is often different from that of other data sets, and therefore, this cannot lead to a generalization. In order to alleviate this problem, a feasible solution is to boost a set of ADA classifiers and combine these boosted classifiers using some fusion scheme in the projected space. AdaBoost [4] is one of the popular boosting algorithms.

A. *AdaBoost*

AdaBoost [4] developed in the computational machine learning area has emerged as a competitive technique that has a theoretically justified ability to improve the performance of any weak classification algorithm in terms of bounds on the generalization error.

The basic idea of AdaBoost is to iteratively re-weight the training examples based on the outputs of some weak learners. In order to boost the weak learning algorithm, the data are reweighed (the relative importance of the training examples is changed) before running the weak learning algorithm at each iteration. Training examples that were misclassified by the weak classifier at the current iteration then receive higher weights at the following iteration. The intention is to increase the weights of the incorrectly classified examples and decrease the weights of the correctly classified examples. This forces the classifier to focus more on the incorrectly classified examples in the next iteration. The end result is a final combined classifier. Each component is the weak classifier obtained at each iteration, and each component classifier is weighted according to how this classifier performed during each iteration.

AdaBoost performs better than many state-of-the-art classification algorithms in experiments, and it does not seem to overfit. Theories trying to explain this include the margin theory [6]–[8] and the additive logistic regression [9]. These explanations have in turn given modifications or improvements over the original AdaBoost. Therefore, AdaBoost provides a general way of combining and enhancing a set of ADA classifiers in the parametric space.

B. *i.Boosting*

Although AdaBoost can enhance each classifier's performance by re-weighting and re-training mechanism, it only updates the weights of samples. It does not update any feature element weight, which is important and very useful, especially for image databases using high-dimensional image features [5]. Hence, we incorporate feature re-weighting into boosting and propose a new feature re-weighting approach for ADA. In addition, considering feature re-weighting on small training data set tends to bias to the training set and causes overfitting, we incorporate relevance feedback [10] into boosting scheme and propose a novel integrated boosting framework (*i*.Boosting) [11]. In contrast to AdaBoost, *i*.Boosting not only re-weights the samples but also re-weights the feature elements iteratively. Besides, by incorporating relevance feedback in *i*.Boosting, it can further improve the understanding of visual content and user interests.

1) Classic Feature Re-Weighting: In image database, each image $i \in I$ is represented by its M features $\boldsymbol{f}_i = [f_{i_1}, f_{i_2}, \ldots f_{i_M}]^T$. Let the feature of query image q be $\boldsymbol{f}_q = [f_{q_1}, f_{q_2}, \ldots f_{q_M}]^T$, the Euclidean distance between query image and the image in the database is

$$d = (\boldsymbol{f}_i - \boldsymbol{f}_q)^T W (\boldsymbol{f}_i - \boldsymbol{f}_q). \tag{12}$$

W is the feature weighting matrix indicating the importance of each component of features. After relevance feedback, the user provides the relevance of each image to the query and the feature weights can be updated to make similar images close to each other and dissimilar images far away from each other. Traditional feature re-weighting methods are based on distance metric, e.g., generalized Euclidean distance [5]. In this paper, we propose a dynamic feature re-weighting method for dimension reduction in order to obtain a better projection during relevance feedback.

2) Weighted Adaptive Discriminant Analysis: We incorporate dynamic feature weighting to ADA and construct weighted ADA. It can provide a more accurate model of the complex distribution for positive and negative samples by finding an optimal projection:

$$W_{ADA} = \arg\max_{W} \frac{\left|W^T\left[\lambda\widetilde{S}_{P->N} + (1-\lambda)\widetilde{S}_{N->P}\right]W\right|}{\left|W^T\left[\eta\widetilde{S}_P + (1-\eta)\widetilde{S}_N\right]W\right|} \tag{13}$$

in which

$$\widetilde{S}_{N->P} = \sum_{i\in Negative} \left(\boldsymbol{f}_{N_\bullet} * \boldsymbol{x}_i - \widetilde{\boldsymbol{m}}_P\right) \times \left(\boldsymbol{f}_{N_\bullet} * \boldsymbol{x}_i - \widetilde{\boldsymbol{m}}_P\right)^T \tag{14}$$

$$\widetilde{S}_{P->N} = \sum_{j\in Positive} \left(\boldsymbol{f}_{P_\bullet} * \boldsymbol{x}_j - \widetilde{\boldsymbol{m}}_N\right) \times \left(\boldsymbol{f}_{P_\bullet} * \boldsymbol{x}_j - \widetilde{\boldsymbol{m}}_N\right)^T \tag{15}$$

$$\widetilde{S}_P = \sum_{j\in Positive} \left(\boldsymbol{f}_{P_\bullet} * \boldsymbol{x}_j - \widetilde{\boldsymbol{m}}_P\right) \times \left(\boldsymbol{f}_{P_\bullet} * \boldsymbol{x}_j - \widetilde{\boldsymbol{m}}_P\right)^T \tag{16}$$

$$\widetilde{S}_N = \sum_{i\in Negative} \left(\boldsymbol{f}_{N_\bullet} * \boldsymbol{x}_i - \widetilde{\boldsymbol{m}}_N\right) \times \left(\boldsymbol{f}_{N_\bullet} * \boldsymbol{x}_i - \widetilde{\boldsymbol{m}}_N\right)^T. \tag{17}$$

$\boldsymbol{f}_P$ and $\boldsymbol{f}_N$ are feature element weights of positive and negative samples. $\widetilde{\boldsymbol{m}}_P$, $\widetilde{\boldsymbol{m}}_N$ are means of weighted positive samples and weighted negative samples. $\bullet*$ stands for Hadamard product operation. The rest are defined the same as in Section II.

In order to avoid searching in 2-D parameter (λ, η) space, we use AdaBoost to enhance and combine a set of weak ADA classifiers into a more powerful one. Unlike most of the existing approaches that boost individual features to form a composite classifier, our scheme boosts both the individual features and a set of weak classifiers.

