Face recognition using direct, weighted linear discriminant analysis and modular subspaces

Jeffery R. Price*, Timothy F. Gee

Oak Ridge National Laboratory, P.O. Box 2008, MS-6010, Oak Ridge, TN 37831-6010, USA

Abstract

We present a modular linear discriminant analysis (LDA) approach for face recognition. A set of observers is trained independently on different regions of frontal faces and each observer projects face images to a lower-dimensional subspace. These lower-dimensional subspaces are computed using LDA methods, including a new algorithm that we refer to as direct, weighted LDA or DW-LDA. DW-LDA combines the advantages of two recent LDA enhancements, namely direct LDA (D-LDA) and weighted pairwise Fisher criteria. Each observer performs recognition independently and the results are combined using a simple sum-rule. Experiments compare the proposed approach to other face recognition methods that employ linear dimensionality reduction. These experiments demonstrate that the modular LDA method performs significantly better than other linear subspace methods. The results also show that D-LDA does not necessarily perform better than the well-known principal component analysis followed by LDA approach. This is an important and significant counterpoint to previously published experiments that used smaller databases. Our experiments also indicate that the new DW-LDA algorithm is an improvement over D-LDA.
© 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Face recognition; Linear discriminant analysis; Dimensionality reduction; Pairwise Fisher criteria

1. Introduction

Despite the availability of commercial systems, face recognition continues to be an active topic in computer vision research. Current face recognition systems perform well under nearly ideal circumstances, but tend to suffer when variations in expression, illumination, decoration (i.e., glasses, facial hair), and/or pose are present. Most current face recognition research aims to improve recognition performance in the presence of such confounding factors. Face recognition methods can be classified broadly into two

categories: feature- or template-based, as described in Ref. [1]. The research presented in this paper is a template-based approach where the image pixels themselves serve as the features. Below we survey the research that has motivated our work. We note that there are other confounding factors (aging and/or drastic weight change, for example) that are beyond the scope of our interest. Furthermore, there are many other interesting and effective face recognition approaches that are not related to this work. A few examples include elastic bunch graph matching [2], support vector classification [3], morphable models [4], and light-fields [5]; certainly, the interested reader can find many more.

Illumination: Approaches for dealing with varying illumination are primarily based upon linear discriminant analysis (LDA), sometimes referred to as "Fisherfaces" [6–8]. A motivating principle behind these techniques is the

---

* Corresponding author. Tel.: +1-865-574-5743; fax: +1-865-576-8380.

E-mail address: pricejr@ornl.gov (J.R. Price).

approximation of a face as a Lambertian surface. As noted in Ref. [6], the images of a Lambertian surface under varying illumination lie in a linear subspace of the entire image space and, under ideal conditions, are linearly separable.

Expression: Varying facial expression can be modeled to some degree by the active appearance models (AAMs) presented in Ref. [9]. AAMs characterize shape and texture information using a statistical point distribution approach.

Illumination and expression: Bayesian face recognition [10–12] has been proposed to improve robustness in the presence of varying illumination and expression. These approaches employ probabilistic models to characterize *intra-personal* and *inter-personal* differences with a principal component analysis (PCA) or "eigenface" representation. In Ref. [10], it is noted that the Bayesian approach can be thought of as a general, non-linear extension of LDA. With this in mind, it seems a reasonable hypothesis that LDA should also be able to address both illumination and expression to some degree. Recent research [13] has demonstrated this hypothesis to be true.

Facial decoration: There has been very little work towards explicitly handling facial decoration. In Refs. [12] and [14], it was shown that two "eigenfeature" images—the eyes and the nose—could be used for accurate recognition after a change in facial hair. However, no method for online selection of the appropriate eigenfeatures was suggested. In the LDA approach described in Ref. [7], some promising results were obtained after artificially degrading face images, indicating that LDA might also provide a reasonable solution to handling some degree of decoration, assuming that the registration landmarks can still be located (i.e., no occlusions of landmarks such as dark glasses hiding the eyes or scarves covering the mouth).

Pose: One method to handle varying pose is the view-based eigenspace approach [14], which was recently shown to perform quite well [15]. Each pose is represented by its own subspace and the multiple subspaces act as independently trained "experts" or observers trying to explain the data. Similarly, motivated techniques include characteristic eigenspace curves [16] and view-based AAMs [17].

Pose and expression: AAM methods [17,18] have been proposed to handle both varying pose and expression.

Pose and illumination: Methods were presented in Refs. [19] and [20] to deal with varying pose and illumination. These methods rely upon generative models that can synthesize a given face under varying illumination from different viewpoints. Although the performance in Ref. [19] is quite remarkable, the proposed method employs seven training images for each subject under strictly controlled lighting and does not address expression or decoration.

The research presented in this paper is motivated by the goal of personnel monitoring in critical spaces of secure facilities. In these situations, we will need to recognize between 100 and 150 people and will have access to good training data. Variations in illumination, expression, and decoration (particularly eyeglasses) are expected. Since access to

the spaces in question is generally well-controlled and monitored by video cameras, the acquisition of frontal images is relatively easy compared to less controlled situations, hence pose variation issues are minimal. With all of these facts in mind, we now note the specific contributions of this paper.

- We propose a modular LDA face recognition algorithm, which is an improvement over the modular PCA approach. Through careful analysis of previous research, our approach explicitly aims to address three of the four confounding factors, namely illumination, expression, and decoration. None of the algorithms presented above addresses more than two confounding factors. Assuming an accurate pose estimator (a subject of ongoing research) and adequate training data, we believe the extension of the proposed system to variable pose is straightforward, as discussed briefly in Section 3.4.

