Ensemble-Based Discriminant Learning With Boosting for Face Recognition

Juwei Lu, *Member, IEEE*, K. N. Plataniotis, *Senior Member, IEEE*, A. N. Venetsanopoulos, *Fellow, IEEE*, and Stan Z. Li

***Abstract*—In this paper, we propose a novel ensemble-based approach to boost performance of traditional *Linear Discriminant Analysis* (LDA)-based methods used in face recognition. The ensemble-based approach is based on the recently emerged technique known as "boosting." However, it is generally believed that boosting-like learning rules are not suited to a strong and stable learner such as LDA. To break the limitation, a novel weakness analysis theory is developed here. The theory attempts to boost a strong learner by increasing the diversity between the classifiers created by the learner, at the expense of decreasing their margins, so as to achieve a tradeoff suggested by recent boosting studies for a low generalization error. In addition, a novel distribution accounting for the pairwise class discriminant information is introduced for effective interaction between the booster and the LDA-based learner. The integration of all these methodologies proposed here leads to the novel ensemble-based discriminant learning approach, capable of taking advantage of both the boosting and LDA techniques. Promising experimental results obtained on various difficult face recognition scenarios demonstrate the effectiveness of the proposed approach. We believe that this work is especially beneficial in extending the boosting framework to accommodate general (strong/weak) learners.**

***Index Terms*—Boosting, face recognition (FR), linear discriminant analysis, machine learning, mixture of linear models, small-sample-size (SSS) problem, strong learner.**

I. Introduction

A. Face Recognition

FACE RECOGNITION (FR) has a wide range of applications, such as face-based video indexing and browsing engines, biometric identity authentication, human-computer interaction, and multimedia monitoring/surveillance. Within the past two decades, numerous FR algorithms have been proposed, and detailed surveys of the developments in the area have appeared in the literature [1]–[6]. Among various FR methodologies used, the most popular are the so-called appearance-based approaches, which include the three most well-known FR methods, namely Eigenfaces [7], Fisherfaces [8], and Bayes Matching [9]. With focus on low-dimensional statistical feature extraction, the appearance-based approaches generally operate directly on appearance images of face object and process them as two-dimensional (2-D) holistic patterns to avoid difficulties associated with three-dimensional (3-D) modeling, and shape or landmark detection [5]. Of the appearance-based FR methods, those based on linear discriminant analysis (LDA) have shown promising results as it is demonstrated in [8], [10]–[15]. However, statistical learning methods such as the LDA-based ones often suffer from the so-called "small-sample-size" (SSS) problem [16], encountered in high-dimensional pattern recognition tasks where the number of training samples available for each subject is smaller than the dimensionality of the samples. For example, in the experiments reported here only $L \in [1, 7]$ training samples per subject are available while the dimensionality of the sample space is up to $J = 17\,154$. In addition, the performance of linear appearance-based methods including LDA often deteriorates rapidly when face patterns are subject to large variations in viewpoints, illumination or facial expression. These variations result in a highly nonconvex and complex distribution of face images [17]. Thus, the limited success of these methods should be attributed to their linear nature.

In general, a nonconvex distribution can be handled either by globally nonlinear models or by a mixture of locally linear models (or ensemble-based methods as they are known in the machine learning literature [18]). Globally nonlinear methods are not without problems. Approaches such as those based on kernel machines [19]–[26] require the optimization of many design parameters, tend to overfit easily due to the increased algorithmic complexity, and they are computationally expensive compared to their linear counterparts. The last point is particularly important for tasks such as face recognition, which are performed in a high-dimensional input space. On the other hand, ensemble-based approaches embody the principle of "divide and conquer," by which a complex recognition task is decomposed into a set of simpler ones, in each of which a locally linear pattern distribution can be generalized and dealt with by a relatively simple linear solution. As such, the ensemble-based methods are simpler, easier to implement, and more cost effective compared to the nonlinear ones. However, most existing ensemble-based FR methods are developed based on traditional cluster analysis [27]–[30]. As a consequence, a disadvantage to classification tasks is that the submodels' division/combination criteria used in these clustering techniques are not directly related to the *classification error rate* (CER) of the resulting classifier, especially the true CER (often referred to as the generalization error rate).

Manuscript received March 23, 2004; revised December 24, 2004.

This work was supported in part by the Bell University Laboratories at the University of Toronto.

J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos are with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, ON M5S 3G4 Canada (e-mail: kostas@dsp.toronto.edu).

Stan Z. Li is with the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, P.R. China.

Digital Object Identifier 10.1109/TNN.2005.860853

B. Ensemble-Based Learning With Boosting

Recently, a machine-learning technique known as "boosting" has received considerable attention in the pattern recognition community, due to its usefulness in designing ensemble-based classifiers [31], [32]. The idea behind boosting is to sequentially employ a base classifier on a weighted version of the training sample set to generalize a set of classifiers of its kind. Often the base classifier is also called "learner." These weights are updated at each iteration through a classification-error-driven mechanism. Although any individual classifier produced by the learner may perform slightly better than random guessing, the formed ensemble can provide a very accurate (strong) classifier. It has been shown, both theoretically and experimentally, that boosting is particularly robust in preventing overfitting and reducing the generalization error by increasing the so-called *margins* of the training examples [32]–[35]. The margin is defined as the minimal distance of an example to the decision surface of classification [36]. For a classifier, a larger expected margin of training data generally leads to a lower generalization error.

Since its introduction, AdaBoost became known as the most accurate general purpose classification algorithm available [37]. However, the machine-learning community generally regards ensemble-based learning rules, including boosting and bagging [38], not suited to a strong and stable learner, such as LDA [35], [39]. The reason behind this belief is that the effectiveness of these rules depends, to a great extent, on the learner's "instability," which means that small changes in the training set could cause large changes in the resulting classifier [35]. On the other hand, it has been found in practical applications that boosting may fail given a too weak learner [32]. In recent boosting studies, Murua [40] introduced a useful notion of weak dependence between classifiers constructed with the same training data, and proposed an interesting upper bound on the generalization error with respect to the margins of the classifiers and their dependence. Murua's bound reveals that to achieve a low generalization error, the boosting procedure should not only create the classifiers with large expected margins, but also keep their dependence low or weak. This suggests in theory that there exists a tradeoff between the large margins and the weak dependence.

The requirement for an appropriately weak learner significantly restricts the applicability of the boosting algorithms in practical applications, given the fact that most of state-of-the-art recognition methods involve the utilization of a strong learner. Therefore, it is highly desirable to improve the traditional boosting frameworks, so that they are capable of accommodating more general learners in both the pattern recognition and machine learning communities.

C. Overview of the Contributions

In this paper, a novel weakness analysis theory is developed to overcome the limitation of the weak learners, which are necessary in existing boosting algorithms. To this end, a new variable called "learning difficulty degree" (LDD) is introduced along with a cross-validation method. They are used to analyze and appropriately regulate the weakness of the classifiers generalized by a strong learner via the training data. In addition, a new loss function with respect to the LDD is proposed to quantitatively estimate the generalization power of these produced classifiers. This is achieved in the loss function by balancing the averaged empirical error of the classifiers and their mutual dependence. They are two key factors to the generalization error of the formed ensemble classifier as shown in Murua's theory [40].

The proposed weakness analysis theory is applied to boost the performance of the traditional LDA-based approaches in complex FR tasks. Thus, the learners in this work are the LDA-based ones, which differ from the traditional learners used in boosting at two aspects: 1) They are rather strong and stable and 2) they are feature extractors rather than *pure* classifiers. The latter makes this work similar in spirit to those of Viola, Tieu and Jones [41]–[43], where the boosting process is viewed as a feature selection process. Particularly, to boost the specific LDA-based learners, a new variable called "pairwise class discriminant distribution" (PCDD) is also introduced to build an effective interaction mechanism between the booster and the learner. As a result, a novel ensemble-based discriminant learning method is developed here under the boosting framework through the utilization of the PCDD and the weakness analysis theory. In the proposed method, each round of boosting generalizes a new LDA subspace particularly targeting those examples from the hard-to-separate pairs of classes indicated by its preceding PCDD, so that the separability between these classes is enhanced in the new LDA subspace. The final result obtained by the process is an ensemble of multiple relatively weak but very specific LDA solutions. The ensemble-based solution is able to take advantage of both boosting and LDA. It is shown by the FR experiments to outperform the single solutions created by the LDA-based learners in various difficult learning scenarios, which include the cases with different SSS settings and the case with increased nonlinear variations.

The rest of the paper is organized as follows. In Section II, we briefly review the AdaBoost approach and its multiclass extensions. Then, in Section III, the theory and algorithm of how to boost a LDA-based strong learner are introduced and described in detail. Section IV reports on a set of experiments conducted on the FERET face database to demonstrate the effectiveness of the proposed methodologies. Finally, conclusions are summarized in Section V. In addition, a brief introduction to the adopted LDA-based learners is given in Appendix I.

II. Related Work

Since the boosting method proposed here is developed from AdaBoost [31], we begin with a brief review of the algorithm and its multiclass extensions.

