Face Recognition with Weighted Locally Linear Embedding

Nathan Mekuz, Christian Bauckhage, and John K. Tsotsos
Department of Computer Science, Center For Vision Research
York University
Toronto, Ontario, Canada
{mekuz,bauckhage,tsotsos}@cs.yorku.ca

Abstract

We present an approach to recognizing faces with varying appearances which also considers the relative probability of occurrence for each appearance. We propose and demonstrate extending dimensionality reduction using locally linear embedding (LLE), to model the local shape of the manifold using neighboring nodes of the graph, where the probability associated with each node is also considered. The approach has been implemented in software and evaluated on the Yale database of face images [1]. Recognition rates are compared with non-weighted LLE and principal component analysis (PCA), and in our setting, weighted LLE achieves superior performance.

Keywords: face recognition, nonlinear dimensionality reduction, locally linear embedding

1. Introduction

Face detection and recognition are an extremely active area in computer vision, resulting in a large number of publications, as can be evidenced by the meta reviews [5], [16], and [17].

Face recognition can be characterized as a high-level pattern recognition problem in which humans are very skilled, whereas for machines, it presents a considerable challenge. In addition to authentication and recognition, recent efforts attempt to utilize face recognition to improve user interfaces.

Given a set of training images, a face recognition system needs to independently train to recognize a person from a new image. To be useful, such a system needs to capture image detail well enough to enable reliable recognition, with minimal sensitivity to variations in the test image, such as pose of the face and lighting. Also, since data tends to be large, scalability is of great importance.

1.1. Related Research

Due to the overwhelming number of approaches that have been proposed in the literature, an extensive survey of face recognition techniques is far beyond the scope of this paper. Rather, we will restrict our discussion to the popular and important class of appearance-based approaches.

Kirby and Sirovich [7] first proposed a low-dimensional reconstruction technique for face images which is optimal in the sense of preserving the direction of maximum variance based on a method known as the Karhunen-Loeve expansion or principal component analysis (PCA, see [6]). This technique was later extended for face recognition by Turk and Pentland [14]. The key idea is a pixel-wise comparison of the input image with the images in the database. To reduce the storage requirements of the database, and to facilitate faster comparisons, linear dimensionality reduction is performed, using PCA. This approach has had a tremendous impact on object recognition and remains popular to this day. While dimensionality reduction techniques such as Independent Component Analysis [3] and Linear Discriminant Analysis [1] may outperform PCA in some settings, PCA's applicability in the general case is widely recognized [3, 8] and it maintains its popularity as a face recognition tool, thanks to its performance, efficiency and simplicity.

However, PCA also suffers from some inherent disadvantages. For example, being a least-squares estimation technique, PCA is sensitive to outliers. Murase and Nayar [9] use a clever extension of PCA to represent different appearances of an object as a continuous manifold in the low dimensional space. Their technique first projects the image onto a low dimensional space, to identify the object. Once the object is recognized, it is projected onto a new coordinate system, defined specifically for that object. Murase and Nayar's system produces very impressive results, but is still sensitive to slight changes in the object's shape and lighting conditions.

Recently, several studies have demonstrated that the space of face images is likely nonlinear (cf. [4],[10],[11] [13]). If so, the application of PCA to face recognition, is sub-optimal. Significant effort has been devoted to the pursuit of appropriate dimensionality reduction techniques that would exploit this idea. Some of the techniques that were developed to consider the nonlinear structure of the manifold, particularly locally linear embedding

(LLE, see [10]), seem to be better discriminators of face images for classification purposes. This could suggest that in the problem of face recognition, local manifold structure carries more discriminating power than the global Euclidean structure. One pitfall of these techniques is that the local structure at a data point is learned from its neighbors. In other words, these techniques require a large training set.

The above approaches have produced promising results, albeit at the cost of higher computational complexity than PCA. Also, the underlying premise has been that the images in the training set equally represent the test image domain, an assumption that is in many cases flawed.

Several schemes have been proposed to extend PCA to address the issues of missing data and robustness against outliers (e.g., [2]). Skočaj et al. [12] propose a technique to extend PCA to include weights for images in the training set as well as individual pixels in the image. These weights are used both in the recognition of the test face, and the computation of the reduced space. This is a very powerful idea, and is a key element of our approach.