- We propose a new LDA algorithm called direct, weighted LDA (DW-LDA) that simultaneously provides the advantages of both direct LDA [21] and weighted pairwise Fisher criteria [22]. A point of significant interest is that we find experimentally that the direct LDA methods do not perform as well in terms of classification accuracy as PCA plus LDA methods. This is in contrast with earlier results in the literature [21]. The direct LDA methods do, however, provide the means to perform subspace computation when there is abundant training data (i.e., no small sample size problem) and many subjects, where PCA might be computationally intractable due to the dimensionality of the full rank covariance matrix. For example, if we had 1000 subjects and 10,000 training images of 10,000 pixels each, PCA would require the eigen-decomposition of a $10,000 \times 10,000$ matrix. D-LDA and DW-LDA, however, would only require the eigen-decomposition of a 1000×1000 matrix.

- In Section 2.3, we note what seems to be contradiction between direct LDA and the weighted, pairwise Fisher criteria regarding the importance of the nullspace of the within-class scatter matrix. (Understanding this contradiction is the subject of ongoing research.)

- We provide experimental results comparing several subspace approaches, both with and without classifier combination. These experiments are the first for modular LDA and, for modular PCA, provide results on a larger database than has been previously published. These results demonstrate the significant performance improvements achievable using simple classifier combination with modular LDA (or modular PCA) subspaces. Perhaps most importantly, these results are the first to indicate that, although computational benefits are indeed provided, direct LDA methods do not necessarily perform better than PCA-first methods in terms of classification accuracy.

- We describe the computation and use of a simple confidence metric. We show experimentally how this confidence metric can be employed to significantly improve

accuracy in situations where multiple observations of a given subject are expected.

The remainder of this paper is organized as follows. In Section 2, we first review traditional LDA (which we will refer to as T-LDA), direct LDA (D-LDA), and weighted LDA (W-LDA). We then present an algorithm that combines both direct and weighted LDA in a unified algorithm we refer to as DW-LDA. In Section 3, we present the multiple-observer, modular LDA subspace system. We then provide some experimental results in Section 4 and conclude in Section 5 with some closing remarks.

2. DW-LDA

The aim of traditional LDA (T-LDA) is to project high-dimensional feature vectors in $\mathbb{R}^n$ onto a lower-dimensional subspace $\mathbb{R}^m$, where $m < n$, while preserving as much discriminative information as possible. One formal expression for the corresponding optimization criterion (see Ref. [23] for equivalents) can be written

$$\arg \max_A \frac{\text{tr}(A^{\mathrm{T}} S_b A)}{\text{tr}(A^{\mathrm{T}} S_w A)}, \tag{1}$$

where $A \in \mathbb{R}^{n \times m}$ is the projection matrix we seek, $\text{tr}(\cdot)$ is the trace operator, $S_w \in \mathbb{R}^{n \times n}$ is the *within-class* scatter matrix, and $S_b \in \mathbb{R}^{n \times n}$ is the *between-class* scatter matrix. The within-class scatter matrix is given by

$$S_w = \sum_{i=1}^{C} P_i \sum_{j=1}^{N_i} (\mathbf{x}_j^{(i)} - \mu_i)(\mathbf{x}_j^{(i)} - \mu_i)^{\mathrm{T}}, \tag{2}$$

where C is the total number of classes, N_i is the number of samples in class C_i, P_i is the prior probability of C_i, $\mathbf{x}_j^{(i)} \in \mathbb{R}^n$ is the jth vector of C_i, and $\mu_i \in \mathbb{R}^n$ is the mean of C_i. The between-class scatter matrix is given by

$$S_b = \sum_{i=1}^{C} P_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}, \tag{3}$$

where $\mu \in \mathbb{R}^n$ is the ensemble mean. We note that $\text{rank}(S_b) \leqslant C - 1$ since it is the sum of C rank-one or zero (if $\mu_i = \mu$) matrices, where at most $C - 1$ are linearly independent. For convenience, and without loss of generality, we assume that $\text{rank } S_b = C - 1$ for the remainder of this paper. The intuitive interpretation of Eq. (1) is that T-LDA attempts to simultaneously minimize the within-class scatter and maximize the between-class scatter. Perhaps, the most common approach for solving Eq. (1) is to solve the generalized eigen-problem of S_b and S_w. This solution can be achieved by simultaneously diagonalizing S_w and S_b [23]. The simultaneous diagonalization process is accomplished (assuming S_w is non-singular) by whitening S_w, diagonalizing the resulting S_b, and then taking the largest

eigenvalue eigenvectors of S_b. Intuitively, this process can be described as whitening the denominator of Eq. (1) and then maximizing the numerator over a reduced dimensionality. The converse approach of whitening the numerator and minimizing the denominator is equivalent, but recall that S_b is generally singular and cannot be whitened.

2.1. W-LDA

As alluded to in Ref. [23] and discussed in Refs. [22] and [24], the class separability criteria that T-LDA maximizes is the Euclidean distance between the class means. Euclidean distance, of course, is not necessarily representative of classification accuracy, and its use as the separability measure can cause some classes to unnecessarily overlap in the reduced space. Two similarly motivated solutions to this problem have been proposed: weighted pairwise Fisher criteria [22] and fractional-step LDA [24]. Although quite effective [25], fractional-step LDA is iterative and very time-consuming; hence, we adopt the weighted pairwise Fisher criteria in this paper which allows for a direct solution. To begin, we first note an alternate expression for S_b [22] (equivalence is proven in Appendix A):

$$S_b = \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} P_i P_j \alpha(\varDelta_{ij})(\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}}, \tag{4}$$

where P_i and P_j are the class priors, $\varDelta_{ij}$ is a measure of the separation between classes C_i and C_j, $\alpha(\cdot)$ is some weighting function, and setting $\alpha(\cdot) = 1$ makes Eqs. (4) and (3) equivalent.