For each weak ADA classifier, to find a better projection, the ratio of $trace(\lambda\widetilde{S}_{P->N} + (1-\lambda)\widetilde{S}_{N->P})/trace(\eta\widetilde{S}_P + (1-\eta)\widetilde{S}_N)$ needs to be maximized. Intuitively, it is to minimize the "within-class scatter" and maximize the "between-class scatter". Therefore, the criterion can be redefined to maximize

$$\begin{aligned} &trace\left(\lambda\widetilde{S}_{P->N} + (1-\lambda)\widetilde{S}_{N->P}\right) \\ &\quad - trace\left(\eta\widetilde{S}_P + (1-\eta)\widetilde{S}_N\right) \\ &= trace(\lambda\widetilde{S}_{P->N} - \eta\widetilde{S}_P) \\ &\quad + trace\left((1-\lambda)\widetilde{S}_{N->P} - (1-\eta)\widetilde{S}_N\right). \end{aligned} \tag{18}$$

Hence, re-weighting scheme for ADA is to update $\boldsymbol{f}_P$ by maximizing $trace(\lambda\widetilde{S}_{P->N} - \eta\widetilde{S}_P)$ and update $\boldsymbol{f}_N$ by maximizing $trace((1-\lambda)\widetilde{S}_{N->P} - (1-\eta)\widetilde{S}_N)$ based on (14)–(17).

3) Relevance Feedback: To efficiently incorporate user feedback and enhance the retrieval accuracy, relevance feedback can also be integrated in the boosting.

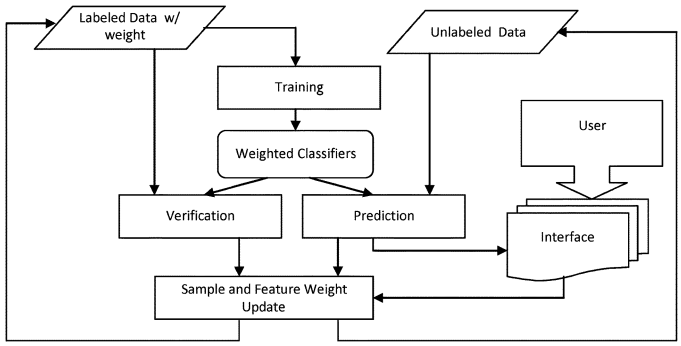


Fig. 4. Integrated boosting framework.

Relevance feedback was initially developed in document retrieval [12] and widely applied in content-based image retrieval [10], [13]. The basic idea of relevance feedback is to get human in the loop. At first, computer processing provides initial retrieval results. Users are then asked to evaluate the current retrieval results according to degrees that are relevant or irrelevant to his/her request. The system then applies the user's feedback to update the training examples to improve performance for the next round. This learning process can be applied iteratively if the user desires. Relevance feedback algorithms have been shown to provide dramatic performance improvement in image retrieval systems [12].

4) i.Boosting for ADA: Motivated by the strength and successes of AdaBoost, dynamic feature re-weighting, and relevance feedback, we propose an integrated boosting framework called *i*.Boosting. It can integrate relevance feedback, AdaBoost (sample re-weighting), and feature re-weighting in the loop of boosting and better bridge the gap between semantic concept and image features. Fig. 4 gives an illustration of the integrated boosting framework.

Based on the above framework, the brief algorithm below shows how the *i*.Boosting is implemented with multiple ADA classifiers.

Algorithm: *i*.Boosting with ADA as weak classifiers

Input: Labeled Sample Set $\boldsymbol{X}$ and label Y

Unlabeled Sample Set $\boldsymbol{U}$

Feature vector D and Feature element d

K ADA classifiers with different (λ, η)

M: The dimension of feature (feature size)

p: positive samples n: negative samples

T: The total number of runs

Initialization: sample weight $w_{k,t=1}(\boldsymbol{x}) = 1/|X|$ and feature weight $f_{p,t=1}(d) = 1$, $f_{n,t=1}(d) = 1$

Boosting

For each classifier $k = 1, \ldots, K$ do

For $t = 1, \ldots, T$

- Find the optimal projection

$$W_{opt} = \arg\max_{W} \frac{\left|W^T \left[\lambda \widetilde{S}_{P->N} + (1-\lambda)\widetilde{S}_{N->P}\right] W\right|}{\left|W^T \left[\eta \widetilde{S}_P + (1-\eta)\widetilde{S}_N\right] W\right|}$$

based on weighted mean for positive samples, negative samples, and all the samples and weighted scatter matrices in the following way. Note that $\sum_{x\in X} w_{k,t}(x) = 1$

$$\sum_{d\in D} f_{p,t}(d) = \sum_{d\in D} f_{n,t}(d) = M$$

1) Update weighted mean μ_p, μ_n, and μ_{all}

$$\mu_p = \sum_{x\in p} w_{k,t}(\boldsymbol{x}) \cdot (f_{p,t_\bullet} * \boldsymbol{x}) \Big/ \sum_{x\in p} w_{k,t}(\boldsymbol{x})$$

$$\mu_n = \sum_{x\in n} w_{k,t}(\boldsymbol{x}) \cdot (f_{n,t_\bullet} * \boldsymbol{x}) \Big/ \sum_{x\in n} w_{k,t}(\boldsymbol{x})$$

$$\mu_{all} = \left(\sum_{x\in p} w_{k,t}(\boldsymbol{x}) \cdot (f_{p,t_\bullet} * \boldsymbol{x}) + \sum_{x\in n} w_{k,t}(\boldsymbol{x}) \cdot (f_{n,t_\bullet} * \boldsymbol{x})\right) \Big/ \sum w_{k,t}(\boldsymbol{x})$$

2) Update within-class and between-class scatter matrices

$$\widetilde{S}_{N->P} = \sum_{x\in n} (f_{n,t_\bullet} * \boldsymbol{x} - \mu_P)\, w_{k,t}(\boldsymbol{x})\, (f_{n,t_\bullet} * \boldsymbol{x} - \mu_P)^T \Big/ \sum_{x\in n} w_{k,t}(\boldsymbol{x})$$

$$\widetilde{S}_{P->N} = \sum_{x\in p} (f_{p,t_\bullet} * \boldsymbol{x} - \mu_N)\, w_{k,t}(\boldsymbol{x})\, (f_{p,t_\bullet} * \boldsymbol{x} - \mu_N)^T \Big/ \sum_{x\in p} w_{k,t}(\boldsymbol{x})$$

$$\widetilde{S}_P = \sum_{x\in p} (f_{p,t_\bullet} * \boldsymbol{x} - \mu_P)\, w_{k,t}(\boldsymbol{x})\, (f_{p,t_\bullet} * \boldsymbol{x} - \mu_P)^T \Big/ \sum_{x\in p} w_{k,t}(\boldsymbol{x})$$

$$\widetilde{S}_N = \sum_{x\in n} \left(\boldsymbol{f}_{n,t_\bullet} * \boldsymbol{x} - \mu_N\right) w_{k,t}(\boldsymbol{x})\, (f_{n,t_\bullet} * \boldsymbol{x} - \mu_N)^T \Big/ \sum_{x\in n} w_{k,t}(\boldsymbol{x})$$

Train weak classifiers on the data projected by the optimal projection W_{opt}.