In the case of pattern classification, the task of learning from examples can be formulated in the following way: Given a training set, $\mathcal{Z} = \{\mathcal{Z}_i\}_{i=1}^{C}$, containing C classes with each class $\mathcal{Z}_i = \{(\mathbf{z}_{ij}, y_{ij})\}_{j=1}^{L_i}$ consisting of a number of examples $\mathbf{z}_{ij}$ and their corresponding class labels y_{ij}, a total of $N = \sum_{i=1}^{C} L_i$ examples are available in the set. Let $\mathbb{Z}$ be the sample space: $\mathbf{z}_{ij} \in \mathbb{Z}$, and $\mathbb{Y} = \{1, \ldots, C\}$ be the label set: $y_{ij}(=i) \in \mathbb{Y}$. Taking as input such a set $\mathcal{Z}$, the objective of

Input: A set of training images $\mathcal{Z} = \{(\mathbf{z}_{ij}, y_{ij})_{j=1}^{L_i}\}_{i=1}^{C}$ with labels $y_{ij} = i \in \mathbb{Y}$, where $\mathbb{Y} = \{1, \cdots, C\}$; a LDA-style learner; and the iteration number, T.

Let $B = \{(\mathbf{z}_{ij}, y) : \mathbf{z}_{ij} \in \mathcal{Z}, \mathbf{z}_{ij} \in \mathbb{Z} = \mathbb{R}^J, y \in \mathbb{Y}, y \neq y_{ij}\}$.

Initialize $\Upsilon_1(\mathbf{z}_{ij}, y) = \frac{1}{|B|} = \frac{1}{N(C-1)}$, the mislabel distribution over B.

(For simplicity, we denote the LDA-based feature extractor as a function $\mathcal{L}(\cdot)$, which has $(\Psi_t, \{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C}) = \mathcal{L}(\mathcal{R}_t, \hat{D}_t, A_t)$. See Appendix I for details.)

Do for $t = 1, \cdots, T$:

1. Update the pseudo sample distribution: $\hat{D}_t(\Upsilon_t)$, and the PCDD: A_t with Eq.2.
2. **If** $t = 1$ then
   randomly choose r samples per class to form a learning set $\mathcal{R}_1 \subset \mathcal{Z}$.
   else choose r hardest samples per class based on $\hat{D}_t$ to form $\mathcal{R}_t \subset \mathcal{Z}$.
3. Train a LDA-style feature extractor with $\mathcal{L}(\mathcal{R}_t, \hat{D}_t, A_t)$ to obtain $(\Psi_t, \{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C})$.
4. Build a gClassifier $h_t = d(\Psi_t, \{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C})$ with Eq.8, apply it into the entire training set $\mathcal{Z}$, and get back corresponding hypotheses, $h_t : \mathbb{R}^J \times \mathbb{Y} \to [0, 1]$.
5. Calculate the pseudo-loss produced by h_t:
   $$\hat{\epsilon}_t = \frac{1}{2} \sum_{(\mathbf{z}_{ij}, y) \in B} \Upsilon_t(\mathbf{z}_{ij}, y) \left(1 - h_t(\mathbf{z}_{ij}, y_{ij}) + h_t(\mathbf{z}_{ij}, y)\right).$$
6. Set $\beta_t = \hat{\epsilon}_t / (1 - \hat{\epsilon}_t)$. If $\beta_t = 0$, then set $T = t - 1$ and abort loop.
7. Update the mislabel distribution Υ_t:
   $$\Upsilon_{t+1}(\mathbf{z}_{ij}, y) = \Upsilon_t(\mathbf{z}_{ij}, y) \cdot \beta_t^{(1 + h_t(\mathbf{z}_{ij}, y_{ij}) - h_t(\mathbf{z}_{ij}, y))/2}.$$
8. Normalize Υ_{t+1} so that it is a distribution,
   $$\Upsilon_{t+1}(\mathbf{z}_{ij}, y) \leftarrow \frac{\Upsilon_{t+1}(\mathbf{z}_{ij}, y)}{\sum_{(\mathbf{z}_{ij}, y) \in B} \Upsilon_{t+1}(\mathbf{z}_{ij}, y)}.$$

Output the final composite gClassifier,

$$h_f(\mathbf{z}) = \arg\max_{y \in \mathbb{Y}} \sum_{t=1}^{T} \left(\log \frac{1}{\beta_t}\right) h_t(\mathbf{z}, y).$$

Fig. 1. Algorithm of boosting a LDA-style learner (simply replacing the learner with JD-LDA or EFM in step 3 to obtain B-JD-LDA or B-EFM). Either $A_t(p, q)$ or $\hat{A}_t(p, q)$ can be used to replace each other during the boosting process.

learning is to estimate a function or classifier $h(\mathbf{z}) : \mathbb{Z} \to \mathbb{Y}$, such that h will correctly classify unseen examples $(\mathbf{z}, y)$.

To this end, AdaBoost works by repeatedly applying a given weak learner to a weighted version of the training set in a series of rounds $t = 1, \ldots, T$, and then linearly combining these weak classifiers $\{h_t\}_{t=1}^{T}$ constructed in each round into a single strong classifier h_f. The most interesting feature of AdaBoost is its surprising ability to reduce the amount of overfitting and the generalization error of classification, even as T becomes large [31], [34]. To explain the property, quite a number of perspectives on AdaBoost have emerged since its introduction [44]. The dominant amongst them is the margin theory, which regards AdaBoost to be an efficient method for maximizing the margin [34]. However, many researchers have shown that the margin theory provides only partial answers to the puzzle [45], [46]. As a result, AdaBoost still remains as a mysterious algorithm, which is considered one of the most important unsolved problems in machine learning [37]. On the other hand, the limitation in the theoretical explanation does not seem to hamper the success of AdaBoost-style approaches in practical applications. For example, Viola and Jones [43] build the first real-time face detection system by using AdaBoost, which is considered a dramatic breakthrough in the face detection research.

AdaBoost is originally developed to support binary classification tasks. Its multiclass extensions include two variants, AdaBoost.M1 and AdaBoost.M2 [31]. AdaBoost.M1 is the most straightforward generalization. However, the algorithm halts if the classification error rate (CER) of the weak classifier h_t produced in any iterative step is $\geq 50\%$. Research indicates that this limitation often terminates the procedure too early, resulting in insufficient classification capabilities [32], [34]. To avoid the problem, rather than the ordinary CER, AdaBoost.M2 attempts to minimize a more sophisticated error measure called "pseudoloss," $\hat{\epsilon}_t$, which is expressed as

$$\hat{\epsilon}_t = \frac{1}{2} \sum_{(\mathbf{z}_{ij}, y) \in B} \Upsilon_t(\mathbf{z}_{ij}, y)(1 - h_t(\mathbf{z}_{ij}, y_{ij}) + h_t(\mathbf{z}_{ij}, y)) \tag{1}$$

where $\Upsilon_t(\mathbf{z}_{ij}, y)$ (see steps 7,8 of Fig. 1 for definition) is the so-called "mislabel distribution" defined over the set of all mislabels: $B = \{(\mathbf{z}_{ij}, y) : \mathbf{z}_{ij} \in \mathcal{Z}, \mathbf{z}_{ij} \in \mathbb{R}^J, y \in \mathbb{Y}, y \neq y_{ij}\}$. With the pseudoloss, the boosting process can continue as long as the weak classifier produced has pseudoloss slightly better than random guessing. In addition, the introduction of the mislabel distribution enhances the communication between the

learner and the booster. In this way, AdaBoost.M2 can focus the learner not only on hard-to-classify examples, but more specifically, on the incorrect labels [31]. For all these reasons, we develop the ensemble-based discriminant algorithm proposed in the next section following the AdaBoost.M2 paradigm.

There are two LDA-based FR approaches (or learners) that are boosted in this work. One is the so-called "Enhanced Fisher LDA Model" (hereafter EFM) [13], and the other is called "Revised Direct LDA" (hereafter JD-LDA) [47] proposed by the authors recently. The EFM method is an improvement of the Fisherfaces method [8], while the JD-LDA method is a LDA variant introduced specifically for face recognition in high-dimensional, small-sample-size scenarios. For completeness, the details of the two learners are described in Appendix I. Compared to traditional learners used in the boosting algorithms, the two LDA-based learners should be emphasized again at the following two points. 1) They are strong and stable learners, which can be successfully used as stand-alone procedures in FR tasks [13], [47], [48]. That obviously contradicts the general belief that boosting solutions should operate only on top of weak learners. 2) The EFM or JD-LDA learner is composed of a LDA-based feature extractor and a nearest center classifier. As it can be seen in Appendix I, the learning focus of such a learner is on the feature extractor rather than the classifier. It is rather different at this point from the original boosting design where the weak learners are used only as *pure* classifiers without concerning feature extraction. This makes the AdaBoost learning tend to be an adaptively feature selection process, some of the ideas seen in [43]. Therefore, accommodating a learner such as JD-LDA or EFM requires a generalized boosting framework, which is not restricted by the assumption of the weak learner availability. To highlight these difference, we call "*gClassifier*" the more general classifier produced by the LDA-based learners in the rest of the paper.

