Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 416318, 7 pages
doi:10.1155/2008/416318

Research Article

Face Recognition Using Classification-Based Linear Projections

Moshe Butman[1] and Jacob Goldberger[2]

[1] *Computer Science Department, Bar-Ilan University, Ramat-Gan 52900, Israel*
[2] *School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel*

Correspondence should be addressed to Moshe Butman, butmanm@cs.biu.ac.il

Received 24 July 2007; Revised 17 January 2008; Accepted 19 February 2008

Recommended by C. Charrier

Subspace methods have been successfully applied to face recognition tasks. In this study we propose a face recognition algorithm based on a linear subspace projection. The subspace is found via utilizing a variant of the neighbourhood component analysis (NCA) algorithm which is a supervised dimensionality reduction method that has been recently introduced. Unlike previously suggested supervised subspace methods, the algorithm explicitly utilizes the classification performance criterion to obtain the optimal linear projection. In addition to its feature extraction capabilities, the algorithm also finds the optimal distance-metric that should be used for face-image retrieval in the transformed space. The proposed face-recognition technique significantly outperforms traditional subspace-based approaches particulary in very low-dimensional representations. The method performance is demonstrated across a range of standard face databases.

1. INTRODUCTION

In recent years, automatic face recognition has become one of the most active research fields in computer vision and a large number of different recognition algorithms have been developed. Face recognition algorithms can be categorized into feature-based, holistic-based and hybrid-matching algorithms. In feature-based methods, local features such as the eyes, nose, and mouth are first extracted and their locations and local description are fed into the recognition system (e.g., [1, 2]). Hybrid-matching methods use a combination of global and local features for face recognition (e.g., [3, 4]). In another aspect, face recognition algorithms can be categorized into 2D, 3D and multimodal algorithms [5]. A comprehensive survey of face-recognition algorithms is given by Zhao et al. [6]. The most successful approaches, however, seem to be those appearance-based methods that operate directly on the face images. An image is considered as a high-dimensional vector, that is, a point in a high-dimensional vector space and the set of all faces is assumed to form a low-dimensional manifold. Following this paradigm, face image matching can be viewed as a two-step process of subspace projection followed by classification in the low-dimensional space (see [7] for a recent survey on face recognition in subspaces). In a simple yet successful approach, face recognition is implemented as a linear subspace projection followed by a nearest-neighbour classifier. In particular, the eigenfaces method which is based on principal component analysis (PCA) [8] and the Fisherfaces method based on the Fisher linear discriminant analysis (LDA) [9] have been applied to face recognition with impressive results. PCA-based algorithms select a subspace with maximum variation and they are optimal for object reconstruction. While PCA minimizes sample covariance (second-order dependency), independent component analysis (ICA) minimizes higher-order dependencies as well. The ICA selects a linear projection that maximizes the degree of statistical independence of output variables based on various contrast functions (see [10] for an application of ICA to face recognition). It was experimentally found that face recognition algorithms based on ICA do not offer much improvement over PCA [7]. When a substantial variability in illumination and expression is present, similarity in the transformed space is not necessarily determined by the face identity. Both PCA and ICA construct the reduced face space without using the available face identity information. LDA-based algorithms take the class structure into account and focus on the most discriminant feature extraction. The performance of LDA, however, is often degraded by the fact that its separability criterion is not directly related to the classification accuracy in the transformed space. Instead, the

LDA optimization is based on the assumption that the intra-class distributions are all Gaussian with a common variance. In other words, the LDA assumes, aside from the linearity of the image subspace, a linear separation between classes in the low-dimensional space. (There are many generalizations of the LDA optimization principle but they all impose parametric models on the within-class distributions). The kernel trick can be utilized to form classification algorithms that are based on nonlinear subspaces (e.g., kernel-PCA and kernel-LDA [11]). The basic methodology is to (implicitly) apply a nonlinear mapping on the input images and then apply linear methods on the resulting feature space. Although kernel methods such as SVM achieve state-of-the-art results, in the case of kernel-PCA and kernel-LDA the performance improvement in face recognition tasks over linear methods was not found to be significant.