- Get the probability-rated prediction on labeled and unlabeled sample $h_{k,t}(\boldsymbol{x}) \in (-1, 1)$.
  Suppose the probability of a samples x belongs to positive and negative class is denoted as $P(\boldsymbol{x} \in p)$ and $P(\boldsymbol{x} \in n)$, respectively.
  If $P(x \in p) \geq P(\boldsymbol{x} \in n)$

$$h(\boldsymbol{x}) = \frac{P(\boldsymbol{x} \in p)}{P(\boldsymbol{x} \in p) + P(\boldsymbol{x} \in n)}$$

else

$$h(\boldsymbol{x}) = \frac{-P(\boldsymbol{x} \in n)}{P(\boldsymbol{x} \in p) + P(\boldsymbol{x} \in n)}$$

- Assign the weights of classifiers based on its classification error rate $\varepsilon_{k,t}$ on labeled samples

$$\alpha_{k,t} = \frac{1}{2} \ln\left(\frac{1-\varepsilon_{k,t}}{\varepsilon_{k,t}}\right)$$

- Present samples from the unlabeled data set with their predicted labels to user.
- Obtain user feedback on the ground truth labels.
- Data obtained from user relevance feedback are added to construct an enlarged labeled data set and removed from unlabeled data set.
- Update the sample weight based on their prediction correctness: $w_{k,t+1}(\boldsymbol{x}) = w_{k,t}(\boldsymbol{x}) \exp(-\alpha_{k,t} \cdot h_{k,t}(\boldsymbol{x}) \cdot y)$.
- Update the feature weights based on the proposed feature re-weighting rule:
  - compute the new $|\lambda \widetilde{S}_{P\to N} - \eta \widetilde{S}_P|$ and $|(1-\lambda)\widetilde{S}_{N\to P} - (1-\eta)\widetilde{S}_N|$ in (18).
  - Update the weight of features $f_{p,t}(d)$, $f_{n,t}(d)$ accordingly.

  End for t

End for each classifier

The final prediction $H(x) = sign(\sum_{k,t} \alpha_{k,t} \cdot h_{k,t}(\boldsymbol{x}))$, using sum rule to combine multiple classifiers.

C. Experiments and Results

In this section, we experimentally evaluate the performance of ADA and *i*.Boosting on various benchmark data sets including UCI benchmark data sets, the COREL image data set, and four face data sets. In order to comprehensively evaluate the performance of our proposed method, we compare it with BDA, LDA, AdaBoost, ADA with relevance feedback, and other state-of-the-art projection techniques. In all experiments, a Bayesian classifier is used on the projected data for all projection-based methods.

1) Comparison of ADA and the State-of-the-Art: In the first experiment, we compare ADA with the state-of-the-art linear and nonlinear variants of discriminant analysis, such as DEM [14], kernel DEM (KDEM) [15], BDA [2], and kernel BDA (KBDA) [2] on face and non-face recognition.

The data set used in the experiments contains both face images, which come from MIT facial image data set (2358 images) [16] and non-face images (2958 images) from Corel database [17]. All the face and non-face images are scaled down to 16 $\times$ 16 grayscale images and normalized feature vector of dimension 256 is used to represent each image. The different sizes of the training sets are 100, 200, 400, and 800, respectively. Compared with the feature vector dimension of 256, the training set size is set from relatively small to relatively large. Table II gives the experimental results with the smallest error rates in bold.

TABLE II
COMPARISON WITH DEM, BDA, KDEM, AND KBDA

Error Rate (%)	Size of Training Set			
	100	200	400	800
DEM w/ reg.	10.5	19.3	15.0	9.0
BDA w/ reg.	34.7	25.4	18.5	19.3
KDEM	6.93	1.93	1.7	**0.5**
KBDA	3.04	2.89	2.58	1.44
ADA (λ^*,η^*)	**2.5** (0.35,0.1)	**1.9** (0.55,0.2)	**1.7** (0.1,0.15)	1.6 (0.1,0.1)

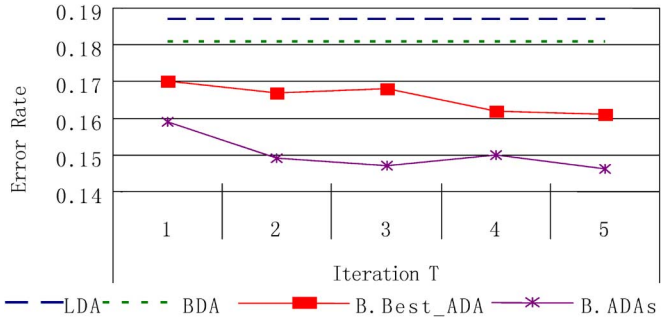


Fig. 5. ADA on heart benchmark data set.

From the results in Table II, we find that our proposed methods perform well when the training set size is small compared to the feature dimensionality. When compared with linear techniques of DEM and BDA with simple regularization [18], [19], ADA performs much better than them and does not require regularization. Even when compared with the nonlinear techniques KDEM and KBDA, ADA still performs better in three out of four tests. It should be noted that only linear transformation is used in our ADA, but it is more efficient than nonlinear algorithms such as KDEM and KBDA. These show the robust performance of the ADA.

2) Comparison of Boosted Single and Multiple ADA: In the second experiment, we evaluate the effectiveness of a boosted single best ADA classifier (B.best_ADA) with boosted multiple ADA classifiers (B.ADAs). The boosted multiple ADA classifiers are trained on 36 ADA classifiers with (λ, η) evenly sampled from 0 to 1 with step size of 0.2. The single best ADA is the best one chosen from these 36 classifiers. For comparison purpose, LDA and BDA are also implemented and tested. The methods are tested on benchmark data sets from UCI repository [20]. Due to the limited space, we only show the results in Fig. 5 on SPECTF heart databases, which describe diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Similar results are obtained on other data sets. This SPECTF data set contains 267 instances (patients) and totally 43 attributes. Each of the patients is classified into two categories: normal and abnormal. The sizes of the training set and testing set are 80 and 187, respectively.

