A Unified Framework of Subspace and Distance Metric Learning for Face Recognition

Qingshan Liu [1,2], Dimitris N. Metaxas [1]

1. The department of computer Sciences, Rutgers University
2. National Laboratory of Pattern Recognition, CAS, China

Abstract. In this paper, we propose a unified scheme of subspace and distance metric learning under the Bayesian framework for face recognition. According to the local distribution of data, we divide the k-nearest neighbors of each sample into the intra-person set and the inter-person set, and we aim to learn a distance metric in the embedding subspace, which can make the distances between the sample and its intra-person set smaller than the distances between it and its inter-person set. To reach this goal, we define two variables, that is, the intra-person distance and the inter-person distance, which are from two different probabilistic distributions, and we model the goal with minimizing the overlap between two distributions. Inspired by the Bayesian classification error estimation, we formulate it by minimizing the Bhattachyrra coefficient between two distributions. The power of the proposed approach are demonstrated by a series of experiments on the CMU-PIE face database and the extended YALE face database.

1 Introduction

Face recognition is a hot topic in the communities of computer vision and pattern recognition due to its potential applications in biometrics, surveillance, human-computer interface, and multimedia. A lot of methods have been proposed in the past decades [31].

Since Principal Component Analysis(PCA) achieved much success in EigenFace [25], subspace learning methods have been widely used for facial feature representation. The general goal of subspace learning is to find some transformation to project high-dimensional data into a low-dimensional subspace. Defining different objective functions will produce different subspaces. We will review some popular subspace methods in Section 2. However, same as most pattern recognition problems, similarity measurement or classification scheme is needed to further analyze the relationship of the data or to predict their labels based on the extracted features for face recognition. The simple Euclidean distance is often used to measure the similarities between two face images in the subspace, but it is not a better metric in most cases. Distance metric learning is a technique to learn a distance based similarity measurement and classification scheme, and has attracted much attention in machine learning and computer vision in recent years. Its original goal is to directly learn the distance metric from the available training data, in order to improve the performance of distance-based classifiers. Due to the encouraging effectiveness of the simple nearest neighbor rule, most studies focused on learning the similarity matrix of the Mahalanobis distance to improve the performance

of the nearest neighbor classification. A common strategy is to minimize various separation criteria between the classes assuming equivalent relations over all the data or the k-nearest neighbors. A brief review will be given in Section 2. However, for high dimensional data, such as face image data (the dimension of an image with the size of 100×100 is up to 10^4), learning the metric matrix directly in such a high dimensional space, not only results in high computational cost, but also is sensitive to noise.

In this paper, we propose a unified scheme of subspace and distance metric learning for face recognition under the Bayesian framework. In order to learn a local distance metric with subspace dimensionality reduction, we divide the k-nearest neighbors of each sample into the intra-person set and the inter-person set according to the local distribution of the data, and we aim to make the distances between the sample and its intra-person set smaller than the distances between it and its inter-person set in the embedding subspace, so as to handle the high-dimensional data well. We define two variables in the subspace, i.e., the intra-person distance and the inter-person distance, and model them with two different probability distributions. Thus, the problem can be converted to minimize the overlap between two distributions. Inspired by the Bayesian classification error estimation, we formulate it by minimizing the Bhattachyrra coefficient measurement between two distributions, and the solution can be obtained by the gradient descent optimization. The proposed work has some special characteristics: 1) It is based on the local neighbors, so it does not make assumption on the global distribution of the data like Linear Discriminant Analysis (LDA). 2) It can be directly used for multi-class problems without any modification or extension. 3) It links to Bayesian classification error and has an intuitionistic geometric property due to adoption of the Bhattachyrra coefficient measurement. We conduct the experiments on two benchmarks, the CMU-PIE face database [21] and the extended YALE face database [15], and the experimental results show the promising performance of the proposed work compared to the state-of-the-arts.

2 Related Work

Subspace learning is a popular approach of face recognition. It maps the high dimensional face image data into a low dimensional subspace based on some criteria. Eigenface [25]and Fisherface [5] [30]are two classic methods, which are based PCA and LDA respectively. PCA seeks to maximize the covariance over the whole data, so it is optimal for data reconstruction, but it is not optimal for classification. The idea of LDA is to find a linear subspace projection that maximizes the between-class scatter and minimizes the within-class scatter. However, LDA assumes that each class has a similar within-class distribution of samples. Kernel PCA (KPCA) and Kernel LDA (KDA) combine the nonlinear kernel trick with PCA and LDA to get nonlinear principal component and discriminant subspaces [19] [16]. However, for the kernel methods, the kernel function design is still an open problem, and different kernels will give different performances. Manifold based subspace methods, such as LLE [17] and ISOMAP [23], aim to preserve the local geometric relations of the data in both the original high dimensional space and the transformed low dimensional space, while they often have a problem of "out of sample". Local Preserving Projection (LPP) gives a linear approx-

imation of manifold structure to deal with this problem [13]. In [8], the idea of LDA is integrated into LPP to enhance the discriminating performance of LPP. In [22], M. Sugiyama proposed to compute the within-class scatter and between-class scatter in LDA with a weighting scheme inspired by LPP. A generalized interpretation for these methods based on graph analysis is discussed in [28]. From the view of subspace dimensionality reduction, our work is similar to LDA, which aims to find a transformation of separating one class from the others, and it can be also extended with the kernel trick. However, our work is different from LDA in that: no constraints are made on the global distribution of the data, because it is based on the local neighbors' distribution, and it preserves the neighborhood relationship of the data during the dimension reduction as in manifold learning.