The LDA approach is based on two assumptions of linearity. It assumes that the face subspace is linear and that there is a linear separation between classes. Kernel methods are based on relaxation of the first assumption. In this paper, we take a different approach. While keeping the linear subspace assumption, we assume no parametric model for the class distributions or the boundaries between them. In this paper, we apply a recently proposed linear subspace method, the neighbourhood component analysis (NCA) [12], to the task of face recognition. The NCA algorithm explicitly utilizes the classification performance criterion to obtain the optimal linear projection. In the original NCA paper [12], the method was applied to standard databases from the UCI repository. In this study, we systematically analyze the benefits of utilizing several variants of the NCA method for face recognition tasks. Unlike other classification problems, these tasks are generally characterized by small sample size on one hand and large sample dimensionality on the other hand. We show experimentally that the NCA approach yields a significant improvement in face-recognition tasks compared to currently used subspace methods.

There is yet another major advantage to the linear subspace method presented here. The fact that the optimization criterion of current subspace methods is not explicitly related to the classification target results in a need for an additional learning procedure that should find a suitable distance function in the transformed subspace [13–15] (e.g., the best results for ICA are obtained using the cosine distance [10]). In the proposed method, the distance measure, that should be used in the transformed subspace, is explicitly stated in the optimization cost function. The optimal transformation is selected such that using the Euclidean distance in the transformed space yields optimal classification results.

We start by presenting several variants of the NCA algorithm in Section 2. Comparative face-recognition experiments on several standard face databases are presented in Sections 3 and 4 contains concluding remarks.

2. LEARNING A LINEAR PROJECTION

In this section, we review the NCA algorithm [12] and focus on a variant that was found to be suitable for face-recognition tasks which often have problems of small sample size and high-dimensional samples. We begin with a labelled dataset consisting of n real-valued input vectors $x_1, \ldots, x_n$ in R^D and corresponding class labels $c_1, ..., c_n$. In the case of face recognition, the vectors are the face images and the labels are the face identities. We want to find a low-dimensional linear transformation $A : R^D \rightarrow R^d$ that maximizes the performance of nearest neighbour classification in the reduced space. Ideally, we would like to optimize performance on future test data, but as we do not know the true data distribution we instead attempt to optimize leave-one-out (LOO) performance on the training data. Given a finite set of linear transformations to choose from, we can easily select the best one, namely the one that minimizes the number of classification errors. The nearest-neighbour classification error, however, is quite a discontinuous function of the transformation $\mathbf{A}$, given that an infinitesimal change in $\mathbf{A}$ may change the neighbour graph and thus affect LOO classification performance by a finite amount. Hence, we can not use this optimization criterion in our case where there is a continuously parameterized family of linear transformations which must be searched. Instead, we adopt a more well-behaved measure of nearest-neighbour performance, by introducing a differentiable cost function based on stochastic ("soft") neighbour assignments in the transformed space. In particular, each point i selects another point j as its neighbour with some probability p_{ij}, and inherits its class label from the point it selects. We define the p_{ij} using a softmax over Euclidean distances in the transformed space:

$$p_{ij}(\mathbf{A}) = \frac{\exp\left(-(1/2)\left\|\mathbf{A}x_i - \mathbf{A}x_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-(1/2)\left\|\mathbf{A}x_i - \mathbf{A}x_k\right\|^2\right)}, \qquad p_{ii} = 0. \tag{1}$$

Note that the norm of matrix A controls the softness of the neighbour assignments. Replacing $\mathbf{A}$ with $\alpha\mathbf{A}$, it can easily be shown that as α tends to infinity, the probabilistic assignment is reduced to deterministic nearest-neighbour assignment in the same transformed space. Denote the set of points in the same class as i by $C_i = \{j \mid c_i = c_j\}$. Under the stochastic selection rule (1), we can compute the probability p_i that a point i will be correctly classified:

$$p_i = \sum_{j\in C_i} p_{ij}. \tag{2}$$

The objective function we maximize is the following:

$$C(\mathbf{A}) = \sum_i \log\left(\sum_{j\in C_i} p_{ij}\right) = \sum_i \log\left(p_i\right). \tag{3}$$