III. Boosting a LDA-Style Learner

A. Interaction Between the LDA Learner and the Booster

To boost a learner, we first have to build a strong connection between the learner and the boosting framework. In AdaBoost, this is implemented by manipulating the so-called "sample distribution," which is a measure of how hard to classify an example. However, we need a more specific connecting variable in this work, given the fact that the nature of LDA is a feature extractor, which goal is to find a linear mapping to enhance the between-class separability of the samples under learning. For this purpose, a new distribution called "pairwise class discriminant distribution" (PCDD), A_{pq}, is introduced here. The PCDD is developed from the mislabel distribution Υ_t of AdaBoost.M2. Defined on any one pair of classes $\{(p,q) : p,q \in \mathbb{Y}\}$, the PCDD can be computed at the tth iteration as (2), shown at the bottom of the page, where L_p and L_q are the number of elements in classes $\mathcal{Z}_p$ and $\mathcal{Z}_q$, respectively. As it is known from the AdaBoost.M2 developments, the mislabel distribution $\Upsilon_t(\mathbf{z}_{ij}, y)$ indicates the extent of difficulty in distinguishing the example $\mathbf{z}_{ij}$ from the incorrect label y based on the feedback information from the preceding $(t-1)$ gClassifiers. Thus, $A_t(p,q)$ can be intuitively considered as a measure of how important it is to discriminate between the classes p and q when designing the current gClassifier h_t. Obviously, a larger value of $A_t(p,q)$ implies worse separability between the two classes. It is, therefore, suitable to drive a LDA-based learner through $A_t(p,q)$, so that it is focused specifically on the hard-to-separate *pairs* of classes. To this end, rather than the ordinary definition of the between-class scatter matrix $\mathbf{S}_b(= (1/N)\sum_{i=1}^{C} L_i(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})^T$ where $\bar{\mathbf{z}}_i = (1/L_i)\sum_{j=1}^{L_i}\mathbf{z}_{ij}$ is the mean of the class $\mathcal{Z}_i$ and $\bar{\mathbf{z}} = (1/N)\sum_{i=1}^{C}\sum_{j=1}^{L_i}\mathbf{z}_{ij}$ is the average of the ensemble $\mathcal{Z}$), we introduce a variant of $\mathbf{S}_b$, which can be expressed as

$$\hat{\mathbf{S}}_{b,t} = \sum_{p=1}^{C}\phi_p\phi_p^T, \quad \text{with}$$
$$\phi_p = (L_p/N)^{1/2}\sum_{q=1}^{C}A_t^{1/2}(p,q)(\bar{\mathbf{z}}_p - \bar{\mathbf{z}}_q). \tag{3}$$

It should be noted at this point that the variant $\hat{\mathbf{S}}_{b,t}$ weighted by A_t embodies the design principle behind the so-called "fractional-step" LDA presented in [49]. According to this principle, object classes that are difficult to be separated in the low-dimensional output spaces $(\Psi_1, \ldots, \Psi_{t-1})$ generalized in previous rounds can potentially result in misclassification. Thus, they should be paid more attention by being more heavily weighted in the high-dimensional input space of the current (tth) round, so that their separability is enhanced in the resulting feature space Ψ_t. It can be easily seen that the variant $\hat{\mathbf{S}}_{b,t}$ reduces to $\mathbf{S}_b$ when $A_t(p,q)$ is equal to a constant.

Similarly, the weighted version of the within-class scatter matrix $\mathbf{S}_w$ can be given as follows:

$$\hat{\mathbf{S}}_{w,t} = N\cdot\sum_{i=1}^{C}\sum_{j=1}^{L_i}\hat{D}_t(\mathbf{z}_{ij})(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)^T \tag{4}$$

where $\hat{D}_t(\mathbf{z}_{ij}) = \sum_{y\neq y_{ij}}\Upsilon_t(\mathbf{z}_{ij}, y)$ is defined over $\mathcal{Z}$ as the sample distribution, similar to the one given in AdaBoost. Since $\hat{D}_t(\mathbf{z}_{ij})$ is derived indirectly from the pseudoloss $(\hat{\epsilon})$, we call $\hat{D}_t(\mathbf{z}_{ij})$ a "pseudo sample distribution" for the distinguishing purpose. It can be seen that a larger value of $\hat{D}_t(\mathbf{z}_{ij})$ implies a harder-to-classify example for those preceding gClassifiers.

Recently, it is shown that to achieve a low generalization error, the boosting procedure should not only create classifiers with large expected margins, but also keep their dependence low or weak [40]. Obviously, classifiers trained with more overlapping examples will result in stronger dependence among them. A way to avoid building similar gClassifiers

$$A_t(p,q) = \begin{cases} \frac{1}{2}\left(\sum_{j=1}^{L_p}\Upsilon_t(\mathbf{z}_{pj}, q) + \sum_{j=1}^{L_q}\Upsilon_t(\mathbf{z}_{qj}, p)\right), & \text{if } p\neq q \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

repeatedly is to artificially introduce some randomness in the construction of the training data. To this end, a modified PCDD is introduced as (5), shown at the bottom of the page, where $g_t(\mathbf{z}) = \arg\max_{y\in\mathbf{Y}} h_t(\mathbf{z}, y)$. As a result of using $\hat{A}_t(p,q)$ instead of $A_t(p,q)$, it can be seen that only those subject sets $\mathcal{Z}_i$ that include the examples mislabeled by the last gClassifier h_{t-1} are contributing to the construction of the current gClassifier h_t (through $\hat{\mathbf{S}}_{b,t}$). By manipulating $\hat{A}_t(p,q)$, we can reduce the extent of overlapping between the training examples used to build different gClassifiers, and thus achieve the goal of weakening the dependence among these gClassifiers. Also, this has the effect of forcing every gClassifier to focus only on the hard-to-separate pairs of classes suggested by its preceding gClassifier, resulting in a more diverse committee of gClassifiers to be generalized in the end. On the other hand, the classification ability of the individual gClassifier h_t is weakened to some extent due to less training examples involved in its construction. This weakening results in decrease in the examples' margins. However, it should be noted at this point that there appears to be a tradeoff between weak dependence and large expected margins to achieve a low generalization error [40]. Our experimentation indicates that in some cases, the utilization of $\hat{A}_t(p,q)$ may yield a better balance than that obtained by $A_t(p,q)$, improving the classification performance.

Based on the introduction of $A_t(p,q), \hat{A}_t(p,q), \hat{D}_t(\mathbf{z}_{ij}), \hat{\mathbf{S}}_{b,t}$ and $\hat{\mathbf{S}}_{w,t}$, we can now give a new boosting solution, depicted in Fig. 1, from which it can be seen that the LDA-style learner at every iteration is tuned to conquer a particular subproblem generalized by the feedback Υ_t in a manner similar to "automatic gain control," and the final solution is a mixture of T LDA subspaces by weighted linear combination. Either JD-LDA or EFM can be adopted as the LDA learner in the step 3 during the boosting process. In the remainder of the paper, we call "B-JD-LDA" the algorithm utilizing JD-LDA, while "B-EFM" indicates the one employing EFM.

B. A Cross-Validation Mechanism to Weaken the Strong Learner

As we mentioned earlier, JD-LDA or EFM itself has been a rather strong and stable learner in terms of classification ability. As a consequence, two problems are often encountered: 1) gClassifiers created exhibit a high similarity or mutual dependence, given the same training data; 2) the pseudoloss $\hat{\epsilon}_t = 0$ is often obtained halting the boosting process too early. To solve the problems, we have to artificially weaken the gClassifiers and increase their diversity accordingly. Generally speaking, the learning capacity of any LDA-like algorithm is directly proportional to the number of training examples per subject, L, and reciprocally proportional to the number of the subjects, C. Combining the two factors, we can define a variable called *Learning Difficulty Degree* (LDD): $\rho = (L/C)$, to roughly estimate the degree of difficulty for the discriminant

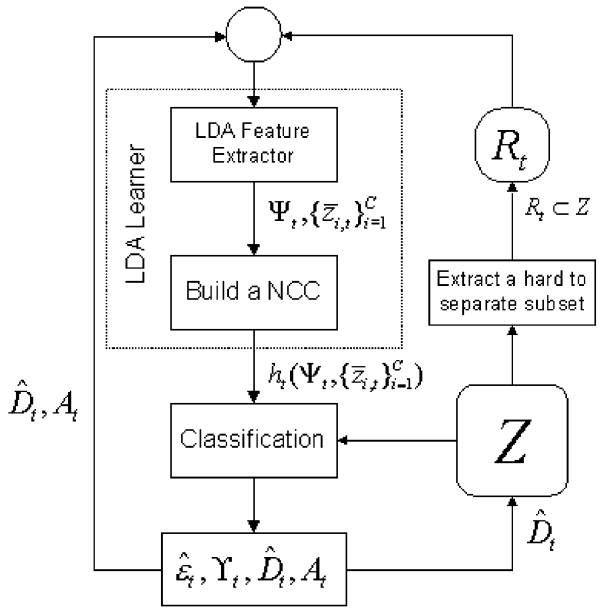

Fig. 2. Flow chart of the cross-validation mechanism embedded in the proposed boosting framework to weaken the LDA-style learner. The flow chart is based on one iteration, and the NCC denotes the nearest center classifier.

learning task on hand. It should be noted that the average $L = (1/C)\sum_{i=1}^{C} L_i$ is considered as subjects are allowed to have different number of training examples, L_i. Obviously, a smaller ρ value implies a more difficult learning task. In other words, if a learner is trained with different sample sets, the classification strength of the obtained gClassifiers will be different: *A sample set with a smaller ρ value leads to a weaker gClassifier*. Thus, from the training data point of view, the LDD provides a qualitative measure of the weakness of the gClassifiers created by the same learner. For the purpose of distinguishing the two meanings, we denote the LDD as ρ_t when it is used to express the degree of difficulty for a learning task, while ρ_l denotes the weakness of a gClassifier.