Shown in Fig. 5, as iteration goes on, the error rates of the B.best_ADA and the B.ADAs decrease. The performance of LDA and BDA are shown for reference as two straight lines. Both B.best_ADA and B.ADAs outperform the LDA and BDA in this experiment. Although B.ADAs starts with a set of weak

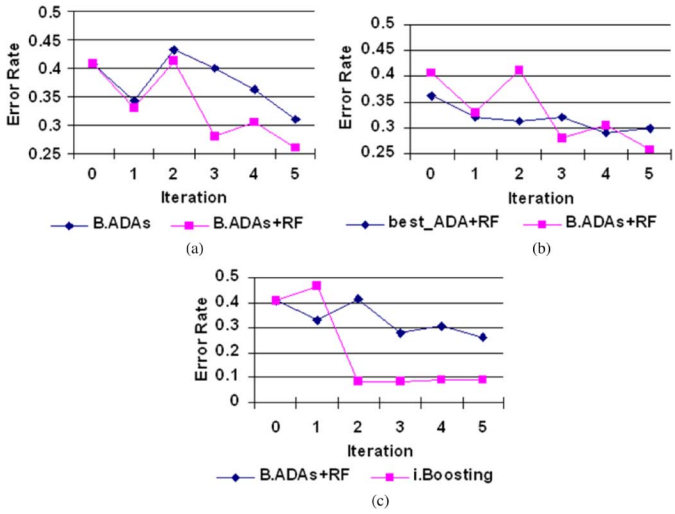


Fig. 6. Comparison between *i*.Boosting and other related variants.

classifiers, after one iteration $(T = 1)$, it outperforms the boosted single best ADA classifier.

3) i.Boosting: In order to have a statistical analysis of our scheme, we perform a **pseudo** relevance feedback. At each relevance feedback, five images are fed to the system automatically based on their ground truth labels. In the following experiments, our boosted ADA is trained on 36 ADA classifiers with (λ, η) evenly sampled from 0 to 1 with step size of 0.2 and the average prediction error rate of 50 runs is reported.

i) i.Boosting on UCI data set: First, we tested the proposed *i*.Boosting on benchmark data sets from UCI repository. For comparison purpose, four independent experiments are designed and implemented to compare *i*.Boosting with other related variants.

Due to the limited space, we only show the results on SPECTF heart databases. The average error rate across five iterations is plotted in Fig. 6, where the x-axis denotes the iteration number (between 0 to 5). 0 stands for the starting status before iterations begin.

a) boosting multiple ADA classifiers with and without relevance feedback

Secondly, we evaluate the effect of integrating user feedback into boosting scheme. From Fig. 6(a), we can find the performance improvement of using AdaBoost alone (B.ADAs) is less than that of using boosted ADAs with relevance feedback (B.ADAs+RF). The performance of B.ADAs+RF is consistently better than that of B.ADAs (without relevance feedback) by up to 30.4% on SPECTF heart set. It is no surprise that user feedback and human judgment could be accumulated iteratively to facilitate learning process.

b) single ADA classifier + RF (without boosting) versus boosting multiple ADA classifiers + RF

The third experiment is designed to verify if the performance improvement of B.ADAs+RF is introduced by relevance feedback only. Hence, we compare the single best ADA classifier with only relevance feedback (best_ADA+RF) and boosted multiple ADA classifiers with relevance feedback (B.ADAs+RF). From the experimental result in Fig. 6(b), we can conclude that: 1) B.ADAs+RF and relevance feedback

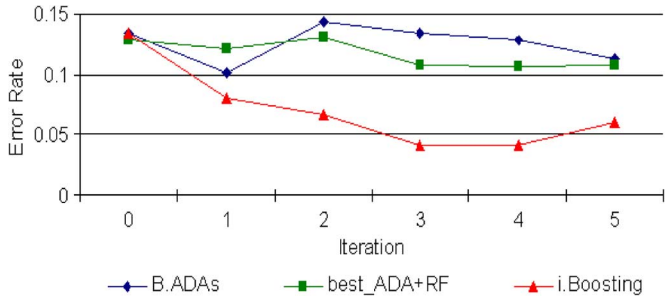


Fig. 7. *i*.Boosting on COREL data set.

(without boosting) only starts with similar performance in iteration 1; 2) After several iterations, simple relevance feedback gains less performance improvement than B.ADAs+RF. In conclusion, B.ADAs+RF has obvious advantage over the simple relevance feedback method in that the classifiers are trained to pay more attention to wrongfully predicted samples in user feedback through a reinforcement training process.

c) boosting multiple ADA classifiers+RF (without feature re-weighting) versus *i*.Boosting

The last experiment is to evaluate the performance of feature re-weighting in integrated boosting. In Fig. 6(c), we can find after two iterations, *i*.Boosting performed much better than B.ADAs+RF (without feature re-weighting). Besides, *i*.Boosting becomes much steadier after several iterations. It is clear that *i*.Boosting boosts not only a set of weak classifiers but also the individual features.

ii) i.Boosting for image classification: In order to evaluate *i*.Boosting for image classification, we test it on the COREL image databases. This database contains 1386 color images, which are categorized into 14 classes. Each class contains 99 images. Each image is represented by 37 feature components including color moments (9) [21], wavelet-based texture (10) [22], and water-filling edge-based structure features (18) [23]. For simplicity, we randomly pick up two classes of images for classification. One-third of the images are used for training while the rest are used for testing.

The experimental result shown in Fig. 7 is consistent with the results on the UCI data set. *i*.Boosting, boosted multiple ADA classifiers (without relevance feedback), and the best ADA classifier with relevance feedback start with similar performance in iteration 1. But as the iteration goes on, *i*.Boosting gains much better performance improvement than the other two. It demonstrates that interactive boosting exploits the favorable attributes of AdaBoost, feature re-weighting, and relevance feedback well.

iii) i.Boosting for face classification: To evaluate *i*.Boosting for face classification, we tested it on four well-known face image databases with change in illumination, expression, and head pose. The Harvard Face Database contains images from ten individuals, each providing total 66 images, which are classified into ten sets based on increasingly changed illumination condition [24]. The AT&T Face Database [16] consists of grayscale images of 40 persons. Each person has ten images with different expressions, open or closed eyes, smiling or non-smiling, and wearing glasses or no glasses. The UMIST Face Database [25] consists of 564 images of 20 people, which covers a range of poses from profile to frontal views. The CMU PIE face database [26] contains 41 368 images of 68 individuals. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination, and expression. We choose the five near frontal poses (C05, C07, C09, C27, C29) and use all the images under different illuminations, lighting, and expressions which leaves us 170 near frontal face images for each individual. Fig. 8 gives some example images from the databases. Sixty image features are extracted to represent these images including histogram (32), wavelet-based texture (10) [22], and water-filling edge-based structure features (18) [23].