Maximizing this objective would correspond to maximizing the probability of obtaining a *perfect (error-free) classification of the entire training set*. Maximizing the objective function $C(\mathbf{A})$ is also equivalent to minimizing the Kullback-Leibler divergence between the true class distribution (having probability one on the true class) and the stochastic class distribution induced by p_{ij} via A. Note that since $\left\|\mathbf{A}x_i - \mathbf{A}x_j\right\|^2 = (x_i - x_j)^\top \mathbf{A}^\top \mathbf{A}(x_i - x_j)$, the optimization criterion depends

Training:
Input: a set of n labeled face images: $\{x_i \in R^D, c_i\}$ and the reduced dimension d.
Output: a linear projection $\mathbf{A}_{d\times D} : R^D \to R^d$ that maximizes the objective:

$$C(\mathbf{A})\sum_i \log\left(\sum_{j\in C_i} pij\right) \quad \text{with } pij(\mathbf{A}) = \frac{\exp\left(-(1/2)\|\mathbf{A}x_i - \mathbf{A}x_j\|^2\right)}{\sum_{k\neq i}\exp\left(-(1/2)\|\mathbf{A}x_i - \mathbf{A}x_k\|^2\right)}$$

Method:
(i) set initial value for $\mathbf{A}$ (e.g., using the LDA or RCA methods).
(ii) apply a Conjugate-Gradient optimization to find the maximum of $C(\mathbf{A}) - \lambda\|\mathbf{A}\|$

$$\text{s.t.}\frac{\partial(C(\mathbf{A}) - \lambda\|\mathbf{A}\|)}{\partial\mathbf{A}} = \sum_{i,k} p_{ik}\Delta_{ik} - \sum_i \frac{\sum_{j\in C_i} p_{ij}\Delta_{ij}}{\sum_{j\in C_i} p_{ij}} - 2\lambda\mathbf{A}$$

(iii) store $\mathbf{A}$ and the projected training set $\{\mathbf{A}x_i, c_i\}$.
Testing:
predict the label at query input face image x.
(i) find the nearest neighbour $i = \arg\min_j \|\mathbf{A}x - \mathbf{A}x_j\|$ and set the label of x to be c_i.

Figure 1: The proposed face recognition method based on a linear subspace-projection learning algorithm.

only on $\mathbf{A}^\top\mathbf{A}$. Hence, every orthogonal matrix $\mathbf{R}_{d\times d}$ yields a solution $\mathbf{R}\cdot\mathbf{A}$ that is completely equivalent to $\mathbf{A}$. To keep the representation parsimonious we can use the Choleski decomposition representation by forcing the entries of $\mathbf{A}$ below the main diagonal to be zero and the entries on the diagonal to be nonnegative. This makes the representation of $\mathbf{A}$ unique.

Differentiating C with respect to the transformation matrix $\mathbf{A}$ yields a gradient rule which we can use for learning. Observing that

$$\frac{\partial p_{ij}(\mathbf{A})}{\partial\mathbf{A}} = p_{ij}\left(\sum_k p_{ik}\Delta_{ik} - \Delta_{ij}\right), \tag{4}$$

where $\Delta_{ij} = \mathbf{A}(x_i - x_j)(x_i - x_j)^\top$, it can be verified that

$$\frac{\partial C(\mathbf{A})}{\partial\mathbf{A}} = \sum_i\left(\sum_k p_{ik}\Delta_{ik} - \frac{\sum_{j\in C_i} p_{ij}\Delta_{ij}}{\sum_{j\in C_i} p_{ij}}\right). \tag{5}$$

Expression (5) can be viewed as the difference between the overall variability and the intraclass variability defined by the probabilistic model (1) induced from $\mathbf{A}$. The learning algorithm therefore is to maximize the above objective (3) using a gradient-based optimizer such as delta-bar-delta or conjugate gradients. Of course, as the cost function above is not convex, some care must be taken to avoid local maxima during training. We have experimentally observed that the linear transformation obtained by the Fisherfaces (LDA) method can serve as a good starting point for the conjugate gradient algorithm. The linear-transformation learning algorithm is summarized in Figure 1.