Subspace learning can be thought as a method of feature representation, while distance metric learning is related to constructing a data classification scheme. It is well known that the nearest neighbor rule is simple and surprisingly effective. However, its performance crucially depends on the distance metric. For different distance metrics, it will produce different nearest neighbor relationships. Most previous studies aim to improve the performance of the nearest neighbor classification by learning a distance metric based on the Mahalanobis distance from the labeled samples. E. Xing et al [27] tried to find an optimal Mahalanobis metric from contextual constrains in combination with a constrained K-means algorithm. B. Hilled et al [4] [20] proposed a much simpler approach called Relevance Component Analysis (RCA), which identities and downscales global unwanted variability within data. However, it does not consider the between class pair-wise information, which will influence its performance on classification [14]. K. Q. Weinberger et al [26] proposed to learn the distance metric by penalizing large distances between each input and its neighbors and by penalizing small distances between each input and all other inputs that do not share the same label. Its solution is based on complex quadratic programming. Torresani and Lee [24] extended this method with dimensional reduction, but its objective function is non-convex. Neighborhood Component Analysis (NCA) aimed at directly maximizing a stochastic variant of the leave one out *K-NN* score on the training set [12]. Later, A. Globerson et al [11] converted the formula of NCA to a convex optimization problem with a strong assumption that all the samples in the same class were mapped to a single point and infinitely far from points in different classes. Actually, this assumption is unreasonable for practical data. In [29], the bound optimization algorithm [18] was adopted to search a local distance metric for the non-convex function. Most of the above methods do not consider the dimensionality reduction for high dimensional data except for RCA [4] [20], NCA [12], and [24]. However, the proposed method is different from them in that it links to Bayesian classification error and has an intuitionistic geometric property due to adoption of the Bhattachyrra coefficient measurement.

3 Our Work

In this section, we propose a new unified framework of subspace and distance metric learning, which is inspired by the Bayesian classification error estimation. We first present our purpose and then give a Bhattacharyya coefficient based solution.

3.1 The Purpose

Let $X = \{x_1, x_2, \cdots, x_n\} \in R^D$ denote the training set of n labeled samples in C classes. Let $l(x_i)$ be the label of sample x_i, i.e.,$l(x_i) \in \{1, 2, \ldots, C\}$. Most distance metric learning methods seek to directly find a similarity matrix Q based on the Mahalanobis distance to maximize the performance of the nearest neighbor classification. The Mahalanobis distance between samples x_i and x_j is defined as follows:

$$P_{i,j} = (x_i - x_j)^T Q (x_i - x_j). \tag{1}$$

However, learning Q directly in a high dimensional space, such as the image space, will be sensitive to noise to some extent, besides being computationally expensive.

Since Q is a $D \times D$ semi-definite matrix, it can be rewritten as: $Q = AA^T$. If the dimension of A is $D \times d$, $d < D$, (1) is equivalent to calculating the Euclidean distance in the transformed subspace with A.

$$P_{i,j} = \|A^T x_i - A^T x_j\|^2 = (x_i - x_j)^T AA^T (x_i - x_j). \tag{2}$$

Thus, the distance metric Q for high dimensional data can be computed by an explicit embedding transformation A. In this paper, we will focus on how to first learn this transformation A, and then compute Q. Actually the transformation A is corresponding to subspace dimension reduction, so this idea is equivalent to integrating the subspace and distance metric learning together.

Before presenting the details of our approach, we first give some definitions. The set $N_r(x_i)$ is the k-nearest neighbors of sample . Same as in [26] [24], the neighbors are computed by the Euclidean distance in the original data space. We divide $N_r(x_i)$ into two sets using the labels of the samples, $N_r(x_i) = S_i \bigcup D_i$, where the labels of the set S_i are same as the label of x_i, $x_s \in S_i, l(x_s) = l(x_i)$, and the samples in D_i have different labels from the sample x_i, $x_d \in D_i, l(x_d) \neq l(x_i)$,. We call them the intra-person set and inter-person set respectively in this paper.