In Ref. [22], weighted pairwise Fisher criteria are proposed and we refer to the resulting algorithm as *weighted* LDA or W-LDA. In W-LDA, the Mahalanobis distance is selected for the class separation measure $\varDelta_{ij}$:

$$\varDelta_{ij} = \sqrt{(\mu_i - \mu_j)^{\mathrm{T}} S_w^{-1} (\mu_i - \mu_j)} \tag{5}$$

and the weighting function, $\alpha(\cdot)$ in Eq. (4) above, is selected so that the contribution of each pair of classes depends (approximately) upon the Bayes error rate between the classes, yielding:

$$\alpha(\varDelta_{ij}) = \frac{1}{2\varDelta_{ij}^2} \, \text{erf}\left(\frac{\varDelta_{ij}}{2\sqrt{2}}\right). \tag{6}$$

Once S_b has been computed in this manner, we simply apply the same procedure as T-LDA.

2.2. D-LDA

One problem often encountered with LDA in practice is that the original feature vectors may be of such high dimensionality that the storage and/or eigen-analysis of S_b and S_w may be impractical. In such applications some other form of dimensionality reduction—usually PCA in

the face recognition case [6,7,26]—is performed prior to LDA. PCA, however, does not consider class labels and can decrease discriminative capability. In Ref. [21], an LDA algorithm—*direct* LDA or D-LDA—that can be directly applied to high-dimensional data is presented.

The critical idea that enables D-LDA is to first project all samples in $\mathbb{R}^n$ onto the $C - 1$ dimensional column-space of S_b (i.e., discard the nullspace of S_b). This is motivated by assuming that directions along which there is no between-class scatter are not useful for discrimination. Although this assumption is not entirely true—since the scatter matrix is parameterized by only the class means—results [21] indicate the approach is still effective. In many high-dimensional problems, the number of classes, C, is much smaller than the dimensionality of the vectors, n. Recalling that rank$(S_b) = C - 1$, we can reduce the dimensionality of the problem from n to $C - 1$ by projecting onto the column-space of S_b. By discarding the nullspace of S_b, the between-class scatter matrix in the reduced space is full rank. We may then use the simultaneous diagonalization approach mentioned above, where we whiten the numerator of Eq. (1) and minimize the denominator. According to Refs. [27,21], this permits us to preserve the nullspace of S_w, which contains the most discriminative information.

As stated above, the first step in D-LDA is to find a basis for the $C - 1$ dimensional column-space of S_b. Recall that S_b is an $n \times n$ matrix, which might imply a significant computational burden if n is large. Fortunately, the $C - 1$ eigenvectors of S_b corresponding to the $C - 1$ non-zero eigenvalues can be found by solving a much more tractable $C \times C$ problem [23] which we review at the end of Section 2.3.

2.3. Combining D-LDA and W-LDA

From the discussion in the previous section, it would certainly be desirable to exploit the benefits of W-LDA and D-LDA simultaneously. There are, however, a couple of potential issues that must be recognized and overcome. First, we note that the computation of S_b for W-LDA, as given by Eq. (4), first requires the computation of S_w, which is a large $n \times n$ matrix; this computation would defeat the computational savings of D-LDA. S_w is required since Mahalanobis distance is used for Δ_{ij} and, as shown in Eq. (5), S_w^{-1} is needed in the computation. Noting the need for S_w^{-1} leads us to another potential difficulty; one of the primary motivations for D-LDA was the preservation of the nullspace of S_w. If the nullspace of S_w is non-empty, however, then S_w^{-1} does not exist.

Noting that a step-by-step algorithmic description is given at the end of this section, we propose the following approach to address these problems. First, recalling (4)–(6), we make the mild assumption that $\alpha(\Delta_{ij}) > 0$. Note that this assumption implies that no two classes have equal means ($\mu_i \neq \mu_j$) and that S_w^{-1} exists, or is replaced with an alternative. In this

case, the nullspaces of S_b from Eqs. (4) and (3) are equivalent (see proof in Appendix B). Hence, we can remove the nullspace by projecting all the data onto the $C - 1$ dimensional column-space of S_b. Recall that the column-space of S_b can be found by eigen-analysis of a much more tractable $C \times C$ matrix. Once we have projected to the $C - 1$ column-space, we compute S_w in the reduced space and, if it is non-singular, we simply proceed with W-LDA as described above.

If, however, S_w is indeed singular in the column-space of S_b, we must compute Δ_{ij} differently for the W-LDA portion. At this point it is interesting to note that there is an apparent contradiction between D-LDA and W-LDA regarding the nullspace of S_w. Suppose, for example, that S_w (in the $r = C - 1$ dimensional column-space of S_b) is full rank, but with one small eigenvalue $\lambda_r = \varepsilon \to 0$. Now suppose there exist two classes whose mean difference vector, $\mu_i - \mu_j$, has a non-zero component in the direction of ϕ_r, the eigenvector of S_w corresponding to the small eigenvalue λ_r. Recalling the weighting function for W-LDA, as specified by Eqs. (5) and (6), we see that as $\varepsilon \to 0$, $\Delta_{ij} \to \infty$, and hence $\alpha_{ij} \to 0$. In other words, any vectors with components in the nullspace of S_w receive minimal weighting in W-LDA. This, of course, contradicts D-LDA which claims these directions to be the most important. The solution we propose is to alter Δ_{ij} so that it is equal to Mahalanobis distance in the column-space of S_w and Euclidean distance in the nullspace. We do this by simply setting the zero eigenvalues of S_w to 1 when computing S_w^{-1} to yield a regularized inverse

$$\hat{S}_w^{-1} = \sum_{i=1}^{r-d} \frac{1}{\lambda_i} \phi_i \phi_i^{\mathrm{T}} + \sum_{i=r-d+1}^{r} \phi_i \phi_i^{\mathrm{T}}, \tag{7}$$

which is then used in place of S_w^{-1} for Δ_{ij} in Eq. (5).