Based on the above analysis, we can introduce into the proposed boosting framework the cross-validation mechanism depicted in Fig. 2 With the mechanism in place, only a subset of the entire training set $\mathcal{Z}, \mathcal{R}_t \subset \mathcal{Z}$, is used to train the LDA-style learner. The subset $\mathcal{R}_t$ is formed in each iteration by choosing the $r \leq L$ hardest-to-classify examples per class based on current values of $\hat{D}_t(\mathbf{z}_{ij})$. Please note that $|\mathcal{R}_t| = C \cdot r$, where $|\mathcal{R}_t|$ denotes the size of $\mathcal{R}_t$. In the sequence, the obtained LDA feature extractor $\mathcal{L} : (\Psi_t, \{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C}) = \mathcal{L}(\mathcal{R}_t, \hat{D}_t, A_t \text{ or } \hat{A}_t)$ (see Appendix I for details of $\mathcal{L}(\cdot)$) are used to build a gClassifier, $h_t = h(\Psi_t, \{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C})$ based on the nearest center rule. The gClassifier is applied to the entire training set $\mathcal{Z}$ including those unseen, to the learner, examples $(\mathcal{Z} - \mathcal{R}_t)$. All the variables defined on $\mathcal{Z}$ such as $\hat{\epsilon}_t, \Upsilon_{t+1}, \hat{D}_t$, and A_t(or $\hat{A}_t$) are then reported and used in the next iteration. The detailed implementation steps of the mechanism have been embedded in Fig. 1.

$$\hat{A}_t(p,q) = \begin{cases} \frac{1}{2}\left(\sum_{j:g_t(\mathbf{z}_{pj})=q} \hat{D}_t(\mathbf{z}_{pj}) + \sum_{j:g_t(\mathbf{z}_{qj})=p} \hat{D}_t(\mathbf{z}_{qj})\right), & \text{if } p \neq q \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

It can be seen that under the proposed strategy, the LDD (ρ_t) value of the sample set used to train the strong learner decreases to (r/C) from (L/C) (note: $r \leq L$) in each iteration. Following the weakness analysis described above, this equivalently weakens the gClassifiers produced by the learner. At the same time, since each iteration feeds the learner a different subset of the entire training set, this essentially increases the diversity among these gClassifiers. Also, it should be added at this point that one of side-effects of using only r examples per subject during the construction of each gClassifier is obtaining a better estimate of the pseudoloss $\hat{\epsilon}_t$. This is achieved by using what Leo Breiman calls the "out-of-bag" samples (those samples not used during the training of the classifier) to estimate the error rate [50]. Hence finding the optimal r also provides a balance between good classifier performance and an improved estimate of the misclassification.

C. Estimation of Appropriate Weakness

The cross-validation mechanism introduced above greatly enhances the strength of the proposed boosting algorithm, but also raises the problem of model selection, that is, the determination of the optimal $\rho_l(r)(=r/C)$. As we know from the analysis in last section, a smaller/larger ρ_l value will equivalently lead to a weaker/stronger gClassifier, given the same learner. However, boosting may fail when either too weak (e.g., $r = 2$) or too strong (e.g., $r = L$) gClassifiers are constructed for combination [32]. Consequently, we can conjecture that a gClassifier with appropriate weakness should have a ρ_l value in between $(2/C)$ and (L/C). Intuitively, it is reasonable to further assume that a stronger gClassifier should lead to a lower empirical CER, while a learner, trained on a smaller fraction of the training set i.e., a smaller size $\mathcal{R}_t$), should generalize a weaker but more diverse committee of gClassifiers with each one having a more honest estimate of misclassification. Thus, a sort of loss function with respect to r that balances the two factors can be used to drive the model selection process. The proposed here function is defined as

$$\mathbf{R}(r) = \left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i,j} \Pr[h_{t,r}(\mathbf{z}_{ij}) \neq y_{ij}] \right) + \lambda \cdot \sqrt{\frac{\rho_l(r)}{\rho_l(L)}} \tag{6}$$

where $\sum_{i,j} \Pr[h_{t,r}(\mathbf{z}_{ij}) \neq y_{ij}]$ is the empirical CER obtained by applying the gClassifier h_t constructed by $\mathcal{L}(\mathcal{R}_{t,r})$ to the training set $\mathcal{Z}, \rho_l(r) = (r/C), \rho_l(L) = (L/C)$, and λ is a regularization parameter that controls the tradeoff between the weakness and the diversity of the gClassifiers. It can be seen that the tradeoff embodied in (6) implements the design principles described earlier in the sense that in order to compensate for high empirical error, the gClassifiers should have low mutual dependence, and *vice versa*. With the introduction of the loss, determining the set of gClassifiers with the optimal $\rho_l(r)$ value is equivalent to minimizing $\mathbf{R}(r)$ with respect to r. As will be seen in the experiments reported here, the estimation results through $\mathbf{R}(r)$ look rather accurate across various settings of the parameters (r, L).

In this paper, the weakness analysis theory, including the cross-validation mechanism of weakening a strong learner and the subsequent estimation method of appropriate weakness, is developed for the LDA-style learners. However, it can be seen from the previous presentations that both the two methods are dependent only on the training set, where each subject is required to have at least two examples. As a result, a traditional boosting framework enhanced with the weakness analysis theory is applicable to work with any general (weak/strong) learners. This exhibits a considerably promising approach to break the traditional limitation of the weak learners in the boosting literature.

IV. Experimental Results

We evaluate the performance of the proposed boosting methodology by applying it to a challenging pattern classification task, namely face recognition. Due to space limitations, only the results of B-JD-LDA and B-EFM with the PCDD $A_t(p, q)$ are reported here (The interested readers can refer to [51] for the results obtained with $\hat{A}_t(p, q)$).

A. The Face Database and FR Evaluation Design

To show the high complexity of the face pattern distribution, the two evaluation databases used in the experiments are taken from the well-known FERET database, which has been considered the largest, most comprehensive and representative face database to be used for evaluating the state-of-the-art in face recognition [52], [53]. The evaluation databases are constructed in two stages. First, for the purpose of preprocessing, we find in the FERET database all grayscale face images that are supplied along with the coordinate information of eyes, nose tip and mouth center to form a set $\mathcal{A}$. The set $\mathcal{A}$ contains in total 3817 face images of 1200 subjects. In the sequence, the first evaluation database denoted as $\mathcal{G}_1$ is formed by choosing in the set $\mathcal{A}$ all (1051) images of 104 subjects with each one having at least seven images. Thus we can generalize a set of SSS learning tasks, ranging from $\rho_t(L) = (3/C)$ to $\rho_t(L) = (7/C)$, to study the corresponding performance changes of the boosting algorithms. Similarly, the second evaluation database denoted as $\mathcal{G}_2$ (including $\mathcal{G}_1$) is constructed by choosing all (1703) images of 256 subjects in $\mathcal{A}$ with at least four images per subject. This database $\mathcal{G}_2$ is utilized to test the learning capacity of the algorithms as the size of the evaluation database becomes larger. The details of the images included in $\mathcal{G}_1$ and $\mathcal{G}_2$ are depicted in Table I, where the naming convention for the imagery categories can be found in [53], [54].

The original images in the FERET database are raw face images that include not only the face, but also some irrelevant, for face recognition, data, such as hair, neck, shoulder and background, as shown in Fig. 3: Left. To avoid incorrect evaluations [55], we follow the preprocessing sequence recommended in [53], which includes four steps: 1) Images are translated, rotated and scaled (to size 150×130) so that the centers of the eyes are placed on specific pixels; 2) a standard mask, as shown in Fig. 3: Middle is applied to remove the nonface portions; 3) histogram equalization is performed in the masked facial pixels; 4) face data are further normalized to have zero mean and unit standard deviation. Figs. 3: Right and 4 depict some examples after the

TABLE I
NUMBER OF IMAGES DIVIDED INTO THE STANDARD FERET IMAGERY CATEGORIES IN THE EVALUATION DATABASES, AND THE POSE ANGLE (DEGREE) OF EACH CATEGORY

Category	fa	fb	ba	bj	bk	ql	qr	rb	rc	sum
$\mathcal{G}_1$	521	292	4	4	4	68	64	32	62	1051
$\mathcal{G}_2$	849	609	6	6	6	68	65	32	62	1703
PoseAngle	0	0	0	0	0	-22.5	+22.5	10	-10	-

Fig. 3. Left: Original samples in the FERET database. Middle: The standard mask. Right: The samples after preprocessing.

Fig. 4. Examples for six subjects drawn from the two normalized FERET subsets, $\mathcal{G}_1$ and $\mathcal{G}_2$.

preprocessing sequence is applied. For the computational purpose, each image is finally represented as a column vector of length $J = 17\,154$.

Following standard FR practices [56], the database $\mathcal{G}(=\mathcal{G}_1$ or $\mathcal{G}_2)$ is randomly partitioned into two subsets: The training set $\mathcal{Z}$ and test set $\mathcal{Q}$. The training set is composed of $|\mathcal{Z}| = L \cdot C$ images: L images per subject are randomly chosen, where $|\mathcal{Z}|$ denotes the size of $\mathcal{Z}$. The remaining images are used to form the test set $\mathcal{Q} = \mathcal{G} - \mathcal{Z}$. Any FR method evaluated here is first trained with $\mathcal{Z}$, and the resulting face recognizer is then applied to $\mathcal{Q}$ to produce a *classification error rate* (CER), which is defined as the fraction of the test examples wrongly classified. To enhance the accuracy of the assessment, all the CERs reported below are averaged over five runs. Each run is executed on such a random partition of the database $\mathcal{G}$ into $\mathcal{Z}$ and $\mathcal{Q}$.

B. The Comparison of FR Performance in Terms of CER

Besides the two proposed boosting methods, B-JD-LDA and B-EFM, their corresponding stand-alone JD-LDA and EFM methods (without boosting) were performed to measure the improvement brought by boosting. Meanwhile, three FR algorithms, the Eigenfaces method [7], the Fisherfaces method [8] and the Bayes matching method [9], were also implemented to provide performance baselines. Both Eigenfaces and Fisherfaces are considered to be among the most cited and influential FR algorithms [5], while the Bayes method is the top performer in the 1996/1997 FERET competitions [53].