In face recognition tasks we often observe the problem of small sample size where the number of the images in the training set (denoted by n) is significantly smaller than the dimensionality of the samples (D). Utilizing the NCA, the small sample size can cause another degeneracy. Assume that $n\cdot d < D$ where d is the dimensionality of the transformed space. In that case, we can easily find a transformation $\mathbf{A}$ that sends all the face images with the same label l to the same (prespecified) point $y_l \in R^d$. We need to solve the linear system

$$\mathbf{A}x_i = y_l, \quad i = 1,\ldots,n \tag{6}$$

such that l is the label of x_i. Since $nd < D$, there are more variables than equations and solutions exist (except for degenerate cases) and can be easily found. Suppose $\mathbf{A}$ solves the linear system (6), then multiplying all the points y_l by a large constant λ, we can obtain a solution $\lambda\mathbf{A}$ such that $p_{ij} = 0$ whenever the labels of x_i and x_j are different. Thus, we can find a transformation that yields a perfect (error-free) classification of the entire training set. To prevent this degeneracy, which can reduce the generalization capabilities of the learning algorithm, we can penalize large-norm transformations $\mathbf{A}$ by adding a regularization term $-\lambda\|\mathbf{A}\|^2$ to the cost function we are maximizing such that λ is a prespecified positive constant that can be set in a cross validation step. The derivative for the regularized cost function is

$$\sum_{i,k} p_{ik}\Delta_{ik} - \sum_i \frac{\sum_{j\in C_i} p_{ij}\Delta_{ij}}{\sum_{j\in C_i} p_{ij}} - 2\lambda\mathbf{A}. \tag{7}$$

Other objective functions based on classification performance can be also considered [12], for example, we can search for a linear transformation that maximizes the expected number of points that are correctly classified. In other words, we can maximize the cost function $\sum_i p_i$. In Section 3, we provide face-recognition results for the two variants of the cost function (for other variants of NCA see [16, 17]).

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method we have conducted a comparative recognition experiments on several standard face databases. It is beneficial to use different kinds of databases because some properties of classification methods, for example, their generalization abilities change depending on the number of classes under consideration and

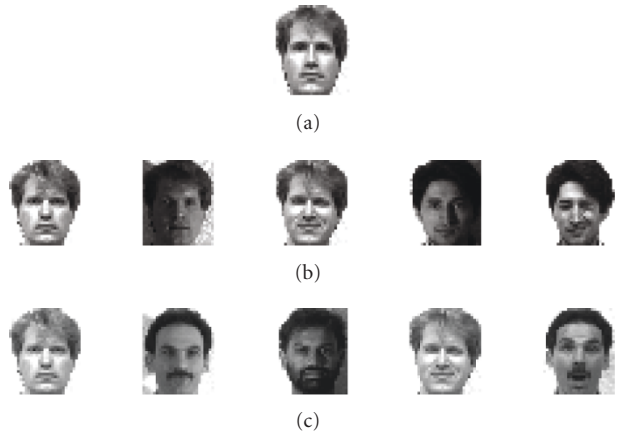


Figure 2: Example of a retrieval query from the Yale database. (a) The query image. Five nearest-neighbour retrieval results obtained using the (b) NCA-based transformation and (c) the Fisherfaces (LDA) method.

the number of training images within each class. For example, Martinez and Kak [18] have shown that when the training is based on a small and nonrepresentative dataset, nondiscriminant methods may outperform to discriminant ones. The datasets we used are the Yale [19], Weizmann [20], and FERET [21]. The Yale University Face Image Database consists of 165 images of 15 different classes (11 images per class). The Weizmann data-base consists of 1916 images of 28 different classes. In the case of the FERET database, many classes consist of exactly two images. Having just one training image is useless for LOO-based training algorithms. We have used two subsets of FERET. The first subset consists of persons with more than 4 images, and the second subset of FERET consists of persons with more than 10 images. The preprocessing included scaling the size of images to be 32×27 in the case of Yale and Weizmann. The FERET images were first scaled to 150×130 and then the first 256 PCA coefficients were taken. For each class, (person identity) half of the images were randomly selected for training and the rest of the images were used for recognition. Each experiment was repeated 10 times.