We note that further investigation is perhaps warranted to determine an optimal method to replace the zero eigenvalues of S_w for the purposes of computing Δ_{ij}. W-LDA seems to suggest that the zero eigenvalues should be replaced with a small number (perhaps the minimum of the non-zero eigenvalues) so that any vectors with components in the nullspace of S_w are minimally weighted. D-LDA, which is premised on the importance of the nullspace of S_w, seems to suggest that zero (or even near zero) eigenvalues should be replaced with a large number (perhaps the maximum eigenvalue) so that such vectors are heavily weighted. As a compromise, we replace the zero eigenvalues with 1, which simply equates to Euclidean distance (which has been effective for T-LDA) in the nullspace of S_w.

As a final point of interest concerning this issue, we note that the nullspace of S_w is generally empty in the column-space of S_b. This is somewhat ironic, considering the claimed significance of this nullspace expressed in Ref. [21]. We can see from Eq. (2) that S_w is the sum of several outer products and that generally rank$(S_w) \geqslant C - 1$, even when projected onto the column-space of S_b, so long as

we have at least two samples per class. Recalling that the column-space of S_b is of dimension $C-1$, it seems that S_w will usually be non-singular in this space with as little as two training samples per class. In fact, in our experiments, S_w was never singular in the column-space of S_b. Although this fact seems to minimize the claimed D-LDA benefit of allowing the preservation of the nullspace of S_w, it does not diminish the computational benefit. We need only to find the eigenvalues and eigenvectors of a $C \times C$ matrix ($C = 128$ in our experiments), as opposed to an $n \times n$ matrix (where n is the number of pixels; as many as 10,580 in our experiments), even if abundant training data exists. Furthermore, it is conceivable that there will exist problems where the feature distributions do indeed make S_w singular in the column-space of S_b. In such situations, D-LDA will certainly be beneficial.

We can now describe the complete DW-LDA algorithm with the following six steps.

1. Let $B \in \mathbb{R}^{n \times r}$ be a orthonormal basis for the column-space of S_b^o, the between-class scatter matrix in the original space. Remove the nullspace of the between-class scatter matrix by projecting all samples onto B

   $$\mathbf{x} \in \mathbb{R}^n \rightarrow B^{\mathrm{T}} \mathbf{x} \in \mathbb{R}^r.$$

2. Compute S_w in the reduced space $\mathbb{R}^r$. If S_w is full rank compute S_w^{-1}, otherwise compute $\hat{S}_w^{-1}$ using Eq. (7).
3. Compute S_b using Eq. (4) with α_{ij} given by Eq. (6) and Δ_{ij} given by Eq. (5). If S_w is singular, then use $\hat{S}_w^{-1}$ when computing Δ_{ij}.
4. Whiten S_b:

   $$S_b \rightarrow W^{\mathrm{T}} S_b W = I_{r \times r},$$
   $$S_w \rightarrow \tilde{S}_w = W^{\mathrm{T}} S_w W,$$

   where $W = \Psi \Gamma^{-1/2}$ is the whitening transformation of S_b with Ψ being the eigenvectors of S_b and Γ the diagonal eigenvalue matrix.

5. Diagonalize $\tilde{S}_w$:

   $$\tilde{S}_w \rightarrow D_w = V^{\mathrm{T}} \tilde{S}_w V,$$

   where D_w is the diagonal eigenvalue matrix of $\tilde{S}_w$ and V contains the corresponding orthonormal eigenvectors.

6. Assume that the eigenvalues and eigenvectors of D_w and V are sorted in ascending order, possibly with some zeros in D_w. To maximize the LDA criterion in Eq. (1) while reducing to dimensionality m, take the first m columns of V which correspond to the m lowest (some possibly zero) eigenvalues. The overall resulting transformation (i.e., projection) matrix $A \in \mathbb{R}^{n \times m}$ can then be written as follows:

   $$A = BWV \begin{pmatrix} I_{m \times m} \\ 0_{(r-m) \times m} \end{pmatrix}. \tag{8}$$

Note that in Step 1 above we project onto the column-space of S_b^o; this requires us to compute its eigenvectors corresponding to non-zero eigenvalues. Although S_b^o can be very large ($n \times n$), recall that $\mathrm{rank}(S_b^o) \leqslant C-1$ and generally $C \ll n$. Referring back to Eq. (3), we note that S_b^o can also be written as follows:

$$S_b^o = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}} = UU^{\mathrm{T}}, \tag{9}$$

where the ith column of $U \in \mathbb{R}^{n \times C}$ is $(\mu_i - \mu)$. The eigenvectors of the possibly very large UU^{T} can be computed simply from the eigenvectors of the generally much smaller $U^{\mathrm{T}}U$ as follows [23]. The eigenvector decomposition of $U^{\mathrm{T}}U$ yields

$$(U^{\mathrm{T}}U)\Psi = \Psi E. \tag{10}$$

Pre-multiplying both sides by U gives

$$(UU^{\mathrm{T}})(U\Psi) = (U\Psi)E, \tag{11}$$

where the columns of $(U\Psi)$ corresponding to non-zero eigenvalues in E give the eigenvectors of UU^{T} that we seek. Although orthogonal, note that the ith column of $(U\Psi)$ must be normalized by $(e_i)^{-1/2}$—where e_i (non-zero) is the ith diagonal element of E—to make the columns orthonormal.

3. Modular Fisherfaces

Perhaps, the earliest suggestions for the use of modular subspaces can be found in Ref. [1] and also in the view-based and modular eigenspaces of Ref. [14]. It is noted in Ref. [14], and observed in much research since, that recognition from a frontal face image is sensitive to changes in expression, decoration, and illumination. By decomposing the full face image into modular subregions, Ref. [14] shows that improved accuracy can be obtained with respect to expression and decoration variation. An extension to the view-based eigenspaces of Ref. [14], along with more comprehensive experimental data showing much promise, is found in Ref. [15].