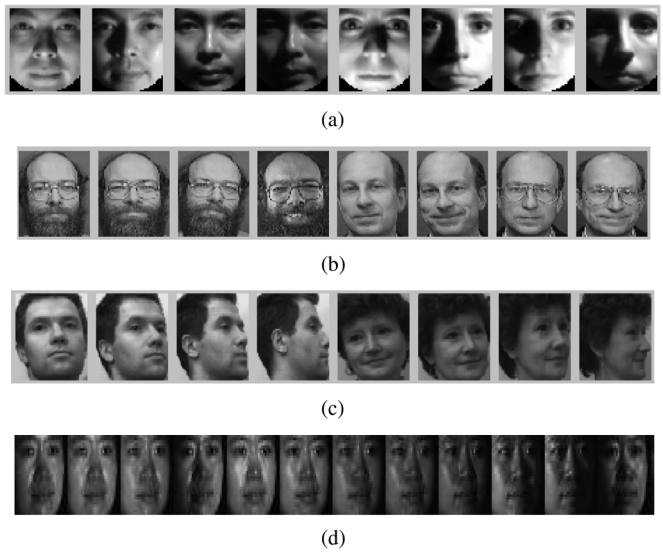


Fig. 8. Example images from four face databases. (a) Change of illumination condition, size is 84 × 96. (b) Change of expressions, size is 92 × 112. (c) Change of head pose, size is 92 × 112. (d) Change of head pose and illumination, size is 64 × 64.

For each database, we randomly choose one person's face images as positive and the rest face images of others are considered as negative. For comparison purpose, six state-of-the-art projection-based techniques: Eigenface [24], LDA, BDA [2], DEM [15], KDEM [16], and ADA are tested on the same databases. To be consistent, the results for these techniques are obtained after five iterations of relevance feedback accumulation.

The results are listed in Table III with the smallest error rate in bold. It is clear that *i*.Boosting performs best in five out of six tests and second to the best ADA in one test. Therefore, *i*.Boosting provides more robustness to the changes of illumination, expression, and pose than other techniques.

IV. Fast Adaptive Discriminant Analysis

To reduce computational cost of ADA, a very simple but effective variant of ADA, fast adaptive discriminant analysis (FADA), is proposed. Instead of searching the parametric space, FADA provides a novel and stable solution to find close-to-optimal ADA projection very quickly [27].

The basic idea of FADA is to find projections to cluster positive samples and negative samples, respectively. Then adjust these projections to separate two classes as far as possible. Fig. 9 gives an illustration of the basic idea of the FADA in two-dimensional space.

The scenario can be described in the following steps (Fig. 9):

TABLE III
COMPARISON OF *i*.BOOSTING WITH STATE-OF-THE-ART TECHNIQUES ON DIFFERENT FACE DATABASES

Error Rate (%)		Harvard Database			ATT Database	UMIST Database	PIE Database
		Subset 1	Subset 2	Subset 3			
Methods	Eigenface	6.33	9.1	4.16	0.31	3.81	36.21
	Fisherface	4.02	5.71	1.19	2.38	3.3	19.26
	LDA	15.06	15.17	15.33	2.07	0.51	14.62
	BDA	1.42	4.0	1.43	0.83	1.36	14.13
	DEM	14.96	15.18	15.26	3.35	1.28	15.4
	KDEM	11.21	13.33	11.18	1.67	2.64	12.72
	Best single ADA	0.33	**2.7**	0.84	0.04	0.17	10.51
	i.Boosting	**0.16**	3.0	**0.58**	**0.02**	**0.11**	**10.38**

1) Find a projection W_1 to cluster positive samples (P) first. Obviously, W_1 is the eigenvector(s) corresponding to the smallest eigenvalue(s) of covariance matrix of positive samples.
2) Project all positive and negative data to W_1, calculate the number of samples R_1 within the overlapping range L_1 of these two classes after projection. The smaller R_1, the more separated of these two classes. If $R_1 = 0$, the positive samples and negative samples can be completely separated by the projection W_1.
3) Similarly, find a projection W_2 to cluster negative samples (N). W_2 is the eigenvector(s) with the smallest eigenvalue(s) of covariance matrix of negative samples.
4) Project all data to W_2 and calculate the number of samples R_2 that belong to the overlapping range L_2 of the two classes.
5) Calculate the ratio

$$\lambda = \frac{\mathrm{R}_2}{\mathrm{R}_1 + \mathrm{R}_2}, \quad 1 - \lambda = \frac{\mathrm{R}_1}{\mathrm{R}_1 + \mathrm{R}_2}. \tag{19}$$

6) The final projection W is a linear combination of W_1 and W_2:

$$W = \lambda W_1 + (1 - \lambda) W_2. \tag{20}$$

Obviously, final W depends on the value of R_1 and R_2 (separability of two classes after projected by W_1 and W_2). If W_1 can better separate two classes than W_2, W will approach W_1. Shown in Fig. 9, after projection by the calculated W, there is no overlapping between positive samples and negative samples. Hence, in the low-dimensional space, these two classes can be separated well.

A. FADA for Image Classification

In this section, we experimentally evaluate the performance of FADA on real image data sets: COREL image set and four popular face image sets in Section III. The use of ADA, LDA, BDA, and other state-of-the-art methods have been investigated on the same data set [3]. The congruent results are that ADA outperformed the other algorithms with Bayesian as the base classifier. Therefore in our experiments, we focused on comparing ADA with FADA in terms of classification accuracy and efficiency (computational time). In COREL data set, ADA searches

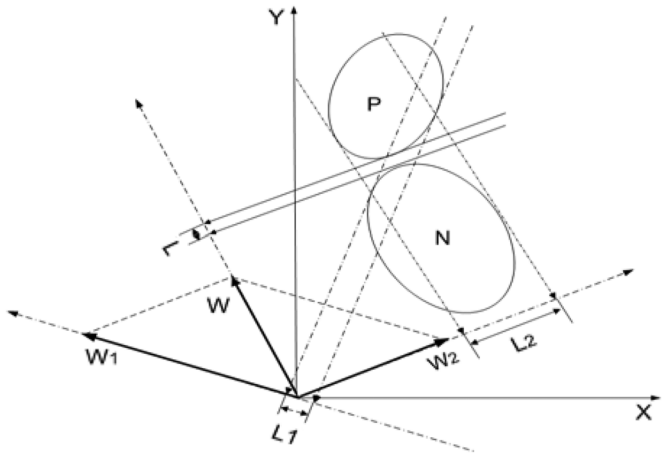


Fig. 9. Illustration of FADA algorithm.