TABLE II
COMPARISON OF THE CERs (%) AS A FUNCTION OF $(\rho_t(L), \rho_l(r))$ OBTAINED ON THE DATABASE $\mathcal{G}_1$

$\rho_t(L) = L/C$	$\rho_l(r) = r/C$	B-JD-LDA $\bar{e}_{20}(\bar{T}^*)$	B-EFM $\bar{e}_{15}(\bar{T}^*)$	JD-LDA $\bar{e}^*(M^*)$	EFM $\bar{e}^*(M^*)$
$3/104$	2/104	15.53(43)	15.89(57)	24.01(91)	22.81(57)
$4/104$	2/104	12.79(45)	11.97(65)	16.25(96)	16.57(40)
	3/104	8.69(53)	10.08(52)		
$5/104$	2/104	9.42(50)	8.06(58)	10.66(76)	11.68(46)
	3/104	5.91(40)	6.44(65)		
	4/104	6.55(44)	7.04(68)		
$6/104$	2/104	8.01(45)	5.90(64)	8.01(73)	8.29(33)
	3/104	3.93(40)	3.98(56)		
	4/104	3.70(46)	3.70(65)		
	5/104	4.07(52)	4.22(66)		
$7/104$	2/104	7.37(53)	5.33(51)	6.81(69)	7.18(38)
	3/104	4.58(40)	3.78(38)		
	4/104	3.84(37)	3.22(66)		
	5/104	4.02(50)	3.34(57)		
	6/104	4.46(36)	3.59(66)		

The first experiment conducted on $\mathcal{G}_1$ is designed to test the sensitivity of the CER measure to $\rho_t(L)$ (i.e., various SSS learning tasks arising from different database partitions) and $\rho_l(r)$ (i.e., various weakness extents of gClassifiers in each task). For all the seven methods compared here, the CER is a function of the number of extracted feature vectors, M, and the number of available training examples per subject, L. In addition, the performance of B-JD-LDA and B-EFM is affected by r, the number of examples per subject that is used to control the weakness of the produced gClassifiers during the boosting process. Considering the huge computational cost, we simply fixed the feature number $M = 20$ for B-JD-LDA and $M = 15$ for B-EFM rather than seek their respective optimal M^*s. The maximal iteration number used in boosting was set as $T = 80$, beyond which it was empirically observed that boosting was very likely to overfit. The lowest CERs finally obtained by the seven methods under various settings of $\rho_t(L)$ and $\rho_l(r)$ are depicted in Tables II and III, where $\bar{e}_x(\bar{T}^*)$ denotes the CER of B-JD-LDA or B-EFM with the best found iteration number $\bar{T}^*$ and $M = x$, while $\bar{e}^*(\bar{M}^*)$ denotes the CER of the five nonboosting methods with the best found feature number $\bar{M}^*$. All these variables have been averaged over five runs as we mentioned earlier. To further facilitate the comparison of boosting performance, we define a quantitative statistic regarding the biggest CER improvement achieved by boosting, denoted as $\bar{\xi}^*(L) = \bar{e}_x^{\{b\}}(r^*, \bar{T}^*, L) - \bar{e}^{*\{s\}}(\bar{M}^*, L)$, where $(\cdot)^{\{b\}}$ and $(\cdot)^{\{s\}}$ denote the CERs of a boosting-based method (B-JD-LDA or B-EFM) and its corresponding nonboosting version (JD-LDA or EFM), respectively, and $r^* = \arg\min_r\{\bar{e}_x^{\{b\}}(r, L)\}$. The results are summarized in Table IV, from which it can be clearly seen that B-JD-LDA and B-EFM with appropriate r values have boosted the performance of JD-LDA and EFM, respectively, across various SSS learning

TABLE III
THE CERS, $\bar{e}^*(M^*)$, OBTAINED BY THE THREE BENCHMARK METHODS ON $\mathcal{G}_1$

$\rho_t(L) =$	$3/104$	$4/104$	$5/104$	$6/104$	$7/104$
Eigenfaces	34.72(244)	32.35(258)	30.81(327)	28.48(376)	28.11(388)
Fisherfaces	24.98(103)	18.96(103)	14.35(103)	12.65(103)	10.65(103)
Bayes	19.92(192)	13.61(282)	9.11(361)	5.90(299)	5.63(376)

TABLE IV
THE BIGGEST CER IMPROVEMENT ACHIEVED BY B-JD-LDA AND B-EFM IN VARIOUS TASKS

Method	$\rho_t(L)$	$3/104$	$4/104$	$5/104$	$6/104$	$7/104$
B-JD-LDA	$\bar{\xi}^*/r^*$	-8.47/2	-7.56/3	-4.75/3	-4.31/4	-2.97/4
B-EFM	$\bar{\xi}^*/r^*$	-6.93/2	-6.49/3	-5.24/3	-4.59/4	-3.96/4

scenarios ranging from $\rho_t = 3/104$ to $\rho_t = 7/104$. Particularly, the performance enhancement is more aggressive in the worse SSS scenarios. This demonstrates the effectiveness of the two boosting approaches against the SSS problem. The biggest improvement, $\bar{\xi}^* = -8.47\%$, is achieved by B-JD-LDA when $r^* = 2$ in the most difficult task $\rho_t = 3/104$.

The second experiment conducted on $\mathcal{G}_2$ is designed to test the CER performance changes as the size of the evaluation dataset increases. Since the two boosting methods require at least three training samples per subject, we are allowed to create only one partition case from $\mathcal{G}_2$, i.e., $L = 3$, which leads to an SSS learning task with $\rho_t = 3/256$. Correspondingly, the lowest CERs obtained by the seven methods are shown in Table V. It can be seen from these results that a stable boosting performance is achieved by both boosting approaches. The quantitative statistic $\bar{\xi}^*$ goes up to -8.64% and -7.49% for B-JD-LDA and B-EFM, respectively. The results indicate only a slightly better boosting performance compared to that achieved under the assumption $(\rho_t = 3/104)$ in the first experiment, although in theory a higher performance improvement is expected when more pattern variations are introduced. The reason may be explained by the fact that most new samples added to the database $\mathcal{G}_2$ come from the fa and fb sets, the simplest categories in the FERET database, as shown in Table I. As a result, the performance margins between different methods is reduced to some extent.

In both of the two experiments, Eigenfaces is the worst performer among the seven methods. From the results delivered by the most popular benchmark method, we can roughly learn how difficult it is to conduct face recognition on the two evaluation datasets, $\mathcal{G}_1$ and $\mathcal{G}_2$. Also, it is of interest to compare the performance of B-JD-LDA and B-EFM with that of the Bayes method. Published results indicate that the latter generally outperforms, in terms of CER, most subspace-based FR approaches including those using traditional LDA, Kernel Principal Component Analysis (K-PCA) or Independent Component Analysis techniques (ICA) by a margin of at least ten percent [57]. However, as it can be seen from Tables II, III, and V, both boosting methods are superior to the Bayes method. Especially, B-JD-LDA leads the state-of-the-art method up to 4.38% and 4.91% in the two most difficult learning tasks, $\rho_t(L) = 3/104$ and $\rho_t(L) = 4/104$ in the first experiment. Although we admit that our implementation of the Bayes method[1] may not be as good as the original implementation of Moghaddam *et al.* [9], this comparison still provides a promising perspective: It is possible to boost a traditional FR algorithm to the state-of-the-art level under the proposed framework. Moreover, it should be mentioned again at this point that unlike the five nonboosting methods, we did not seek the CERs with the optimal M^* values for the two boosting approaches. Obviously $\bar{e}_{15}$ or $\bar{e}_{20}$, as a substitute for $\bar{e}_{M^*}$, is only suboptimal. We expect that a higher boosting performance gain can be obtained when a better M value is used.

C. Weakness Analysis of the gClassifiers

As it was mentioned earlier, the proposed boosting approaches would fail, in theory, to perform well when too weak or too strong gClassifiers are utilized. Clearly, it can be experimentally observed at this point from the example shown in Fig. 5, where the results are obtained by B-JD-LDA in the task $\rho_t(L) = 5/104$. In this example, the weakest and strongest gClassifiers are produced by B-JD-LDA when $r = 2$ and $r = 5$, respectively. However, the generalization performance of the former is only slightly better than that of the single JD-LDA, while the latter tends to overfit quickly, although it yields the lowest training error. In contrast with this, appropriately weak gClassifiers are produced when $r = 3, 4$ are used. In these cases, it can be seen from Fig. 5 that the B-JD-LDA exhibits the beautiful property of boosting: The test CER is continuously improved, even long after the training error has dropped down to zero. Similar phenomena have been also observed with the B-EFM method.

Based on the theory developed in Section III-C, the gClassifiers with the best weakness or the optimal $\rho_l(r^*)$ can be found by minimizing a generalization loss function $\mathbf{R}(r)$ (6) with respect to r, i.e., $r^* = \arg\min_r \mathbf{R}(r)$. To test the estimation accuracy of the method, we applied the loss function to the various learning tasks designed in the first experiment. The obtained results including $\mathbf{R}(r), r^*$, and the worst r value (r^-) are depicted in Tables VI and VII for B-JD-LDA and B-EFM, respectively, where the values of λ were found empirically. It should be mentioned here that it is not a difficult task to find an appropriate λ value within [0, 1]. In fact, our experiments reveal that there exist a range of λ values which produce the same estimation for the preference rankings of the r values, for example, $\lambda \in [0.57, 0.65]$ for B-JD-LDA found in the experiment. Comparing the r rankings to the CER results shown in Table II, it can

[1]To reduce the effect of reimplementation related issue, we use the maximum likelihood (ML) version of the Bayes method instead of the maximum a posteriori (MAP) version. The former is much easier to be implemented than the latter. However, there is only a very slight performance difference between the ML and MAP versions as shown in the works of [58].