We propose a face recognition system employing modular LDA subspaces, i.e., modular Fisherfaces. Through the survey of previous work in Section 1 and the previous modular subspace efforts discussed above, this system should provide improved robustness to three of the four confounding factors—illumination, expression, and decoration—simultaneously, as described shortly. This framework employs a parallel system of observers, each of which is trained on a specific (modular) region of the face from a specific viewpoint (all frontal for the work herein). Each such observer is a linear subspace classifier and the outputs of all the observers are combined using a simple sum-rule, classifier combination strategy [28]. The modular observers [14] along with the use of LDA should provide
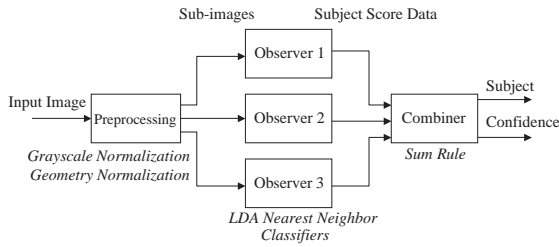
Fig. 1. Example of the proposed framework with three observers.

Fig. 2. Examples of the modular face image regions used by the three observers in the proposed system.

some robustness to decoration. Improved robustness to illumination variation should also be provided by LDA [6]. Expression should be addressed through the combination of multiple observers and LDA—some observers will be approximately invariant to expression, while LDA will discount expression variability in the training set (since it is not discriminatory) when constructing the projection matrices. Finally, additional overall robustness should be provided by integrating the responses of all the observers to obtain a final classification. A simple illustration of a system with three observers is illustrated in Fig. 1. In the following subsections, we describe the components of this system in more detail.

3.1. Preprocessing

The preprocessing stage indicated in Fig. 1 assumes a frontal face image as input with previously labeled landmarks. After geometry normalization based upon the labeled landmarks, an elliptical mask is applied to remove background information. Each image is subsequently normalized to have zero-mean and unit variance to account for gross gray-scale variation. The resulting images are raster scanned to create the feature vectors for input into the observers.

3.2. Observers

Observer 1 uses 92×115 pixel full frontal face images implying an original feature vector length of $n_1 = 10,580$. Observer 2 uses 92×56 pixel images of the eyes and nose face region, implying an original feature vector length of $n_2 = 5152$. Observer 3 works with 92×40 pixel images of the eyes region, implying an original feature vector length of $n_3 = 3680$. Examples of images corresponding to each observer are shown in Fig. 2. These observers were selected based upon earlier work in Refs. [8] and [14] and intuition. For example, in the presence of unexpected lower facial hair, humans can still recognize a face by focusing upon the eyes and nose region. Similarly, in the presence of confusing eyeglasses, humans can still perform recognition based on other, whole-face features. The automatic selection of such observers might be an interesting avenue for further research.
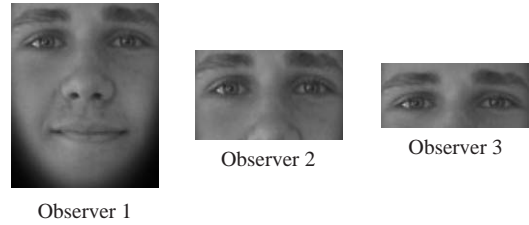
In the current implementation, each observer performs a simple nearest-neighbor search, using Euclidean distance, on its own local database. For query image q and each observer o, a score p_{qs}^o is computed for each subject s and passed on to the combiner. The score for observer o is given by

$$p_{qs}^o = (F_q^o d_{qs}^o)^{-1}, \tag{12}$$

where d_{qs}^o is the distance between the query image q and its nearest neighbor in subject class s, and the normalization factor F_q^o is given by

$$F_q^o = \sum_{s=1}^{S} (d_{qs}^o)^{-1}, \tag{13}$$

where S is the number of subjects, so that

$$\sum_{s=1}^{S} p_{qs}^o = 1 \tag{14}$$

Hence p_{qs}^o is, in some sense, a rough approximation to the probability that query image q belongs to subject class s in the space seen by observer o. Without combining the observers, nearest-neighbor classification is implemented for each observer individually by simply selecting the subject class with the highest score.

3.3. Observer combination

For each query image q, the score data for each subject s from each observer o is passed to the combiner. Since the score data has been normalized so that it approximates a probability, a number of simple classifier combination strategies can be employed. In our implementation, we employ the sum-rule. As noted in Ref. [28], the implicit assumptions that make the sum-rule optimal are quite restrictive. Despite this fact, the sum-rule is reported in Ref. [28] to be the best performing. This is attributed to the sum-rule being more resilient to estimation errors. Obviously, the interested reader could explore various other combination strategies that have been well-documented in the literature. According

to the sum-rule, a combined score P_{qs} for each subject s is computed by simply adding the scores for subject s reported by each observer:

$$P_{qs} = \sum_{o=1}^{3} p_{qs}^{o}. \tag{15}$$

The subject with the highest combined score is then selected as the classification result for the query q.

In our implementation, we also compute a confidence measure m_q as an additional combiner output, similar to the approach of Ref. [29]. For a given query image q, let the highest combined score be denoted P_{qs_1} and the second highest combined score be denoted P_{qs_2}. The confidence measure m_q is computed as the logarithm of the ratio of the highest score to the second-highest score:

$$m_q = \log\left(\frac{P_{qs_1}}{P_{qs_2}}\right). \tag{16}$$

In applications where more than a single observation of a subject is expected and/or multiple poses are observed simultaneously (e.g., video surveillance in a controlled access environment), the confidence factor can be used for weighting the integration of decisions through time and/or across views. In the results of Section 4, we illustrate the potential effectiveness of this confidence metric.

3.4. Extension to pose variation

Finally, recall that we earlier mentioned the extension of the presented work to the pose variation problem. Based on the success of previous work in Ref. [14] and particularly Ref. [15], we hypothesize that the extension of our approach to handle varying pose is quite straightforward. In addition to variable-pose training data, the extension would only require observers (i.e., additional modular LDA subspaces) for alternate poses and a pose estimator in the preprocessing stage. Classifier combination could then be carried out using modular image regions of the same pose as well as across different poses. Since pose variation is not an issue in our problem domain, this topic is presently unaddressed in our research.