1.2. Contributions

The apparent effectiveness of nonlinear, non-global dimensionality reduction and the potential in using weights (which we will motivate in section 2) prompted us to investigate a generalization of such a nonlinear technique to process images with variable probabilities.

The main contributions of the present research are as follows: First, certain appearance variations are modelled as variance around the standard image, with decaying probability of occurrence as the test image diverges from the standard image. Second, we present a unified framework that allows computing a low-dimensional mapping known as locally linear embedding, that channels learning to recognize images with variable probabilities.

1.3. Outline of Paper

This paper proposes a technique for face recognition given variable probabilities based on locally linear embedding (LLE). It is composed of four main sections. This first section has provided motivation for considering the probabilities of face appearances and for using dimensionality reduction based on local features. In section 2 we describe the algorithmic and technical aspects of our approach in general terms. Section 3 presents empirical results of a realization of our algorithm. Section 4 provides a summary and finally, section 5 points out directions for further research.

2. Technical Approach

Our approach is based on the following ideas:

- Variation in face appearance can produce a range of face images, with varying probabilities of occurrence.
- As shown by [10] and [13], face space may not be linear. Both contributions demonstrate recognition rates superior to those generated by PCA. Therefore, the pursuit of a locally linear dimensionality reduction scheme is justified.
- Nonlinear dimensionality reduction techniques such as LLE may be enhanced by introducing image weights that represent a given image's probability of occurrence.

These ideas are discussed in greater detail below.

2.1. Weighted Images

Various generalizations of PCA, including weighted PCA, have been known to statisticians for decades. Skočaj et al. [12] have applied weighted PCA to image recognition, using two types of weights: weights for individual pixels (spatial weights), used to account for parts of the image which are unreliable or unimportant, and weights for images, which they refer to as temporal weights. The main motivation for the latter is the idea that more recent images of a subject (or object) are more reliable. The idea is to maximize the weighted variance in the low dimensional space in order to achieve lower reconstruction error for certain target images or pixels.

Other possible applications for image weights are tuning an authentication system to achieve lower error rates for individuals with high levels of classification.

Our research was motivated by a different hypothesis. Normally, an effort is made to standardize the input images to minimize variations. For example, variation in location can be moderated (alignment) by projecting a sliding window over the image onto the reduced space and looking for minimum distance to the projection hyperplane. Other sources of variance are much more difficult to mitigate. This category includes variation due to scale, orientation (i.e. rotation with respect to the camera's optical axis) and pose (relative to the camera), facial expression, occlusion, and lighting conditions, and the presence or absence of features such as beards, glasses, or clothing articles like scarfs and hats. To this end, face recognition systems are typically trained with multiple images for each individual. Most systems treat all training images equally, implicitly assuming that their probability of occurrence is uniformly distributed. Clearly, this assumption is faulty. If we consider the case of orientation (in-plane rotation), most faces appear more or less upright. In the absence of specific knowledge about the probability distribution of face images with respect to rotation, it is convenient to assume a Gaussian distribution, which is far from uniform.

2.2. Locally Linear Dimensionality Reduction

The quest for nonlinear techniques for dimensionality reduction ([10], [13], [11], [4]), has been driven by several factors. Roweis [10] illustrates some pitfalls of PCA using a few artificial 3-D examples, such as a Swiss Roll. PCA fails to generate a feasible mapping due to its reliance on Euclidean distances instead of geodesic distances. Others have used examples where PCA mistakenly maximizes variance caused by outliers. Beyond these contrived examples, the above works have shown that the face domain indeed lies on a nonlinear manifold.

One method that appears to generate particularly good results for face recognition is locally linear embedding, or LLE [15]. In essence, LLE computes a low dimensional embedding where adjacency of points in the original space is maintained in the low dimensional space. In other words, the local arrangement of the points is preserved. In addition to learning the local structure of the manifold, this technique promotes robustness, since an outlier only affects its neighbors. On the other hand, LLE is heavily dependent on sufficient sampling of the manifold to learn its shape.

If the manifold is indeed well-sampled, then a point and its neighbors lie on an almost linear hyperplane which describes a patch of the manifold. For each point, LLE finds coefficients for its neighbors that best describe it using a linear combination that generates the lowest reconstruction error. After these sets of coefficients are computed, LLE finds a mapping to a low-dimensional space where each point can be approximated with these coefficients while minimizing reconstruction error.