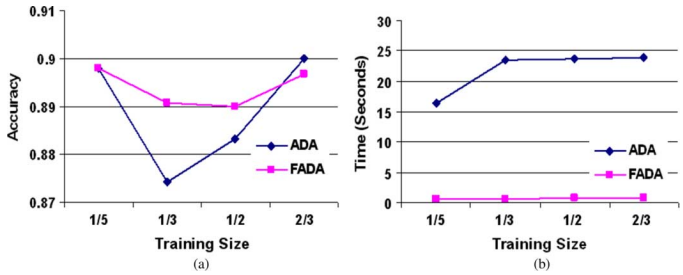


Fig. 10. Comparison of ADA and FADA with different sizes of training set.

36 parameter combinations (λ, η) sampled from 0 to 1 with step size of 0.2 to find the best one. Bayesian classifier is used on the projected data for all projection-based methods. In all experiments, average classification accuracy of 50 runs is reported. We performed our experiments using Matlab on a Pentium IV 2.26-GHz machine with 1 GB of RAM.

Fig. 10 shows the performance of ADA and FADA as the size of training samples changes from 1/5 to 2/3 of the total samples. For example, 1/5 means one-fifth of the images are used for training while the rest are used for testing. In Fig. 10(a), we find that the accuracy of FADA and ADA both change with different training sets. No matter if the training size is small or large, FADA is very comparable to ADA in terms of accuracy. Another key observation from Fig. 10(b) is that FADA is much faster than ADA by an order of 1–2. As the size of the training set increases, the speedup of FADA over ADA significantly increases because ADA spends a lot of time in training and searching. It demonstrates that FADA is a more efficient dimension reduction algorithm than ADA, as it is comparable to ADA in classification while it has much lower time costs.

B. FADA for Face Classification

To evaluate FADA for face classification, we tested it on three face image databases used in Section III. For each database, we randomly chose one person's face images as positive and the rest of the face images of others are considered as negative. In all of these data sets, ADA searches 121 various parameter combinations with the searching step size of 0.1.

Table IV shows the comparison of ADA and FADA on accuracy and efficiency, with the largest accuracy and the smallest computational time in bold. It can be seen that FADA performs

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY AND EFFICIENCY ON THREE DIFFERENT FACE DATABASES

Database	Method	ADA	FADA
Harvard Subset1	*Accuracy (%)*	89.56	**89.67**
	Time (Second)	78.67	**0.72**
Harvard Subset2	*Accuracy (%)*	88.62	**88.70**
	Time (Second)	114.34	**0.98**
Harvard Subset3	*Accuracy (%)*	88.98	**89.58**
	Time (Second)	155.93	**1.31**
ATT Database	*Accuracy (%)*	**97.88**	97.28
	Time (Second)	328.77	**2.89**
UMIST Database	*Accuracy (%)*	95.55	**95.76**
	Time (Second)	471.56	**4.31**

better in four out of five data sets on classification accuracy and at least two orders of magnitude faster than ADA in all five data sets. It is to be noted that the computation requirements of ADA increase cubically with the increase size of data sets (from Harvard to UMIST data set), and the speed difference between ADA and FADA becomes more significant with the increase of face database scale. It is proved that FADA not only reduces the computational cost but also achieves competitive classification accuracy with ADA. It is an efficient dimension reduction scheme for image classification on small or large image data sets.

V. TWO-DIMENSIONAL ADA

In most cases, image data are stored in vector format. However, recent studies indicate that matrix/tensor representation of images has become popular and widely used in face recognition and classification recently. Compared to vector representation (1-D) of the data, 2-D representation works directly with images in their native state, as two-dimensional matrices. Hence, the image does not need to be transformed, which not only saves the computational cost but also preserves all spatial information of the original image.

Since the two-dimensional linear discriminant analysis (2DLDA) was proposed in 2004, up to present, there are several variants. Li *et al.* [28] and Sanguansat *et al.* [29] presented their 2DLDA with only reducing the number of columns and keeping the number of rows unchanged. Instead, Yang *et al.* [30] presented a two-step algorithm, which first reduces the number of columns and then reduces the number of rows. Ye *et al.* [31] proposed to calculate the row and column transformation matrices in an iterative way. Fortunately, they also recommended one iteration is sufficient because the accuracy curves were stable with respect to the number of iterations T. Later, Inoue *et al.* [32] proposed two non-iterative algorithms, namely *selective* and *parallel* algorithm. However, those two algorithms are more complex than the iterative one with $T = 1$.

A. Two-Dimensional ADA

We extend 1DADA to 2DADA, which merges 2DLDA and 2DBDA in a unified framework and offers more flexibility and a richer set of alternatives to each individual method in the parametric space.

2DADA tries to find two transformation matrices $L \in \Re^{r\times l_1}$ and $R \in \Re^{c\times l_2}$ that map each $X \in \Re^{r\times c}$ from original high-dimensional space to a low-dimensional space $Y = L^T X R \in \Re^{l_1\times l_2}$, in which the most useful features are preserved [33].

Mathematically, it could be modeled as finding two optimal projections L and R that maximizes the following ratios:

$$\langle L_{opt}, R_{opt}\rangle = \arg\max_{L,R} \frac{\left|L^T\left[\lambda\cdot S^R_{N\to P} + (1-\lambda)\cdot S^R_{P\to N}\right]L\right|}{\left|L^T\left[\eta\cdot S^R_P + (1-\eta)\cdot S^R_N\right]L\right|} \tag{21}$$

in which

$$S^R_{N\to P} = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P) R R^T (\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{22}$$

$$S^R_{P\to N} = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N) R R^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{23}$$

$$S^R_P = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P) R R^T (\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{24}$$

$$S^R_N = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N) R R^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T. \tag{25}$$

Due to the difficulty of computing the optimal L and R simultaneously, we derive an iterative algorithm similar to Ye's work in [31]. Initially, $R_0 = (I_{l_2}, 0)^T$, we can compute the optimal in (26) at the bottom of the page. Next, with the computed L, calculate the optimal with (27) at the bottom of the next page. This procedure is repeated for T iterations. In real application, T can be set to be 1 according to [29].