4. Experimental results

In this section, we present experimental results using our modular system and compare the performance of several subspace projection algorithms. Recall from Section 1, that our target application requires recognition of between 100 and 150 people, with good training data available for each person. With this fact in mind, we selected two publicly available databases that contained data most appropriate to

our problem of interest. The first database was the CVL database [30], which nominally consisted of 114 persons with three frontal views each and included variations in expression (neutral, smile without teeth showing, smile with teeth showing). In this database, subject number 56 was a duplicate of subject number 50 and was therefore removed. Subject numbers 35, 44, and 93 had only two frontal images available each. One image for subject number 25 was discarded because the pose was significantly deviated from frontal. Hence, from the CVL database we had 113 subjects with a total of 335 images. The other database employed was the Yale Face Database [31]. This database comprised 11 images each of 15 subjects. The 11 images per subject included variations in expression, illumination, and decoration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. All of the "surprised" (open-mouth as if screaming) data was excluded because it differed significantly from what we expect to see in our problem domain of secure facility monitoring. Furthermore, the "with glasses" image for subject number 3 was discarded because the glasses were opaque and occluded the requisite landmarks. Hence, from the Yale Face Database, we have 15 subjects and a total of 149 images. Our complete database, comprising both the CVL and Yale data, hence contained 128 subjects and 484 images.

Five different subspace projection algorithms were tested, each reducing the original features spaces (of dimension 10,580, 3152, and 3680 for observers 1, 2, and 3, respectively) to 50 dimensions. A series of 100 training and classification runs were performed. In each run, all but one randomly selected image for each subject was used for training and testing was performed on the remaining 128 images (one for each subject). The five algorithms tested were: PCA, PCA/LDA, PCA/W-LDA, D-LDA, and DW-LDA. In PCA, which is the well-known "eigenface" approach, we simply select the first 50 principal components. In PCA/LDA, also known as the "Fisherface" approach, we first use PCA to project to a 107-dimensional subspace ($\frac{1}{3}$ of the possible $484 - 128 = 356$ from the training data) and then apply T-LDA to reduce the dimension to 50. PCA/W-LDA is implemented in the same manner, but we replace T-LDA with the W-LDA algorithm. For D-LDA, we use the direct LDA algorithm of Section 2.2 to project to dimension 50. For DW-LDA we employ the new direct, weighted LDA algorithm presented in Section 2.3. Recognition performance for each of the five algorithms is reported in Table 1 for each observer individually as well as all three combined (using the sum-rule combiner). Note that traditional "eigenfaces" is represented by the Observer 1 score for PCA and that traditional "Fisherfaces" is represented by the Observer 1 score for PCA/LDA.

From these results we see that the algorithms employing W-LDA perform better—although only marginally so—than those that employ traditional LDA. It is interesting to note,

Table 1
Correct classification percentages for five different subspace projection algorithms

	1	2	3	Combined
PCA	60.8	72.4	86.6	83.5
PCA/LDA	91.9	88.7	93.0	95.1
PCA/W-LDA	92.1	89.2	93.4	95.3
D-LDA	85.6	78.6	87.3	91.6
DW-LDA	87.3	80.2	88.1	92.3

Results are reported for the three observers individually as well as the combination. Note that traditional "eigenfaces" corresponds to the Observer 1 score for PCA (60.8%) and "Fisherfaces" corresponds to the Observer 1 score for PCA/LDA (91.9%).

however, that in all cases D-LDA performs worse than PCA/LDA and, similarly, DW-LDA performs worse than PCA/W-LDA. These results are significantly different than earlier results [21] that used a smaller facial image database. Our results seem to indicate that D-LDA, while allowing the direct application of LDA to high-dimensional data, does not generalize as well as the PCA-first algorithms. One explanation for this is that the direct LDA-based algorithms tend to preserve noise that is discriminative in the training set, but that does not exist in the testing set; PCA tends to discard such noise. The value of D-LDA or DW-LDA should not necessarily be discounted because of this, though. The direct LDA approaches requires only the eigenvalues and eigenvectors of a $C \times C$ (128×128 in our experiments) matrix regardless of the total number of the training samples. PCA, on the other hand, requires the eigenvalues and eigenvectors of at least an $N \times N$ matrix in the small sample size situation (recall N is the number of training samples; $N = 356$ in our experiments) and up to those of an $n \times n$ matrix with a complete training set (recall n is the original dimensionality, up to $n = 10,580$ in our experiments). Obviously, PCA can become computationally challenging as N increases, while the complexity of D-LDA or DW-LDA only increases with the number of classes, C. As a topic for future research, this characteristic allows for augmentation of the training set, by adding synthetically altered versions of the training images (i.e., misalignments, occlusions, noise, etc.), to account for potential error sources such as additive noise and/or misalignment. Also evident from the data in Table 1 is that Observer 3 (the eyes-only region) invariably performs better than either Observer 1 or Observer 2 alone. This indicates experimentally the fact that the eyes tend to be somewhat invariant to illumination, expression, and decoration (such as eyeglasses) that does not occlude the eyes and/or eyebrows. A similar result was reported previously [14] in the PCA case. Another factor that contributes to the improved performance of Observer 3 is the fact that the Observer 3 images are smaller and hence the retained 50 dimensions
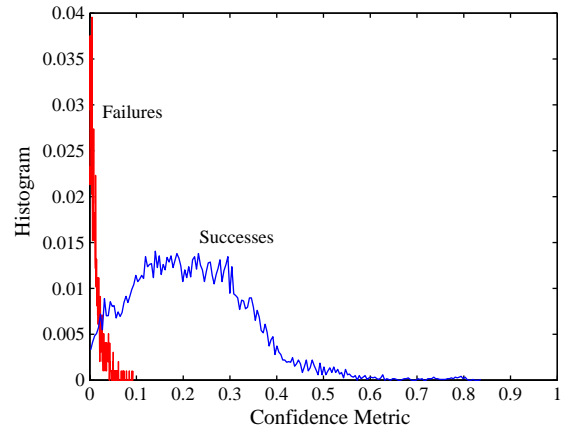

Fig. 3. Histograms of confidence metric for successful identifications and failed identifications using DW-LDA. In our area of interest, where multiple surveillance cameras will be installed, low-confidence identifications can be discarded.

represent a larger percentage of the total variation in the image set.