The goal of our experiments is to asses the relative performance of NCA as a (supervised) method in a face-recognition task. The face-recognition methods we compared are Eigenfaces (PCA) [8], Gaussian RBF Kernel-PCA [11], and Fisherfaces (LDA) [9]. All these subspace projection methods are followed by a whitening step in the transformed space, which is equivalent to utilizing the Mahalanobis distance in the transformed space. Another recently suggested distance metric to be used in the transformed space is the relevant component analysis (RCA) method [13] where only the within-class variability is used for whitening. It was shown in [13] that utilizing the RCA distance metric can enhance the performance of LDA. We also show recognition results for the LDA followed by RCA. We have implemented two variants of the NCA. The first (denoted by NCA1) is based on the cost function $\sum_i \log(p_i)$ and the second variant (denoted by NCA2) is based on the cost function $\sum_i p_i$ where p_i is the probability of correct classification.

The recognition task in the following experiments is to classify face images with respect to the identity of the person. We consider the retrieval paradigm reminiscent of nearest-neighbour classifier in which a query image leads to the retrieval of its nearest neighbour in the training data set. The distance measure we used in the transformed space (after whitening) is the Euclidean distance. Note that when using the NCA to obtain a subspace, there is no need for a whitening process as the distance learning is combined with the linear-subspace searching. An example of a face-recognition retrieval query from the Yale database is presented in Figure 2 where 5 nearest-neighbour retrieval results, based on LDA and NCA1, are shown.

The recognition results are presented in Figure 3. It can be verified that both variants of the NCA algorithm significantly outperform previously suggested subspace methods across all the databases that were used. The competitive advantage of the NCA method is even more significant in the case of projection into very low-dimensional space (e.g., when $d = 5$ or $d = 10$). Aside from improved performance, this fact can yield a better recognition-system in terms of computational complexity and memory size. Following the results of Bar-Hillel et al. [13] we have found that in some cases using the RCA distance metric can improve the performance of recognition systems based on LDA. The RCA is useful in cases where there are many face-image examples from each subject and we can obtain a good estimation of the within-class variability. In such cases (e.g., Yale and Weizmann) using RCA and NCA, we obtained similar classification results when the dimensionality of the reduced representation was relatively high. In very low dimensions the NCA was found to be significantly better. In the case of the FERET database the RCA has no advantage over LDA with Mahalanobis distance (FERET-10) and it can even be worse (FERET-4).

To further exemplify the significant improved performance gained from the NCA in very low dimensions, we show an example of linear projection into the 2-dimensional plane. The database used comprised the first five subjects of the Weizmann face database. For each subject there are 66