B. 2DADA on Hand Digits Recognition

In this section, we experimentally evaluate the performance of the 2DADA algorithm on handwritten digit recognition. In all experiments, 2DADA is tested with (λ, η) evenly sampled from 0 to 1 with step size of 0.1. Besides, 10-fold cross-validation is used to report the mean accuracy of a K-NN query with $K = 10$.

$$L_{t+1} = \arg\max_{L} \frac{\left|L^T_{t+1}\left[\lambda \sum_{i\in Negative} (\mathbf{x}_i - \mathbf{m}_P) R_t R_t^T (\mathbf{x}_i - \mathbf{m}_P)^T + (1-\lambda) \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N) R_t R_t^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T\right] L_{t+1}\right|}{\left|L^T_{t+1}\left[\eta \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P) R_t R_t^T (\boldsymbol{x}_i - \boldsymbol{m}_P)^T + (1-\eta) \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N) R_t R_t^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T\right] L_{t+1}\right|} \tag{26}$$

Fig. 11. Examples of handwritten images.

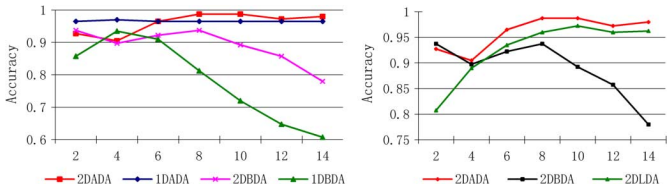


Fig. 12. Comparison of accuracy with different dimensions.

First, we tested 2DADA, 2DBDA, 2DLDA, and their 1-D-based methods on a subset of MNIST data set [34], which contains 400 similar handwritten 1's (200) and 7's (200). Some example images are shown in Fig. 11.

In this experiment, the original dimension of images 28×28 is reduced to $d \times d$ ($d < 28$) by all 2-D-based methods. Correspondingly, the reduced dimension p in their 1-D-based methods is chosen such that both 1-D and 2-D methods use the same amount of storage for the transformation matrices and the reduced presentations [31]. For examples, on the MNIST data set, $d = 2$, 4, 6, 8, 10, 12, and 14 for 2DADA are used, corresponding to $p = 2$, 6, 12, 22, 34, 49, and 67 for 1DADA.

The average accuracy rate across 10-fold cross-validation over the variation of dimension d is plotted in Fig. 12, where the x-axis denotes the values of d (between 2 to 14).

From Fig. 12, we can clearly find: 1) 2-D-based approaches (i.e., 2DADA, 2DBDA) achieve higher or comparable accuracy with their 1-D-based approach (1DADA and 1DBDA). In addition, in our experiments, we found 2-D methods are almost one order of magnitude faster than 1-D methods. It justifies that 2-D techniques have lower loss of information and are computationally efficient with the same amount of storage. 2) 2DADA consistently outperforms others irrespective of variation in dimensions. Its stableness verifies that it is a powerful dimension reduction method for classification.

VI. Conclusion and Future Work

This paper proposes ADA and its three variants for image classification. These approaches address the high dimensionality problem by applying adaptive discriminant projection in an optimal linear discriminant subspace. These proposed methods are applied on UCI benchmark data sets, four facial image data sets, and COREL color image data sets. Their superior performance demonstrates that ADA and its variants are promising, effective, and efficient approaches to image classification.

The main contributions of this work are as follows.

1) ADA provides a richer set of alternatives to classic dimension reduction methods such as LDA and BDA. As a result, it not only compensates for regularization that is afflicted by all sample-based estimation methods but also finds an optimal projection with adaptation to different sample distributions.
2) In order to improve the searching time of parameter space, we propose a novel integrated boosting framework—*i*.Boosting. It boosts not only the individual features but also a set of weak classifiers. Compared to the traditional boosting scheme, the proposed method updates both sample weights and feature weights iteratively. It obtains more performance improvement from the relevance feedback by putting human in the loop to facilitate the learning process. It has obvious advantage over the simple relevance feedback method in that the classifiers are trained to pay more attention to wrongfully predicted samples in user feedback through a reinforcement training process. With affordable computational cost, *i*.Boosting can provide a unified and stable solution to find better or close to optimal ADA prediction result.
3) The novelty of FADA lies in that without searching a parametric space, it can automatically calculate a good projection based on sample distributions information. FADA has asymptotically lower time complexity than ADA, which is desirable for large image data sets, while it achieves competitive classification accuracy with ADA.
4) 2DADA is an extension of ADA. The key difference between 2DADA and ADA is that 2DADA works on the matrix representation of images directly, while ADA uses a vector representation. 2DADA has asymptotically minimum memory requirements, e.g., the size of scatter matrix is much smaller than its vector form, and lower time complexity than ADA, which can improve the speed of image feature extraction.

Our future work includes testing different relevance feedback schemes such as active learning techniques [35] in the interactive boosting. Different base classifiers and their corresponding feature re-weighting schemes will be implemented. We will also explore correlation metric and graph embedding-based techniques [36]–[38] in the boosting and feature fusion schemes in the future.

$$R_{t+1} = \arg\max_{L} \times \frac{\left| R_{t+1}^T \left[\lambda \sum_{i \in Negative} (\mathbf{x}_i - \mathbf{m}_P) L_{t+1} L_{t+1}^T (\mathbf{x}_i - \mathbf{m}_P)^T + (1-\lambda) \sum_{i \in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N) L_{t+1} L_{t+1}^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T \right] R_{t+1} \right|}{\left| R_{t+1}^T \left[\eta \sum_{i \in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P) L_{t+1} L_{t+1}^T (\boldsymbol{x}_i - \boldsymbol{m}_P)^T + (1-\eta) \sum_{i \in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N) L_{t+1} L_{t+1}^T (\boldsymbol{x}_i - \boldsymbol{m}_N)^T \right] R_{t+1} \right|} \tag{27}$$

REFERENCES

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[2] X. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using biasMap," in *Proc. IEEE CVPR*, 2001.

[3] J. Yu and Q. Tian, "Adaptive discriminant projection for content-based image retrieval," in *Proc. Int. Conf. Pattern Recognition*, Hong Kong, Aug. 2006.

[4] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, 1999.

[5] Y. Wu and A. Zhang, "A feature re-weighting approach for relevance feedback in image retrieval," in *Proc. IEEE Int. Conf. Image Processing*, 2002.

[6] Schapire *et al.*, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.

[7] L. Breiman, Prediction Games and Arcing Algorithms, Statist. Dept., Univ. California, Tech. Rep. 504, 1997.

[8] C. Rudin *et al.*, "The dynamics of AdaBoost: Cyclic behavior and convergence of margins," *J. Mach. Learn. Res.*, vol. 5, pp. 1557–1595, 2004.

[9] J. Friedman *et al.*, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–374, 2000.

[10] Y. Rui, T. S. 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, 1998.