We now turn our attention briefly to the confidence measure discussed in Section 3.3. In Fig. 3, we plot the distribution of the confidence metric for successful identifications and failed identifications for the DW-LDA algorithm. We can use the confidence metric to reject decisions below a certain confidence level. Using this idea, we plot the accuracy of the different algorithms versus the rate of low-confidence rejected images in Fig. 4. Note that the acceptable confidence threshold increases to the right along with the rate of rejected images. The plot in Fig. 4 indicates that if 10% of the images were rejected due to low confidence, DW-LDA (dash–dot line) would achieve about 99% accuracy on the remaining data, while PCA/LDA and PCA/W-LDA would achieve nearly 100%. At the same low-confidence rejection rate, PCA would achieve slightly less than 90% accuracy. Also evident in the plot is that DW-LDA consistently performs better than D-LDA (dashed line), but only by about 0.7%. This use of the confidence metric can be very effective in secure-facility video surveillance applications, where multiple observations of a subject can easily be obtained via multi-camera video surveillance. We can continue to acquire data until a high-confidence observation is obtained or alert security personnel if no such observation is acquired after some preset time interval.

5. Conclusions

In this paper, we present a modular LDA subspace approach for template-based face recognition that performs
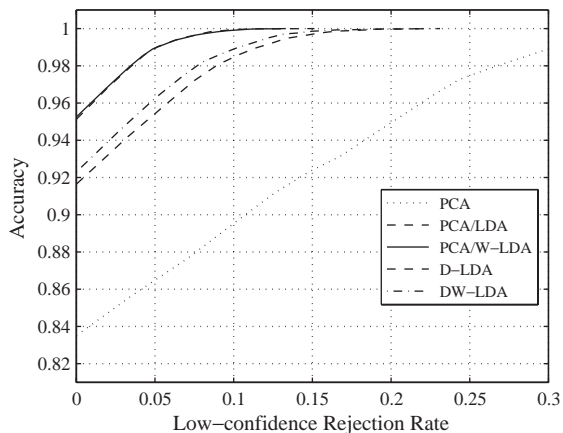
Fig. 4. Accuracy of different subspace projection algorithms versus rate of images rejected due to low confidence. This plot indicates that if 10% of the images were rejected due to low confidence, DW-LDA (dash–dot line) would achieve about 99% accuracy on the remaining data, while PCA/LDA and PCA/W-LDA would achieve nearly 100%. At the same low-confidence rejection rate, PCA would achieve slightly less than 90% accuracy. Note that PCA/LDA and PCA/W-LDA are the best performing methods and are essentially indistinguishable on this plot (top curves). Also note that DW-LDA consistently performs better than D-LDA (dashed line), but only by about 0.7%.

significantly better than traditional "eigenfaces" or "Fisherfaces." This approach is specifically aimed at addressing three (illumination, expression, and decoration) of the four confounding factors of interest. Although pose is not specifically addressed in the presented work, we briefly describe how the system might be extended to handle pose variation. We also present a new LDA-based subspace projection algorithm that unifies direct LDA (D-LDA) and the weighted pairwise Fisher criteria (W-LDA) in a single algorithm we refer to as direct, weighted LDA (DW-LDA). Experimental results indicate that DW-LDA performs better than D-LDA. Our results also indicate that D-LDA-based approaches, while providing significant potential computational savings, are actually outperformed by the well-known, PCA-first LDA approaches in terms of classification accuracy.

Acknowledgements

Appendix A

Here we show that Eq. (4) is equivalent to Eq. (3) when $\alpha(\Delta_{ij}) = 1$ for all (i, j). We begin with Eq. (4):

$$
\begin{aligned}
S_b &= \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} P_i P_j (\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}} \\
&= \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j (\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}} \\
&= \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j ((\mu_i - \mu) + (\mu - \mu_j)) \\
&\quad \times ((\mu_i - \mu) + (\mu - \mu_j))^{\mathrm{T}} \\
&= \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j ((\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}} \\
&\quad + (\mu_i - \mu)(\mu - \mu_j)^{\mathrm{T}} + (\mu - \mu_j)(\mu_i - \mu)^{\mathrm{T}} \\
&\quad + (\mu - \mu_j)(\mu - \mu_j)^{\mathrm{T}}).
\end{aligned}
$$

Since $\sum_{i=1}^{C} P_i = 1$, we can combine the first and last outer product terms above to get

$$
\begin{aligned}
S_b &= \sum_{i=1}^{C} P_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}} \\
&\quad + \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j (\mu_i - \mu)(\mu - \mu_j)^{\mathrm{T}} \\
&\quad + \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j (\mu_j - \mu)(\mu - \mu_i)^{\mathrm{T}}.
\end{aligned}
$$

Examining the last two terms above, we note that $\sum_{i=1}^{C} P_i \mu_i = \mu$ and therefore $\sum_{i=1}^{C} P_i (\mu_i - \mu) = 0$. We are then left with only the first term which is exactly Eq. (3).