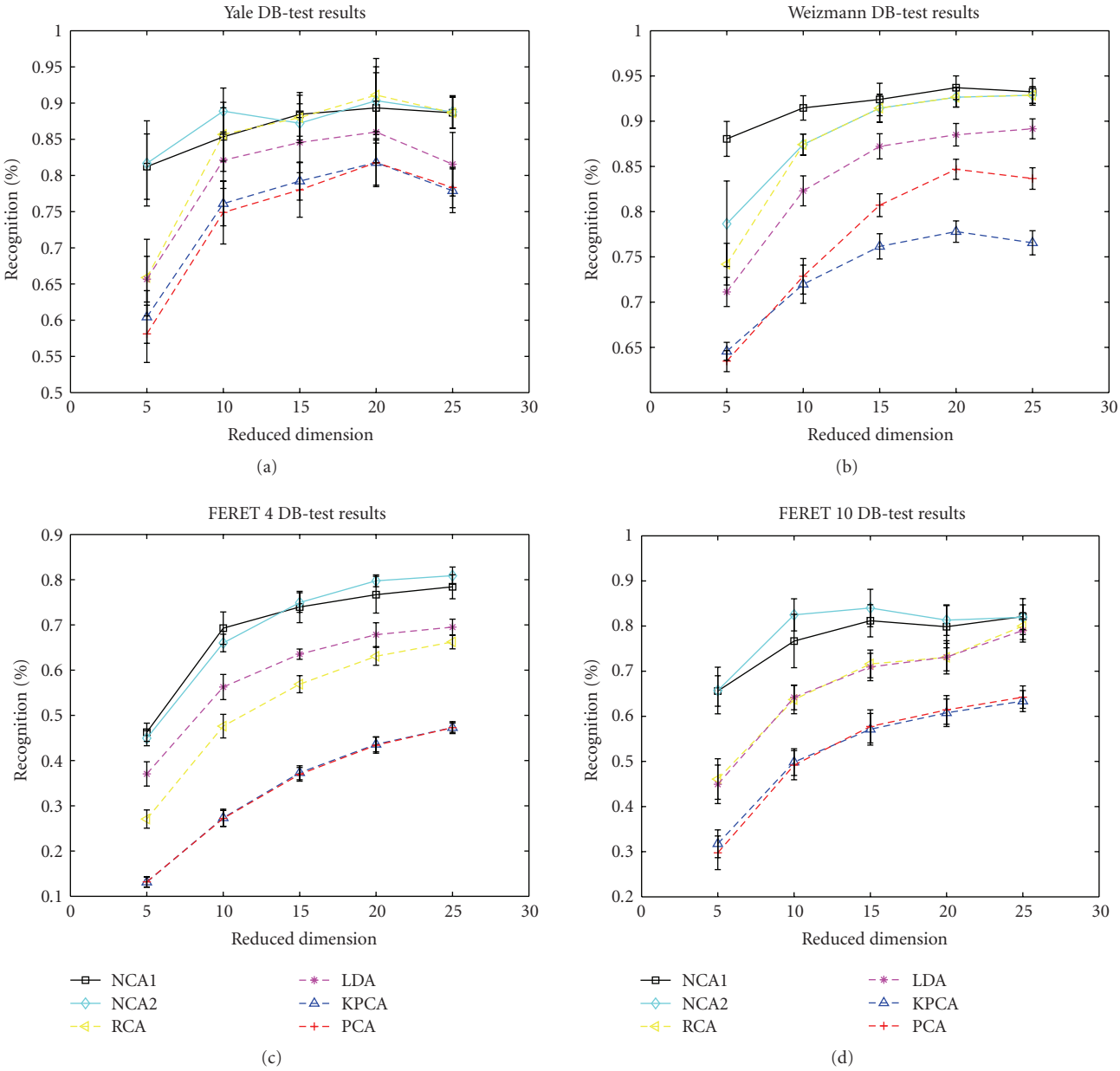
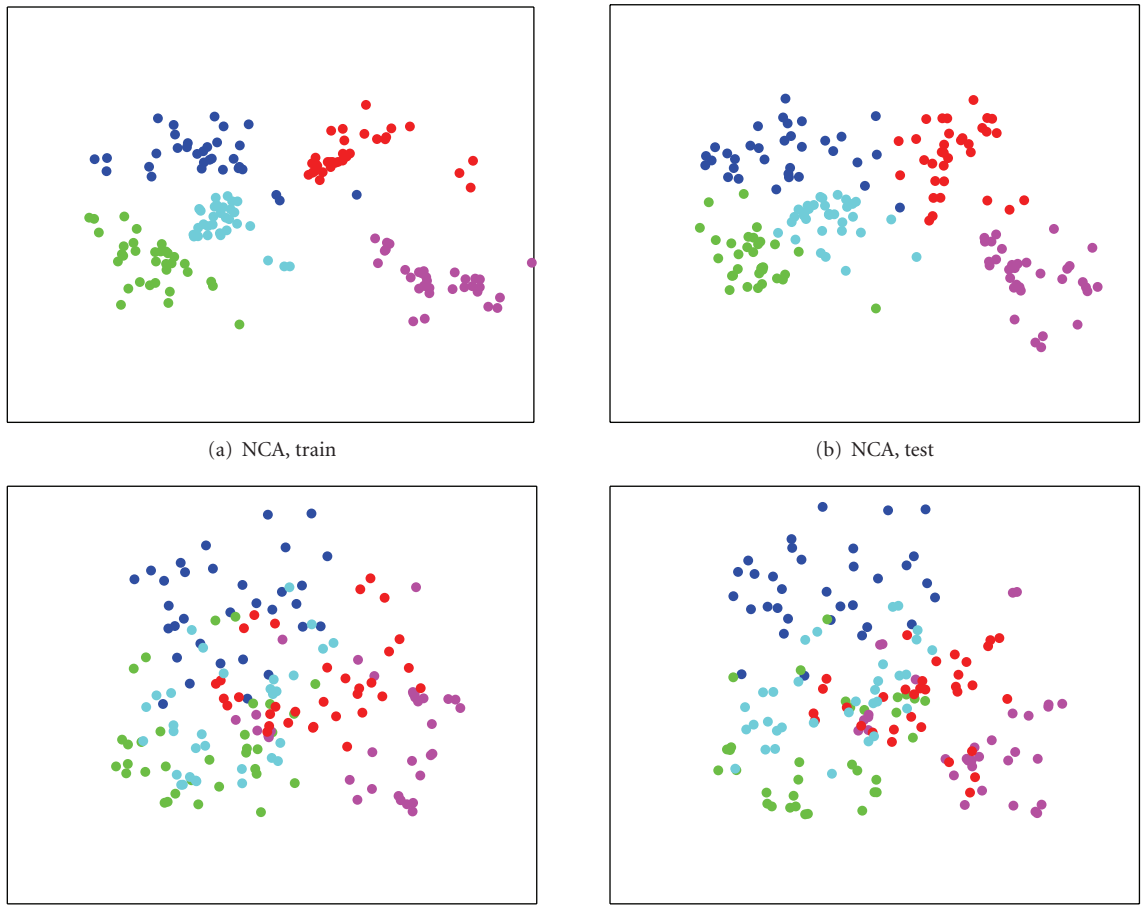