[11] Y. Lu, T. Zhang, and Q. Tian, "i.Boosting for image classification," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Beijing, China, Jul. 2–5, 2007.

[12] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1992.

[13] X. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *ACM Multimedia Syst. J,* Special Issue on CBIR, vol. 8, no. 6, pp. 536–544, 2003.

[14] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant EM algorithm with application to image retrieval," in *Proc. IEEE Computer Vision and Pattern Recognition*, Jun. 13–15, 2000.

[15] Q. Tian, Y. Wu, J. Yu, and T. S. Huang, "Self-supervised learning based on discriminative nonlinear features for image classification," *Pattern Recognit.*, Special Issue on Image Understanding for Digital Photographs, vol. 38, no. 6, pp. 903–917, 2005.

[16] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 23–38, 1998.

[17] COREL color image database. [Online]. Available: http://www.corel.com.

[18] Q. Tian, J. Yu, T. Rui, and T. S. Huang, "Parameterized discriminant analysis for image classification," in *Proc. Int. Conf. Multimedia and Expo*, Jun. 27–30, 2004.

[19] J. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[20] UCI Machine Learning Repository. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[21] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. SPIE Storage and Retrieval for Image and Video Databases*, San Diego, CA, 1995.

[22] J. R. Smith and S. F. Chang, "Transform features for texture classification and discrimination in large image database," in *Proc. IEEE Int. Conf. Image Processing*, Austin, TX, 1994.

[23] X. Zhou, Y. Rui, and T. S. Huang, "Water-filling algorithm: A novel way for image feature extraction based on edge maps," in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, 1999.

[24] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[25] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Applications of Computer Vision*, Sarasota, FL, Dec. 1994.

[26] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, 2003.

[27] Y. Lu, J. Ma, and Q. Tian, "FADA: An efficient dimension reduction scheme for image classification," in *Proc. Pacific-Rim Conf. Multimedia*, Hong Kong, Dec. 11–14, 2007.

[28] M. Li and B. Yuan, "2D-LDA: A novel statistical linear discriminant analysis for image matrix," *Pattern Recognit. Lett.*, vol. 26, no. 5, pp. 527–532, 2005.

[29] P. Sanguansat and W. Asdornwised *et al.*, "Two-dimensional linear discriminant analysis of principle component vectors for face recognition," in *Proc. ICASSP*, 2006, pp. 345–348.

[30] J. Yang, D. Zhang, X. Yong, and J. Yang, "Two-dimensional linear discriminant transform for face recognition," *Pattern Recognit.*, vol. 38, no. 7, pp. 1125–1129, 2005.

[31] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Advances in Neural Information Processing Systems (NIPS2004)*, 2004, vol. 17, pp. 1569–1576.

[32] K. Inoue and K. Urahama, "Non-iterative two-dimensional linear discriminant analysis," in *Proc. ICPR*, Hong Kong, 2006.

[33] Y. Lu, J. Yu, N. Sebe, and Q. Tian, "Two-dimensional adaptive discriminant analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Honolulu, HI, Apr. 16–20, 2007.

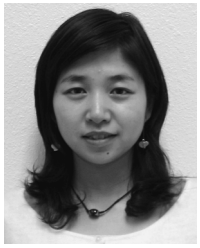
[34] Y. LeCun *et al.* [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[35] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2001, pp. 107–118.

[36] Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2229–2235, 2008.

[37] Y. Fu and T. S. Huang, "Image classification using correlation tensor analysis," *IEEE Trans. Image Process.*, vol. 17, no. 2, pp. 226–234, Feb. 2008.

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



Yijuan Lu (M'05) received the Ph.D. degree in computer science from the University of Texas at San Antonio in 2008.

She is an Assistant Professor in the Department of Computer Science, Texas State University, San Marco, TX. During 2006–2008, she was a summer Intern Researcher at FXPAL lab, Web Search & Mining Group, Microsoft Research Asia (MSRA), National Resource for Biomedical Supercomputing (NRBSC) at the Pittsburgh Supercomputing Center (PSC), Pittsburgh, PA. She was the Intern Researcher at Media Technologies Lab, Hewlett-Packard Laboratories (HP) in 2008, and research fellow of Multimodal Information Access and Synthesis (MIAS) Center at the University of Illinois at Urbana-Champaign (UIUC) in 2007. Her current research interests include multimedia information retrieval, computer vision, machine learning, data mining, and bioinformatics. She has published extensively and serves as a reviewer for top conferences and journals.

Dr. Lu is the 2007 Best Paper Candidate in Retrieval Track of Pacific-Rim Conference on Multimedia (PCM) and the recipient of 2007 Prestigious HEB Dissertation Fellowship, 2007 Star of Tomorrow Internship Program of MSRA. She is a member of ACM.

Qi Tian (SM'03) received the B.E. degree from Tsinghua University, Beijing, China, in 1992 and the Ph.D. degree in electrical and computer engineering from University of Illinois at Urbana-Champaign (UIUC) in 2002.

He is currently an Associate Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA). He has been taking Faculty Leave at Microsoft Research Aisa (MSRA) since fall 2007. He was a Visiting Scholar at UIUC MIAS center (2007), a Visiting Researcher at MSRA (summer 2007), a Visiting Professor in NEC Laboratories America, Inc. (summer 2003), and a Visiting Researcher (2001) in MERL, Cambridge, MA. His research interests include multimedia information retrieval and computer vision. He has published over 80 refereed book chapters, journal, and conference papers in these fields. His research projects were funded by ARO, DHS, HP Lab, SALSI, CIAS, and the Chinese Academy of Science.

Dr. Tian was the coauthor of a Best Student Paper in ICASSP 2006 and coauthor of a Best Paper Candidate in PCM 2007. He was nominated for 2008 UTSA President Distinguished Research Award. He has been serving as Program Chairs and Organization Committee Member for ACM Multimedia (2009), CIVR (2010), ACM ICIMCS (2009), ACM LSMMRM (2009), MMM (2010), VIP (2007, 2008), IMAI 2007, MIR (2005), and Session Chairs and TPC members for over 100 IEEE and ACM Conferences including ACM Multimedia, SIGIR, ICCV, ICME, ICASSP, ICPR, MIR, VCIP, and PCM. He has been a Guest Co-Editors for the IEEE TRANSACTIONS ON MULTIMEDIA, *Journal of Computer Vision and Image Understanding*, and *EURASIP Journal on Advances in Signal Processing* and is on the Editorial Board of *Journal of Multimedia*. He has been a member of ACM since 2004.