Various flavors of LLE define neighbors differently. The simplest form identifies k nearest Euclidean neighbors for each point. A more complex formulation looks at neighbors that fall within a ball of radius ε. Roweis uses the former method, which is simpler and computationally less expensive. We have also used the k nearest neighbor model, extending it to variable image weights. We leave a similar extension to the fixed radius version for future work.

2.3. Image Weights in LLE

Skočaj et al. ([12]) introduce two methods for image and pixel weights in PCA. An EM algorithm which allows online processing of images, and a batch algorithm where all training data is available upfront.

The batch method considers weights on two occasions: in the computation of eigenfaces, and in the calculation of distance of the projected test image in the recognition step. To compute the weighted eigenfaces, an input image x_i, after normalization by subtracting the average face is multiplied by the image's weight w_i^i as follows:

$$\hat{x}_i = \sqrt{w_i^i}(x_i - \Psi), \quad i = 1, \ldots, N \tag{1}$$

where $\hat{x}_i$ is the adjusted image from which the eigenvectors of the covariance matrix are computed, with image weights factored in and Ψ is the average face. Similarly, spatial weights are introduced as follows:

$$\hat{x}_{ij} = \sqrt{w_j^s}(x_{ij} - \Psi_j), \quad i = 1, \ldots, N \ , \ j = 1, \ldots, M \tag{2}$$

where $\hat{x}_i$ is the image adjusted for spatial weights, and w_j^s is the weight for pixel j. So, ultimately the weighted input image, $\hat{x}_i$ from which eigenfaces are derived is computed as follows:

$$\hat{x}_{ij} = \sqrt{w_j^s w_i^i}(x_{ij} - \Psi j), \quad i = 1, \ldots, N \ , \ j = 1, \ldots, M \tag{3}$$

Our aim is to introduce image weights in LLE in a fashion that would achieve similar results. We limit our research at this time to the k nearest neighbors flavor of LLE. The LLE algorithm consists of three steps:

1. identify k nearest neighbors. This step is generally inexpensive. Our implementation employs a brute force algorithm with complexity $O(DN^2)$ (where N is the size of the data and D is the original dimensionality), but the nearest neighbors can be computed in $O(NlogN)$ time using K-D trees or ball trees.
2. compute w_{ij}, the weights that best reconstruct data point x_i from its neighbors. The computational complexity of this step is $O(DNk^3)$.
3. compute low dimensional vectors y_i that best reconstruct the weights computed in step (2). This step runs in $O(dN^2)$ time, where d is the reduced dimensionality.

In our extension, images have associated weights that represent their probability of occurrence, or reliability. Since the local shape of the manifold for a given point is learned from its neighbors, we need to factor in the neighbors' weights in order to reduce the effect of Euclidean neighbors with low weights. To this end, we need to examine the first two steps of the algorithm closer. Once the nearest neighbors for a point x_n have been identified, the local Gram matrix is computed.

$$C_{i,j} = (x_n - x_i)^T (x_n - x_j) \tag{4}$$

where $i, j \in \{neighbor_1, \ldots, neighbor_k\}$. Note that the Gram matrix is symmetric and semipositive definite, and defines the difference vectors $x_i - x_j$ up to isometry. Optimal weights that minimize reconstruction error can be easily computed using Lagrange multipliers, or equivalently by solving

$$Cw_i = \mathbf{1} \tag{5}$$

for w_i where $\mathbf{1} = [1, 1, ..., 1]^T$, and normalizing w_i so that $\sum_{j=1}^{k} w_{ij} = 1$.

Therefore, the Euclidean distance of a point to its neighbors plays a role both in the selection of neighbors, and in the formation of the Gram matrix. Clearly, we need to adjust the distance to incorporate the neighbor's probability of occurrence or weight. Let p_i be the probability of occurrence for point x_i. Then we define the adjusted distance d_{ij} of point i from its neighbor j to be:

$$\hat{d}_{ij}^2 = p_i \cdot d_{ij}^2 \tag{6}$$

This is of consequence both when determining a point's neighbors in step (1) and in the computation of the Gram matrix in step (2). Identifying point i's k nearest neighbors now becomes a search for the k points with the smallest value for $\hat{d}_{ij}^2$, and the adjusted Gram matrix for point i is computed as follows:

$$\hat{C}_{i,j} = \sqrt{p_i p_j}(x_n - x_i)^T(x_n - x_j) \tag{7}$$

where $i, j \in \{neighbor_1, \ldots, neighbor_k\}$. Now the solution to $\hat{C}w_i = \mathbf{1}$ yields the best weights adjusted for probability of occurrence, and the embedding in step (3) can be computed as before from these adjusted weights.