FIGURE 3: Face-recognition performance of several subspace methods, as a function of the representation dimensionality, on standard face databases. Standard errors of the means are shown on curves. The databases are (a) Yale, (b) Weizmann. (c) A subset of FERET consisting of persons with more than 4 images. (d) A subset of FERET consisting of persons with more than 10 images.

face images, half of which are used to find the subspace and the other half is used for testing. Figure 4 shows the low-dimensional representation obtained from LDA and NCA. The LDA transformation was also used as a starting point for the iterative conjugate-gradient algorithm that was applied to find the optimal NCA transformation. The nearest-neighbor recognition results (percentage of correct classification) for the database presented in Figure 4 are PCA-47, LDA-67, and NCA-93.

The NCA was found to be better than all the other methods discussed in this paper in terms of performance. In real world applications the training is done once and the test phase running time is important. The computational complexity of a single-face recognition is the same for all the methods. They are all based on a nearest-neighbor classifier in the projected space. The only difference between the methods is the linear transformation selected at the training phase, which has no implications on the complexity. The NCA training time is larger since the optimization is done iteratively. The training running time based on 200 training images 27×32, 40 classes, and reduced dimensionality 5 (using Pentium(R) 4 CPU 3.2 GHz, 1 GB of RAM) was 0.5 seconds for PCA, 5 seconds for LDA and RCA, and 68 seconds for NCA.



(a) NCA, train (b) NCA, test

(c) LDA, train (d) LDA, test

Figure 4: A two-dimensional linear representation of the first 5 subjects from the Weizmann face database. Images of the same person have the same color. The top and bottom rows show the results for NCA and LDA, respectively.