TABLE V
COMPARISON OF THE CERS (%) OBTAINED ON THE DATABASE $\mathcal{G}_2$

B-JD-LDA	B-EFM	JD-LDA	EFM	Eigenfaces	Fisherfaces	Bayes
$\bar{e}_{15}(\bar{T}^*)$	$\bar{e}_{15}(\bar{T}^*)$	$\bar{e}^*(M^*)$	$\bar{e}^*(M^*)$	$\bar{e}^*(M^*)$	$\bar{e}^*(M^*)$	$\bar{e}^*(M^*)$
13.45(51)	13.58(63)	22.10(110)	21.07(88)	34.87(565)	33.43(254)	16.11(327)

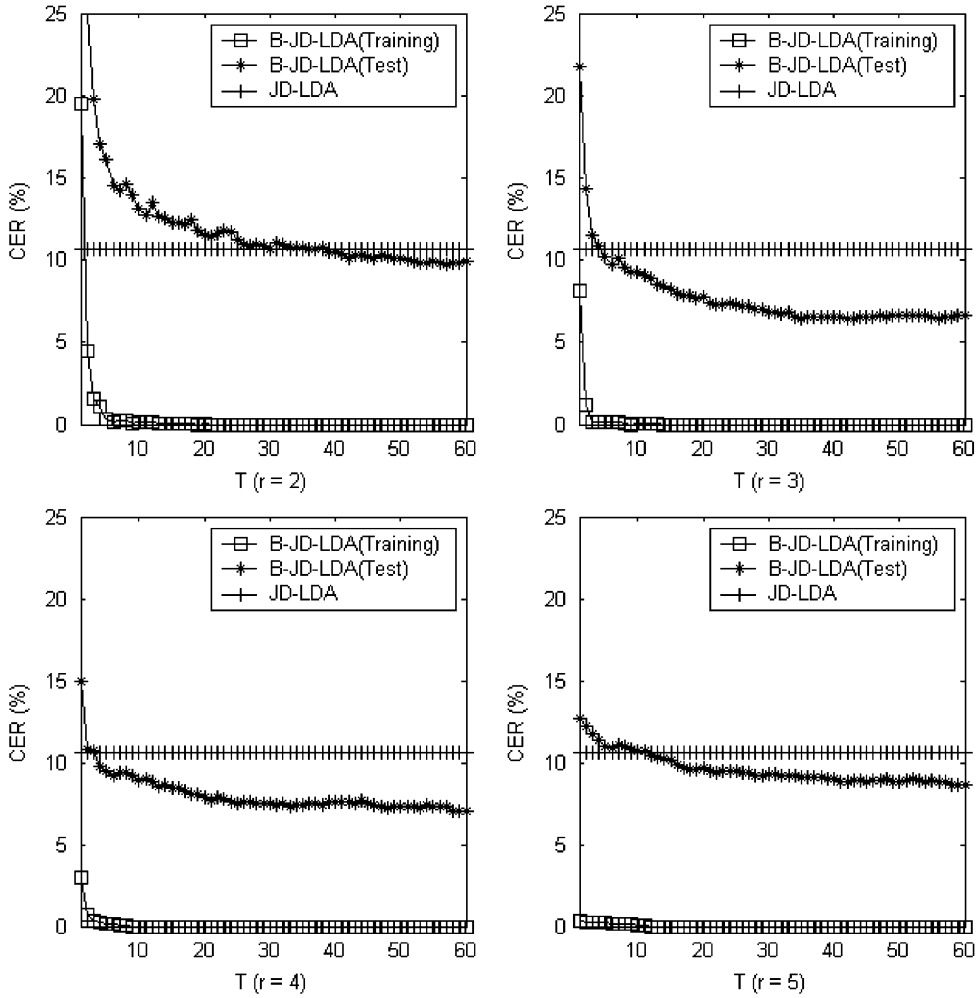

Fig. 5. Training and test CERs of B-JD-LDA with varying weakness extents of gClassifiers as a function of T in the task $\rho_t(L) = 5/104$. The CER of JD-LDA is the one obtained with $M = M^*$.

TABLE VI
THE GENERALIZATION LOSS $\mathbf{R}(r, L)$ WITH $\lambda = 0.62$, THE BEST r ESTIMATE (r^*) AND THE WORST r ESTIMATE (r^-) OBTAINED BY B-JD-LDA ON THE DATABASE $\mathcal{G}_1$

$\rho_t(L)$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	Estimation	
4/104	0.6044	0.5709	–	–	–	$r^*=3$	$r^-=2$
5/104	0.6077	0.5599	0.5669	–	–	$r^*=3$	$r^-=2$
6/104	0.6091	0.5523	0.5402	0.5703	–	$r^*=4$	$r^-=2$
7/104	0.6049	0.5446	0.5221	0.5377	0.5759	$r^*=4$	$r^-=2$

TABLE VII
THE GENERALIZATION LOSS $\mathbf{R}(r, L)$ WITH $\lambda = 0.62$, THE BEST r ESTIMATE (r^*) AND THE WORST r ESTIMATE (r^-) OBTAINED BY B-EFM ON THE DATABASE $\mathcal{G}_1$

$\rho_t(L)$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	Estimation	
4/104	0.6222	0.5889	–	–	–	$r^*=3$	$r^-=2$
5/104	0.6275	0.5808	0.5828	–	–	$r^*=3$	$r^-=2$
6/104	0.6214	0.5730	0.5620	0.5797	–	$r^*=4$	$r^-=2$
7/104	0.6158	0.5626	0.5462	0.5549	0.5819	$r^*=4$	$r^-=2$

be seen that the values of the loss correctly indicate the optimal r^*, the worst r^-, and even the goodness of the r values between them in most cases, for example, the second, third, and fourth best r values.

D. Some Discussions on the Convergence of Boosting

From Fig. 5, it can be seen that given an appropriately weak learner, the generalization error of h_f often continues to drop as T becomes large even long after the training error reaches

zero. However, the phenomenon also leads to the difficulty in determining when the boosting procedure should be stopped in order to avoid possible overfitting.

Considering the relationship between boosting and the margin theory, intuitively, it is reasonable to use the cumulative margin distribution of the training examples as an indicator to roughly estimate an appropriate value of T. In other words, we can observe the changes of the margins of the training examples at every boosting iteration, and consider it convergent when the margins of most training examples stop increasing or are increasing slowly. Our experiments indicate that this approach works well in many cases (see [51] for details). However, as mentioned earlier, the margin theory alone is insufficient to explain the behaviors of boosting [40], [45], [46]. It is therefore unrealistic to expect that the heuristic approach can accurately estimate the optimal value of T. For example, it is found in our experiments that JD-LDA with the best found M^* often yielded much better cumulative margin distributions than its boosting version [51].

Also, compared to many other boosting methods that usually need hundreds of iterations, it should be noted that only around $T^* < 70$ iterations are required to find an excellent result using the B-JD-LDA and B-EFM algorithms in the experiments reported in Tables II and V. Considering that each gClassifier works in a considerably lower-dimensional subspace $(M = 15, 20)$ compared to those nonboosting methods, such a computational cost is affordable for most existing personal computers.

V. Conclusion

In this paper, a novel weakness analysis theory has been developed to overcome the limitation of the weak learners in traditional boosting techniques. The theory proposed here is composed of a cross-validation mechanism of weakening a strong learner and a subsequent estimation method of appropriate weakness for the classifiers created by the learner. With the introduction of the weakness analysis theory, a traditional boosting algorithm can be used to work effectively with a general (strong or weak) learner. To demonstrate the effectiveness, the new boosting framework is applied to two strong LDA-style learners, which are generally believed to be rather difficult to be boosted. To this end, a novel variable, the pairwise class discriminant distribution, is introduced to build an effective connection between the booster and the learners. As a result, two novel ensemble-based discriminant learning methods, B-JD-LDA and B-EFM, are introduced. By manipulating the boosting process, a set of specific LDA feature spaces can be constructed effectively in a manner of similar to "automatic gain control." Unlike most traditional mixture models of linear subspaces that are based on cluster analysis [59], these LDA subspaces are generalized in the context of classification error minimization.

The effectiveness of the proposed B-JD-LDA and B-EFM approaches including boosting power, estimation accuracy of the loss function, and robustness against the overfitting and SSS problems has been demonstrated through the FR experimentation performed on the FERET database. It is further anticipated that in addition to JD-LDA and EFM, other existing traditional face recognizers such as those based on PCA or ICA techniques may be boosted to higher levels through integration into the proposed boosting framework.