One issue to be considered is picking an appropriate value for k that allows a fair comparison between the standard training set and one enlarged by introducing additional variations.

Ultimately, k defines the size of the neighborhood used to learn the local shape of the manifold. Adding more training data (e.g. by introducing variations generated by rotation) increases the sample density which aids learning the manifold (the effectiveness of LLE greatly depends on the size of training set). However, Euclidean neighbors with low weights play a relatively small role in the construction of the embedding so when choosing k, weights need to be considered as well. Another factor that needs to be examined is the similarity of the degree of variation introduced by new data. Consider the extreme case where the training set is simply doubled by duplication, with equal weights. Clearly, in this case, it is appropriate to use $2k$ neighbors excluding each point's twin.

We have presented some basic intuition for choosing an appropriate value for k. We will leave a more rigorous and complete derivation for future research.

3. Empirical Assessment

We have designed a series of performance measurements to test the effectiveness of the proposed approach. We used a standard database of face images and augmented it with variations of the faces generated using in-plane rotation at various small angles. A Gaussian distribution of probabilities was assigned based on rotation angle. The rationale for selecting rotation as a mutator is that rotation cannot be easily corrected by maximizing the projection of a sliding window, as is commonly done to correct location. Also, faces generally appear upright, and the subject's position and camera geometry only allow small variations.

3.1. Testing Methodology

For our testing, we used the Yale face database [1], which consists of 165 images of 15 subjects recorded under different lighting conditions and showing a variety of facial expressions.

In keeping with the testing methodology applied by Roweis and Saul [10], we cropped and aligned the images, to a final size of 80×80. To further reduce computational cost, following an idea proposed by Niyogi et al. [4], we performed a PCA preprocessing step on the images reducing them to their 100 largest principal components, effectively keeping over 99% of their information.

We tested the recognition rates achieved by PCA and LLE on the standard Yale database [1] of 165 upright face images. For LLE, we used $k = 3$, which empirical evidence proved to be a good choice. Next, for each image in the database, we created five variations by performing in-plane rotation by $\varphi_l \in \{-8°, -4°, 0°, 4°, 8°\}$, effectively increasing the size of the database to 825 images. We tested this extended data set with LLE using $k = 7$ (following the intuition provided in section 2.3), and finally with weighted LLE, where the image weights $w(\varphi_l)$ were set according to

$$w(\varphi_l) = e^{-\frac{(l-2)^2}{2\sigma^2}} \tag{8}$$

where $l \in \{0, \ldots, 4\}$ indicates the rank of φ_l in the ordered set of angles and $\sigma = 0.776$.

Recognition rates were measured for each of the above techniques using the leave one out strategy (m fold cross validation): each image was removed from the database to create a training set of size $m - 1$, and then classified.

3.2. Results

We compared the recognition error rates achieved by the above methods, at various dimensionality settings. As figure 2 demonstrates, our weighted LLE algorithm realizes recognition by clustering images of the same individual together in the low dimensional space. In the plot, markers represent face images projected onto the first two dimensions, with a different marker used for each individual. Clusters of different appearances of the same individual are apparent, providing some insight into weighted LLE's recognition power.

The results of our testing are summarized in figure 1, which plots the error rate in face recognition as a function of the dimension d of the low dimensional embedding space.