4. CONCLUSION

We have presented a linear subspace algorithm implicitly combined with a distance-learning method in the transform subspace for face-recognition tasks. We have shown that this method performs well across a range of standard face databases and a range of projection dimensions. It consistently outperforms existing subspace methods for face recognition particulary in the case of very low dimensions. There is a trend in recent years that linear subspace methods may be too limited for difficult classification tasks. A popular nonlinear alternative is based on kernelizing linear methods (e.g., kernel PCA and kernel LDA). The face manifold is definitely nonlinear. However, we have shown in this study that linear subspace can be a good approximation of this nonlinear manifold. We have shown that the space of linear transformations is still large enough to contain good classifiers. When using an appropriate target function, linear subspace methods can yield excellent face recognition results. It should be noted that the proposed method can be easily "kernelized." Instead of defining a projection $\mathbf{A}x_i$ in R^D, we can firstly project the subject in a Hilbert space F using a function ϕ and then using the projection $\mathbf{A}\phi(x_i)$.

This paper is focused on batch learning of the projection transformation. There are also online algorithms for learning a linear projection (e.g., [22]). A future research direction is developing an online version of the NCA algorithm that can be incrementally updated. As a final remark we note that the focus used in this paper is in evaluating a face-recognition approach by means of the performance achieved on a collection of datasets, dividing them into training and test sets. Under this focus, the results are provided as a general rate without taking into consideration if all the identities are similarly recognized or there are variations among them, this can be not good enough in real situations, that is, unrestricted imagery or video [23, 24].

REFERENCES

[1] M. Jones and P. Viola, "Face recognition using boosted local features," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, Nice, France, October 2003.

[2] K. Okada, J. Steffans, T. Maurer, et al., "The Bochum/USC face recognition system and how it fared in the FERET phase III test," in *Face Recognition: From Theory to Applications*, pp. 186–205, Springer, Berlin, Germany, 1998.

[3] J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3D morphable models," in *Proceedings of the 4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA '03)*, pp. 27–34, Guildford, UK, June 2003.

[4] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '94)*, pp. 84–91, Seattle, Wash, USA, June 1994.

[5] A. Mian, M. Bennamoun, and R. Owens, "2D and 3D multimodal hybrid face recognition," in *Proceeding of the 9th European Conference on Computer Vision (ECCV '06)*, pp. 344–355, Graz, Austria, May 2006.

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

[7] G. Shakhnarovich and B. Moghaddam, "Face recognition in subspaces," in *Handbook of Face Recognition*, Springer, Berlin, Germany, 2004.

[8] M. A. Turk and A. P. Pentland, "Face recognition using Eigenfaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '91)*, pp. 586–591, Maui, Hawaii, USA, June 1991.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[10] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 115–137, 2003.

[11] M.-H. Yang, "Kernel Eigenfaces vs. kernel Fisherfaces: face recognition using kernel methods," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '02)*, pp. 215–220, Washington, DC, USA, May 2002.

[12] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*, MIT Press, Cambridge, Mass, USA, 2004.

[13] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, pp. 937–965, 2005.

[14] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 303–321, 2001.

[15] E. Xing, A. Y. Ng, M. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, Mass, USA, 2002.

[16] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Advances in Neural Information Processing Systems 18*, MIT Press, Cambridge, Mass, USA, 2005.

[17] P. W. Keller, S. Mannor, and D. Precup, "Automatic basis function construction for approximate dynamic programming and reinforcement learning," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 449–456, Pittsburgh, Pa, USA, June 2006.

[18] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[19] Yale University Face Database, http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[20] The Weizmann Facebase, http://www.faculty.idc.ac.il/moses/.

[21] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[22] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, p. 94, Banff, Alberta, Canada, July 2004.

[23] A. Adler and J. Maclea, "Performance comparison of human and automatic face recognition," in *Proceedings of Biometric Consortium Conference (BC '04)*, Arlington, Va, USA, September 2004.

[24] A. M. Burton, R. Jenkins, P. J. B. Hancock, and D. White, "Robust representations for face recognition: the power of averages," *Cognitive Psychology*, vol. 51, no. 3, pp. 256–284, 2005.