Appendix I
Two LDA-Style Learners: EFM and JD-LDA

In face recognition applications, each sample $\mathbf{z}_{ij}$ is a face image, represented as a column vector of length $J(=I_w \times I_h)$, i.e., $\mathbf{z}_{ij} \in \mathbb{Z} = \mathbb{R}^J$, where $(I_w \times I_h)$ is the image size, and $\mathbb{R}^J$ denotes the J-dimensional real space. A LDA-style learner mainly functions as a feature extractor, which determines a set of optimal discriminant basis vectors, denoted as $\{\psi_m\}_{m=1}^{M}$ where $\psi_m \in \mathbb{R}^J$ and $M \ll J$, by optimizing the Fisher's discriminant criterion

$$\Psi = \arg\max_{\Psi} \frac{|\Psi^T \mathbf{S}_b \Psi|}{|\Psi^T \mathbf{S}_w \Psi|}, \quad \text{with } \Psi = [\psi_1, \ldots, \psi_M] \tag{7}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between- and within-class scatter matrices of the training set, respectively. However, the estimation for either $\mathbf{S}_b$ or $\mathbf{S}_w$ is extremely ill-posed due to the SSS problem in most FR tasks. Generally, two kinds of discriminant feature bases are considered for the solution: 1) $\psi_i^T \mathbf{S}_b \psi_i > 0$ and $\psi_i^T \mathbf{S}_w \psi_i > 0$; and 2) $\psi_i^T \mathbf{S}_b \psi_i > 0$ and $\psi_i^T \mathbf{S}_w \psi_i = 0$. Some researchers such as [8], [10], [13], prefer the feature (1) based on the consideration that the small/zero eigenvalues of $\mathbf{S}_w$ tend to capture noise. Other researchers such as [11], [12], [47], consider (2) the optimal discriminant feature bases, since they maximize the ratio of (7). Particularly, EFM [13] is an extension to Fisherfaces [8], while JD-LDA [47] is an improvement of [12]. Based on the experience of the authors, it is hard to say which kind of features are better. Different experimental settings often lead to different conclusions as shown in Table II. Both EFM and JD-LDA return $(\Psi, \{\bar{\mathbf{z}}_i\}_{i=1}^{C}) = \mathcal{L}(\mathcal{Z})$, where $\bar{\mathbf{z}}_i = (1/L_i)\sum_{j=1}^{L_i} \mathbf{z}_{ij}$ is the center of the class i. For simplicity, we denote EFM or JD-LDA as a function $\mathcal{L}(\cdot)$, which has $(\Psi, \{\bar{\mathbf{z}}_i\}_{i=1}^{C}) = \mathcal{L}(\mathcal{Z})$. For the limitation of space, only the pseudocode implementation of the less known JD-LDA is depicted as a learner in Fig. 6. With the example, the EFM learner can be implemented easily in a similar way.

For an input face image $\mathbf{z}$, its LDA-based representation $\mathbf{y}$ can be obtained by a linear mapping: $\mathbf{y} = \Psi^T \mathbf{z}, \mathbf{y} \in \mathbb{R}^M$. The subsequent classification in the feature space can be performed using any classifier. However, from the viewpoint of reducing the overfitting chances in the context of boosting, a *simple* discriminant function that explains most of the data is preferable to a complex one. Consequently, a classic *nearest center classifier* (NCC) is adopted here for the classification task. The NCC is based on a normalized Euclidean distance, given by

$$d(\mathbf{z}, i, \mathcal{L}) = \frac{(d_{\max} - d_{\mathbf{z},i})}{(d_{\max} - d_{\min})} \tag{8}$$

where $d_{\mathbf{z},i} = ||\Psi^T(\mathbf{z} - \bar{\mathbf{z}}_i)||, d_{\max} = \max(\{d_{\mathbf{z},i}\}_{i=1}^{C})$, and $d_{\min} = \min(\{d_{\mathbf{z},i}\}_{i=1}^{C})$. Based on the nearest center rule, the class label $y(\mathbf{z})$ of the input $\mathbf{z}$ can be inferred through $y(\mathbf{z}) = \arg\max_i d(\mathbf{z}, i, \mathcal{L})$. The classification score $d(\mathbf{z}, i, \mathcal{L})$ has values in [0, 1], and thus it can fulfill the functional requirement of the boosting algorithm (AdaBoost.M2 [31]), indicating

Input: A training set $\mathcal{Z}_t$ with C classes: $\mathcal{Z}_t = \{\mathcal{Z}_{i,t}\}_{i=1}^{C}$, each class contains $\mathcal{Z}_{i,t} = \{\mathbf{z}_{ij}\}_{j=1}^{L_i}$ face images, where $\mathbf{z}_{ij} \in \mathbb{R}^J$.

Output: A M-dimensional LDA subspace spanned by Ψ_t, a $M \times J$ matrix with $M \ll J$, and the class centers $\{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C}$.

Algorithm:

Step 1. Re-write $\hat{\mathbf{S}}_{b,t}$ of Eq.3: $\hat{\mathbf{S}}_{b,t} = \mathbf{W}_b\mathbf{W}_b^T$, where $\mathbf{W}_b = [\phi_1, \cdots, \phi_c]$.

Step 2. Find the eigenvectors of $\mathbf{W}_b^T\mathbf{W}_b$ with non-zero eigenvalues, and denote them as $\mathbf{E}_m = [\mathbf{e}_1, \ldots, \mathbf{e}_m]$, $m \leq C-1$.

Step 3. Calculate the first m most significant eigenvectors ($\mathbf{V}$) of $\hat{\mathbf{S}}_{b,t}$ and their corresponding eigenvalues (Λ_b) by $\mathbf{V} = \mathbf{W}_b\mathbf{E}_m$ and $\Lambda_b = \mathbf{V}^T\hat{\mathbf{S}}_{b,t}\mathbf{V}$.

Step 4. Let $\mathbf{U} = \mathbf{V}\Lambda_b^{-1/2}$. Find eigenvectors of $\mathbf{U}^T(\hat{\mathbf{S}}_{b,t} + \hat{\mathbf{S}}_{w,t})\mathbf{U}$, $\mathbf{P}$, where $\hat{\mathbf{S}}_{w,t}$ is defined in Eq.4.

Step 5. Choose the $M(\leq m)$ eigenvectors in $\mathbf{P}$ with the smallest eigenvalues. Let $\mathbf{P}_M$ and Λ_w be the chosen eigenvectors and their corresponding eigenvalues respectively.

Step 6. Return $\Psi_t = \mathbf{U}\mathbf{P}_M\Lambda_w^{-1/2}$ and $\{\bar{\mathbf{z}}_{i,t}\}_{i=1}^{C}$.

Fig. 6. Pseudocode implementation of the JD-LDA feature extractor: $\mathcal{L}(\mathcal{Z}_t)$ in the tth boosting iteration, where the input $\mathcal{Z}_t = \mathcal{R}_t$, and $\mathcal{R}_t \subset \mathcal{Z}$ is an adaptively updated subset defined in Section III-B.

a "degree of plausibility" for labeling $\mathbf{z}$ as the class i. Since a classifier h such as the NCC discussed here usually yields two outputs, the classification score $d(\mathbf{z}, i, \mathcal{L})$ and the class label $y(\mathbf{z})$, we denote $h(\mathbf{z}) = y(\mathbf{z})$, and $h(\mathbf{z}, i) = d(\mathbf{z}, i, \mathcal{L})$ for the distinguishing purpose.

ACKNOWLEDGMENT

Portions of the research in this paper use the FERET database of facial images collected under the FERET program [54]. The authors would like to thank the FERET Technical Agent, the U.S. National Institute of Standards and Technology (NIST) for providing the FERET database.

REFERENCES

[1] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, pp. 65–77, 1992.

[2] D. Valentin, H. A. Alice, J. O. Toole, and G. W. Cottrell, "Connectionist models of face processing: A survey," *Pattern Recognit.*, vol. 27, no. 9, pp. 1209–1230, 1994.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[4] S. Gong, S. J. McKenna, and A. Psarrou, *Dynamic Vision From Images to Face Recognition*, Singapore: World Scientific , May 2000.

[5] M. Turk, "A random walk through eigenspace," *IEICE Trans. Inf. Syst.*, vol. E84-D, no. 12, pp. 1586–1695, Dec. 2001.

[6] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[7] M. A. Turk and A. P. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. 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.

[9] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, pp. 1771–1782, 2000.

[10] W. Zhao, R. Chellappa, and J. Phillips, "Subspace linear discriminant analysis for face recognition," Univ. Maryland, College Park, MD, Tech. Rep., CS-TR4009, 1999.

[11] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, pp. 1713–1726, 2000.

[12] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, pp. 2067–2070, Oct. 2001.

[13] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[14] J. Ye and Q. Li, "LDA/QR: An efficient and effective dimension reduction algorithm and its theoretical foundation," *Pattern Recognit.*, vol. 37, no. 4, pp. 851–854, Apr. 2004.

[15] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 679–691, May 2005.

[16] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.

[17] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology," *CVGIP: Image Understanding*, vol. 59, pp. 254–261, 1994.

[18] S. Kutin, "Algorithmic Stability and Ensemble-Based Learning," Ph.D. Thesis, Faculty Div. Phys. Sci., Univ. Chicago, June 2002.

[19] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1999.

[20] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.

[21] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[22] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using feature optimization and ν-support vector learning," in *Proc. IEEE Int. Workshop Neural Networks for Signal Processing*, Falmouth, MA, Sep. 2001, pp. 373–382.

[23] ——, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

[24] M. Wang and S. Chen, "Enhanced fmam based on empirical kernel map," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 557–564, May 2005.