Several observations are apparent from this chart. First, in our testing LLE consistently outperformed PCA. Although in Roweis' tests [10], PCA actually achieves better recognition rates than LLE at higher dimensions, our results are

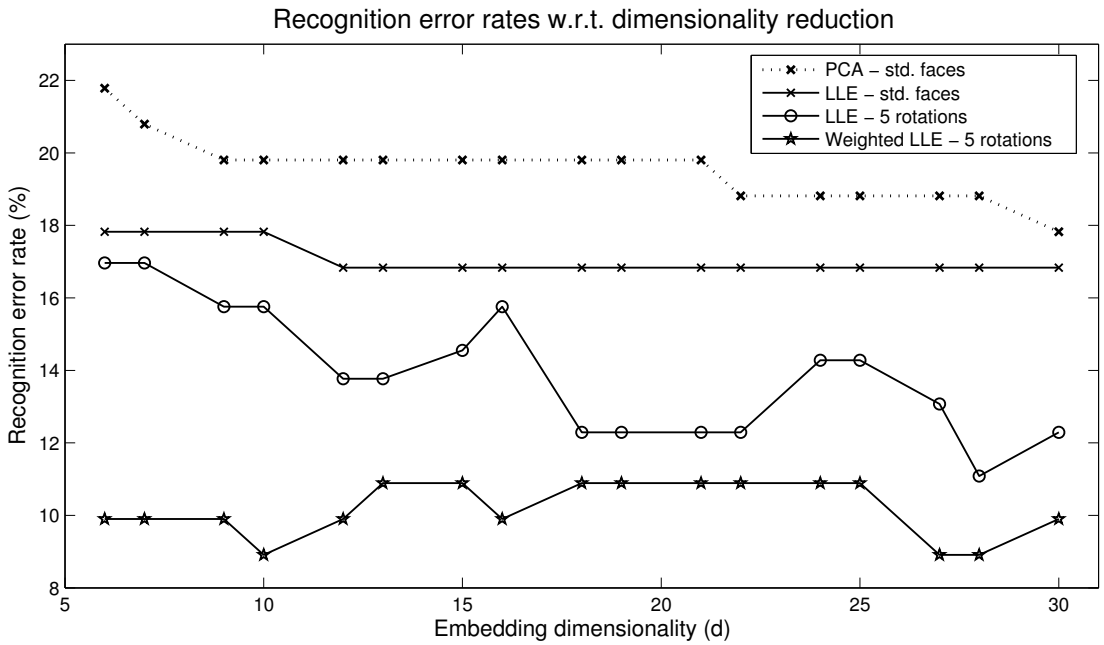

Figure 1. Recognition errors obtained from experiments with different dimension reduction techniques for face recognition using the Yalefaces database.

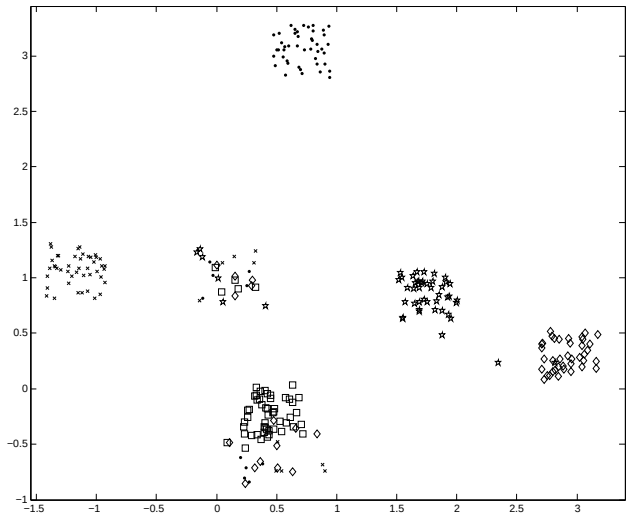

Figure 2. The projection of images by weighted LLE onto the first two dimensions, using $k = 7, d = 15$. Each marker type represents a different individual. Note the formation of clusters by the algorithm.

not entirely at odds, as we clarify below. Next, not surprisingly, LLE's performance improved when a larger training set was introduced. And finally, we note that weighted LLE achieved superior weighted recognition rates compared to the non-weighted version.

As an aside, the computational cost of LLE (both weighted and non-weighted) in our experiments greatly depended on the size of the training set. The target dimensionality d and neighborhood size k affected execution time only moderately. Although a difficulty with all appearance based vision techniques, is that computational complexity heavily depends on the size of the training data, allowing for variable probability of occurrence in training images is still befitting and feasible.

3.3. Discussion

The results are encouraging in that they confirm our hypothesis, that in a setting where images in the training set have varying probabilities of resembling a test image, weighted LLE achieves superior recognition results.