Appendix B

Here we provide a proof that demonstrates that the nullspaces of S_b from Eqs. (3) and (4) are equivalent when $\alpha(\Delta_{ij}) > 0$. Define

$$
A = \sum_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \tag{17}
$$

and

$$
B = \sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}, \quad \alpha_i > 0 \ \forall i. \tag{18}
$$

We denote the nullspace of A as $\mathcal{N}(A)$ and the nullspace of B as $\mathcal{N}(B)$. Now suppose that $\mathbf{v} \in \mathcal{N}(A)$ and $\mathbf{v} \neq 0$.

Then

$$A\mathbf{v} = 0 \implies \mathbf{v}^{\mathrm{T}} A \mathbf{v} = 0$$

$$\implies \mathbf{v}^{\mathrm{T}} \left(\sum_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \mathbf{v} = 0$$

$$\implies \sum_i (\mathbf{x}_i^{\mathrm{T}} \mathbf{v})^2 = 0$$

$$\implies \mathbf{x}_i^{\mathrm{T}} \mathbf{v} = 0 \ \forall i.$$

Therefore,

$$B\mathbf{v} = \sum_i \alpha_i \mathbf{x}_i (\mathbf{x}_i^{\mathrm{T}} \mathbf{v}) = 0$$

so

$$\mathbf{v} \in \mathcal{N}(A) \implies \mathbf{v} \in \mathcal{N}(B).$$

Similarly, now suppose that $\mathbf{v} \in \mathcal{N}(B)$ and $\mathbf{v} \neq 0$. Then

$$B\mathbf{v} = 0 \implies \mathbf{v}^{\mathrm{T}} B \mathbf{v} = 0$$

$$\implies \mathbf{v}^{\mathrm{T}} \left(\sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \mathbf{v} = 0$$

$$\implies \sum_i (\alpha_i \mathbf{x}_i^{\mathrm{T}} \mathbf{v})^2 = 0$$

$$\implies \mathbf{x}_i^{\mathrm{T}} \mathbf{v} = 0 \ \forall i.$$

Therefore,

$$A\mathbf{v} = \sum_i \mathbf{x}_i (\mathbf{x}_i^{\mathrm{T}} \mathbf{v}) = 0$$

so

$$\mathbf{v} \in \mathcal{N}(B) \implies \mathbf{v} \in \mathcal{N}(A).$$

Hence,

$$\mathbf{v} \in \mathcal{N}(A) \iff \mathbf{v} \in \mathcal{N}(B).$$

References

[1] R. Brunelli, T. Poggio, Face recognition: features versus templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993) 1042–1052.

[2] L. Wiskott, J. Fellous, N. Kruger, C. Von der Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 775–779.

[3] J. Fortuna, D. Capson, Improved support vector classification using PCA and ICA feature space modification, Pattern Recognition 37 (6) (2004) 1117–1129.

[4] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1063–1074.

[5] R. Gross, I. Matthews, S. Baker, Appearance-based face recognition and light-fields, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4) (2004) 449–465.

[6] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[7] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 336–341.

[8] W. Zhao, R. Chellappa, P.J. Phillips, Subspace linear discriminant analysis for face recognition, Technical Report CS-TR-4009, Center for Automation Research, University of Maryland, April 1999.

[9] G.J. Edwards, T.F. Cootes, C.J. Taylor, Face recognition using active appearance models, in: Proceedings of the European Conference on Computer Vision, 1998, vol. 2, pp. 581–595.

[10] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognition 33 (11) (November 2000) 1771–1782.

[11] B. Moghaddam, C. Nastar, A. Pentland, Bayesian face recognition using deformable intensity surfaces, in: Proceedings of the IEEE Computer Society Conference on Pattern Recognition, 1996, pp. 638–645.

[12] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 696–710.

[13] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 11 (4) (2002) 467–476.

[14] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: Proceedings of the IEEE Computer Society Conference on Pattern Recognition, 1994, pp. 84–91.

[15] F.J. Huang, Z. Zhou, H.-J. Zhang, T. Chen, Pose invariant face recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 245–250.

[16] D.B. Graham, N.M. Allinson, Face recognition from unfamiliar views: subspace methods and pose dependency, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 348–353.

[17] T.F. Cootes, K. Walker, C.J. Taylor, View-based active appearance models, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 227–232.

[18] T.F. Cootes, G.V. Wheeler, K.N. Walker, C.J. Taylor, Coupled-view active appearance models, in: Proceedings of the British Machine Vision Conference, 2000, vol. 1, pp. 52–61.

[19] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.

[20] W.Y. Zhao, R. Chellappa, SFS based view synthesis for robust face recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 285–292.

[21] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.

[22] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Trans. Pattern Anal. Mach. Intell. 23 (7) (2001) 762–766.

[23] K. Fukunaga, Statistical Pattern Recognition, Morgan Kaufmann, Los Altos, CA, 1990.

[24] R. Lotlikar, R. Kothari, Fractional-step dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 623–627.

[25] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using LDA-based algorithms, IEEE Trans. Neural Networks 14 (1) (2003) 195–200.

[26] D.L. Swets, J. (Juyang) Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.

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

[28] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[29] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Anal. Mach. Intell. 17 (10) (1995) 955–966.

[30] The CVL Face Database, www.lrv.fri.uni-lj.si, Peter Peer (Peter.Peer@fri.uni-lj.si).

[31] The Yale Face Database, cvc.yale.edu, Athos Georghiades (georghiades@yale.edu).

About the Author—JEFF PRICE received the B.S.E.E. degree from the US Naval Academy in 1993, and the M.S. and Ph.D. degrees in Electrical Engineering from the Georgia Institute of Technology in 1997 and 1999, respectively. Dr. Price is currently with the Image Science and Machine Vision Group at Oak Ridge National Laboratory.

About the Author—TIM GEE received the B.S.E.E. degree from Auburn University in 1992, and the M.S. degree in Electrical Engineering from the Georgia Institute of Technology in 1993. Mr. Gee is currently a member of the Image Science and Machine Vision Group at Oak Ridge National Laboratory.