[25] S. Pang, D. Kim, and S. Y. Bang, "Face membership authentication using svm classification tree generated by membership-based lle data partition," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 436–446, Mar. 2005.
[26] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–474, Mar. 2005.
[27] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. Computer Vision and Pattern Recognition Conf.* , June 1994, pp. 1–7.
[28] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, Jan. 1998.
[29] B. J. Frey, A. Colmenarez, and T. S. Huang, "Mixtures of local linear subspaces for face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, Jun. 1998, pp. 32–37.
[30] J. Lu and K. N. Plataniotis, "Boosting face recognition on a large-scale database," in *Proc. IEEE Int. Conf. Image Processing*, Rochester, NY, Sep. 2002, pp. II.109–II.112.
[31] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
[32] R. E. Schapire, "The boosting approach to machine learning: An overview," *MSRI Workshop Nonlinear Estimation and Classification*, pp. 149–172, 2002.
[33] H. Drucker and C. Cortes, "Boosting decision trees," *Adv. Neural Inform. Process. Syst. 8*, pp. 479–485, 1996.
[34] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," in *Proc. 14th Int. Conf. Machine Learning* , 1997, pp. 322–330.
[35] L. Breiman, "Arcing classifiers," *Ann. Statistics*, vol. 26, no. 3, pp. 801–849, 1998.
[36] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[37] L. Breiman, "Population theory for boosting ensembles," *Ann. Statistics*, vol. 32, no. 1, pp. 1–11, 2004.
[38] ——, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[39] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, Jun. 2002.
[40] A. Murua, "Upper bounds for error rates of linear combinations of classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 591–602, May 2002.
[41] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. 9th IEEE Int. Conf. Computer Vision* , vol. 2, Oct. 2003, pp. 734–741.
[42] K. Tieu and P. Viola, "Boosting image retrieval," *Int. J. Comput. Vis.*, vol. 56, no. 1, pp. 17–36, 2004.
[43] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, May 2004.
[44] Y. Freund and R. E. Schapire, "A discussion of "process consistency for adaboost" by wenxin jiang, "on the bayes-risk consistency of regularized boosting methods" by gbor lugosi and nicolas vayatis, "statistical behavior and consistency of classification methods based on convex risk minimization" by tong zhang," *Ann. Statist.*, vol. 32, no. 1, 2004.
[45] A. J. Grove and D. Schuurmans, "Boosting in the limit: Maximizing the margin of learned ensembles," in *Proc. 15th Nat. Conf. Artifical Intelligence* , July 1998, pp. 692–699.
[46] C. Rudin, R. E. Schapire, and I. Daubechies, "Boosting based on a smooth margin," in *COLT (Computational Learning Theory)*, J. Shawe-Taylor and Y. Singer, Eds. New York: Springer-Verlag, 2004.
[47] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.
[48] ——, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognit. Lett.*, vol. 24, no. 16, pp. 3079–3087, Dec. 2003.
[49] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 623–627, Jun. 2000.
[50] *Out-of-Bag Estimation*, L. Breiman. (1996). [Online]. Available: ftp://ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z
[51] J. Lu, "Discriminant learning for face recognition," Ph.D. Dissertation, The Edward S. Rogers Sr. Dept. Elect. Comp. Eng., Univ. Toronto, Toronto, Canada, Jun. 2004.
[52] Image Group, Information Access Division, ITL, NIST (2004, Jan.). [Online]. Available: http://www.itl.nist.gov/iad/humanid/feret/
[53] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[54] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image Vis. Comput. J.*, vol. 16, no. 5, pp. 295–306, 1998.
[55] L.-F. Chen, H.-Y. M. Liao, J.-C. Lin, and C.-C. Han, "Why recognition in a statistics-based face recognition system should be based on the pure face portion: A probabilistic decision-based proof," *Pattern Recognit.*, vol. 34, no. 7, pp. 1393–1403, 2001.
[56] P. J. Phillips and E. M. Newton, "Meta-analysis of face recognition algorithms," in *Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Washinton, DC, May 20–21, 2002, pp. 235–241.
[57] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 780–788, Jun. 2002.
[58] *Evaluation of Face Recognition Algorithms Website*, R. Beveridge and B. Draper. (2004, Dec.). [Online]. Available: http://www.cs.colostate.edu/evalfacerec
[59] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

Juwei Lu (M'00) received the B.Eng. degree in electrical engineering from Nanjing University of Aeronautics and Astronautics, China, in 1994, the M.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 1999, and the Ph.D. degree in electrical and computer engineering from University of Toronto, Canada, in 2004.

From July 1999 to January 2001, he was a Research Engineer with the Center for Signal Processing, Singapore. From April 2004 to December 2004, he was a Postdoctoral Researcher at the Bell Canada Multimedia Laboratory, Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto. Currently, he is a Senior Software Developer at the Epson Canada Limited, Toronto, ON, Canada. His research interests include multimedia signal processing, visual object detection and recognition, kernel methods, support vector machines, neural networks, and boosting technologies. He has published 28 refereed papers and book chapters in these areas.

Dr. Lu is a member IEEE Computational Intelligence Society. He is a reviewer of many journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS - PART B and *Pattern Recognition Letters*.

Konstantinos N. (Kostas) Plataniotis (S'90–M'92–SM'03) received the B. Eng. degree in computer engineering informatics from University of Patras, Greece, in 1988, and the M.S and the Ph.D. degrees in electrical engineering from Florida Institute of Technology (Florida Tech) in Melbourne, Florida, in 1992 and 1994, respectively.

He is an Associate Professor with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto, Toronto, ON, Canada, a Nortel Institute for Telecommunications Associate, a member of the Knowledge Media Design Institute at the University of Toronto and an Adjunct Professor with the School of Computer Science at Ryerson University. His research interests include adaptive systems, biometrics, image and signal processing, stochastic estimation, and pattern recognition.

Dr. Plataniotis is the Vice Chair of the *9th International IEEE Conference on Intelligent Transportation Systems (ISTC 06)*, September 18-20 2006, Toronto, Canada, and the Technical Program Co-Chair for the *IEEE International Conference on Multimedia Expo (ICME)* 2006, July 9–12, Toronto, Canada. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS.

Anastasios N. Venetsanopoulos (S'66–M'69–SM'79–F'88) received the Diploma in engineering degree from the National Technical University of Athens (NTU), Athens, Greece, in 1965, and the M.S., M.Phil., and Ph.D. degrees in electrical engineering from Yale University, New Heaven, CT, in 1966, 1968, and 1969, respectively.

He joined the Department of Electrical and Computer Engineering of the University of Toronto, ON, Canada, in September 1968, as a Lecturer and he was promoted to Assistant Professor in 1970, Associate Professor in 1973, and Professor in 1981. Since July 1997, he has been Associate Chair: Graduate Studies of the Edward S. Rogers Sr. Department of Electrical and Computer Engineering and was Acting Chair during the spring term of 1998–1999. Since July 2001, he has served as the 12th Dean of the Faculty of Applied Science and Engineering of the University of Toronto. He has served as Chair of the Communications Group and Associate Chair of the Department of Electrical Engineering and Associate Chair: Graduate Studies for the Department of Electrical and Computer Engineering. He was on research leave at Imperial College of Science and Technology, the National Technical University of Athens, the Swiss Federal Institute of Technology, the University of Florence and the Federal University of Rio de Janeiro, and has also served as Adjunct Professor at Concordia University. He has served as lecturer in 138 short courses to industry and continuing education programs and as Consultant to numerous organizations; he is a contributor to twenty eight (28) books, a coauthor of *Color Image Processing and Applications* (New York: Springer-Verlag, 2000), *Nonlinear Filters in Image Processing: Principles Applications* (Norwell, MA: Kluwer, 1990), *Artificial Neural Networks: Learning Algorithms, Performance Evaluation and Applications* (Norwell, MA: Kluwer, 1993), and *Fuzzy Reasoning in Information Decision and Control systems* (Norwell, MA: Kluwer 1994). He has served as Chair on numerous boards, councils and technical conference committees of the Institute of Electrical and Electronic Engineers (IEEE), such as the Toronto Section (1977–1979) and the IEEE Central Canada Council (1980–1982); he was President of the Canadian Society for Electrical Engineering and Vice President of the Engineering Institute of Canada (EIC) (1983–1986). He was a Guest Editor or Associate Editor for several IEEE Journals and the Editor of the *Canadian Electrical Engineering Journal* (1981–1983). He was the Technical Program Co-Chair of the *IEEE International Conference on Image Processing (ICIP'01)*. He has published 750 papers in refereed journals and conference PROCEEDINGS on digital signal and image processing, and digital communications.

Prof. A.N. Venetsanopoulos is a member of the IEEE Communications, Circuits and Systems, Computer, and Signal Processing Societies of IEEE, as well as a member of Sigma Xi, the Technical Chamber of Greece, the European Association of Signal Processing, the Association of Professional Engineers of Ontario (APEO) and Greece. He was elected as a Fellow of the IEEE "for contributions to digital signal and image processing." He is also a Fellow of the EIC, "for contributions to electrical engineering," and was awarded an Honorary Doctorate from the National Technical University of Athens, in October 1994. In October 1996, he was awarded the "Excellence in Innovation Award" of the Information Technology Research Center of Ontario and Royal Bank of Canada, "for innovative work in color image processing."

Stan Z. Li received the B.Eng. degree from Hunan University, P. R. China, the M.Eng. degree from National University of Defense Technology, P. R. China, and the Ph.D. degree from Surrey University, U.K.

He is a Researcher at National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China, Beijing, P. R. China, and the Director of the Center for Biometrics and Security Research (CBSR), Beijing, P. R. China. He worked at Microsoft Research Asia, Beijing, P. R. China, as a Researcher, from May 2000 to August 2004 . Prior to that, he was an Associate Professor of Nanyang Technological University, Singapore. His current research interest is in face recognition technologies, biometrics, intelligent surveillance, pattern recognition, and machine learning. He has published several books, including *Handbook of Face Recognition* (New York: Springer-Verlag, 2004) and *Markov Random Field Modeling in Image Analysis* (New York: Springer-Verlag, 2nd edition in 2001), and over 200 refereed papers and book chapters in these areas.