It is interesting to note that for every embedding dimensionality considered in our experiments, weighted LLE produced the lowest recognition error, sometimes by a large margin. Weighted LLE reached especially low recognition errors even for embedding spaces of small dimensionality. Also, in our tests, LLE performed consistently better than

PCA. While in Roweis' experiments [10], PCA actually beat LLE at higher dimension settings, we observed an improvement in PCA's recognition rates as the dimensionality of the embedding space grows but saw no such improvement in LLE. In Roweis' experiments, this crossover occurs at around $d = 18$. We have tested d values of up to 30, and have not reached the crossover point.

Another observation is that with the larger training set, non-weighted LLE appears to improve as the dimensionality increases, whereas weighted LLE's recognition rates are practically constant. This may suggest the existence of some upper-bound on the recognition rate due to poor modelling of the face manifold in some areas. The bumpiness of the sets that include rotation could be due to closer competition for neighbors. This may possibly be alleviated with better parameter tuning.

4. Summary

This paper presented a novel approach to face recognition. Given the observation that face recognition systems generally improve in reliability when presented with multiple training images for each subject and noting that training images are treated equally by most algorithms, we proposed extending the locally linear embedding procedure with a weighting scheme.

In our extension, weights are associated with images to represent their probability of occurrence. As the LLE algorithm recovers the local structure of the face manifold from the neighbors of a given face, we have shown how the impact of neighboring faces with low weights may be reduced. The effectiveness of this approach has been verified by experiments with the Yale face database [1]. We have demonstrated that compared with PCA and standard LLE, our weighted LLE algorithm performs better: for every embedding dimensionality considered in the experiments, weighted LLE produced the lowest weighted recognition error rates, with error rates of about 5% lower than non-weighted LLE.

5. Future Work

It should be noted that in order to realize the above results, we chose values for k empirically. Currently, no definitive method exists to choose optimal values for k in LLE, as well as several other nonlinear dimensionality reduction techniques. Future work may study the response of weighted (and non-weighted) LLE to different values of k.

Also, as mentioned earlier, the present research only covers the k nearest neighbors paradigm. In the future, it may be useful to introduce weights to the more complex (due to variable neighborhood size) fixed radius model.

Another problem left for future research is the rigorous derivation of an appropriate value for k in the variable image probability model. Additionally, we feel that better understanding is needed of the factors that contribute to the assignment of modes of variation to principal axes in the reduced space.

Finally, the idea of weighting can also be extended, to assign different weights to individual pixels in the image, where information is deemed to be more significant, and to other nonlinear reduction algorithms, such as Isomap [13].

Acknowledgments:

The authors thank Konstantinos Derpanis for his helpful comments in reviewing this paper.

References

[1] P. N. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Machine Intelli.*, 19(7):711–720, 1997.

[2] F. De la Torre and M. Black. Robust principal component analysis for computer vision. In *Proc. ICCV*, pages 362–369, 2001.

[3] B. Draper, K. Baek, M. Bartlett, and J. Beveridge. Recognizing faces with pca and ica. *Comput. Vis. Image Underst.*, 92(2):115–137, 2003.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Trans. Pattern Anal. Machine Intelli.*, 27(3):328–340, 2005.

[5] E. Hjelmås and B. Low. Face Detection: A Survey. *Comput. Vis. Image Underst.*, 83(3):236–274, 2001.

[6] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[7] M. Kirby and L. Sirovich. Low-dimensional procedure for the characterization of human faces. *Optical Society of America*, 4(3):519–524, 1987.

[8] A. Martínez and A. Kak. PCA versus LDA. *IEEE Trans. Pattern Anal. Machine Intelli.*, 23(2):228–233, 2001.

[9] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. J. Comput. Vision*, 14(1):5–24, 1995.

[10] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[11] A. Shashua, A. Levin, and S. Avidan. Manifold pursuit: A new approach to appearance based recognition. In *ICPR*, volume 3, pages 590–594, 2002.

[12] D. Skočaj, H. Bischof, and A. Leonardis. A robust pca algorithm for building representations from panoramic images. In *Proc. ECCV*, volume IV, pages 761–775. Springer-Verlag, 2002.

[13] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuro Science*, 3(1):71–86, 1991.

[15] M.-H. Yang. Face Recognition using Extended Isomap. In *Proc. ICIP*, pages 117–120, 2002.

[16] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Trans. Pattern Anal. Machine Intelli.*, 24(1):34–58, 2002.

[17] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.