Intuitively, a good distance metric should make each sample close to the samples in the same class and far from the samples in the different classes. Based on the nearest neighbor classification scheme, we can compare each sample against its k-nearest neighbors. We aim to find a distance metric that makes each sample far from the samples in its inter-person set and close to the samples in its intra-person set. Thus, our goal can be described as follows:

Given any samples x_i and its two kinds of neighbors $x_s \in S_i$ and $x_d \in D_i$, the intra-person distance $P_{is}(A)$ between x_i and x_s should be smaller than the inter-person distance $P_{id}(A)$ between x_i and x_d:

$$P_{is}(A) = \|A^T (x_i - x_s)\|^2, x_s \in S_i, \tag{3}$$

$$P_{id}(A) = \|A^T (x_i - x_d)\|^2, x_d \in D_i, \tag{4}$$

$$P_{is}(A) < P_{id}(A), for \forall i, s, d. \tag{5}$$

3.2 Bhattacharyya Coefficient based Solution

For convenience, we define the variable $P_s(A)$ to represent all the intra-person distances, $P_s(A) = \{P_{is}(A)\}$ for all the i and s, and the variable $P_d(A)$ to represent all the inter-person distances, $P_d(A) = \{P_{id}(A)\}$ for all the i and d. Assuming that $P_s(A)$ and $P_d(A)$ are from two distributions respectively, $P_s(A) \sim \rho_s(P(A))$ and $P_d(A) \sim \rho_d(P(A))$, we can achieve our goal to find a transformation A that minimizes the overlap between these two distributions. Figure 1 gives an illustration, where x represents the distance $P(A)$. It can be found that minimizing the overlap means to separate the intra-person distances $P_s(A)$ from the inter-person distances $P_d(A)$ as much as possible, and it is also equivalent to minimizing the up-boundary of the classification error as much as possible under the Bayesian framework.

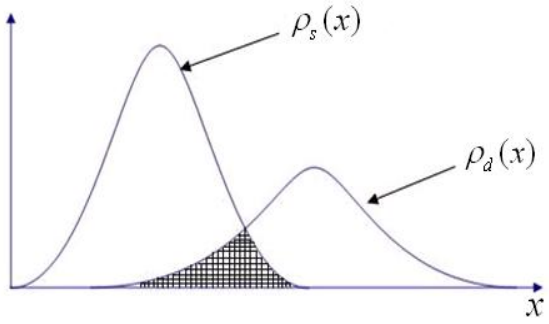

Fig. 1. An illustration of minimizing the overlap

The Bhattacharyya coefficient is a divergence-type measure which has an has an intuitionistic geometric interpretation [9]. Moreover, it is a popular technique to estimate the boundary of the classification error, i.e., the overlap between two distributions [10]. Given two distributions, $\rho_1(x)$ and $\rho_2(x)$, their Bhattacharyya coefficient is $\int \sqrt{\rho_1(x)\rho_2(x)}dx$. A small Bhattacharyya coefficient means a small overlap between two distributions which may lead to a small classification error. Thus, we define the objective function with Bhattacharyya coefficient between $\rho_s(P(A))$ and $\rho_d(P(A))$ as follows:

$$J_B(A) = \max_A(-\ln \int \sqrt{\rho_s(P(A))\rho_d(P(A))}dP(A)). \tag{6}$$

In this paper, we regard the variables of the intra-person distance and inter-person distance as two different Gaussian distributions. We define the mean and variance of all the $P_s(A)$ distances as $\mu_s(A)$ and $\Sigma_s(A)$, and the mean and covariance of all the $P_d(A)$ vectors as $\mu_d(A)$ and $\Sigma_d(A)$ i.e.,

$$\rho_s(P(A)) = N(\mu_s(A), \Sigma_s(A)), \tag{7}$$

$$\rho_d(P(A)) = N(\mu_d(A), \Sigma_d(A)), \tag{8}$$

where $N(\mu, \Sigma)$ represents a Gaussian distribution with mean μ and covariance Σ. Now the objection function (6) can be written as [10]:

$$J_B(A) = \max_A \left\{ \frac{1}{4} \frac{(\mu_s(A) - \mu_d(A))^2}{\Sigma_s(A) + \Sigma_d(A)} + \frac{1}{2} \ln \frac{\Sigma_s(A) + \Sigma_d(A)}{2\sqrt{\Sigma_s(A)\Sigma_d(A)}} \right\} \tag{9}$$

Denote $E(\cdot)$ represents the expectation operation, and $Tr(X)$ is the trace of the matrix X. Since any $\|A^T x_{ij}\|^2 = Tr(A^T x_{ij} x_{ij}^T A)$, where $x_{ij} = x_i - x_j$, we have

$$\mu_s(A) = E(P_s(A)) = E(Tr(A^T x_{is} x_{is}^T A)) = Tr(A^T E(x_{is} x_{is}^T) A) = Tr(A^T M_s A) \tag{10}$$

$$\mu_d(A) = E(P_d(A)) = E(Tr(A^T x_{id} x_{id}^T A)) = Tr(A^T E(x_{id} x_{id}^T) A) = Tr(A^T M_d A) \tag{11}$$

$$\Sigma_s(A) = E(P_s(A) - \mu_s(A))^2 = E(P_s(A))^2 - \mu_s^2(A) \tag{12}$$

$$\Sigma_d(A) = E(P_d(A) - \mu_d(A))^2 = E(P_d(A))^2 - \mu_d^2(A) \tag{13}$$

The solution of (9) can be obtained by the gradient descent algorithm, such as the conjugate gradient method. For simplicity, we ignore (A) in all the $J_B(A)$, $\mu_s(A)$, $\Sigma_s(A)$, $\mu_d(A)$, and $\Sigma_d(A)$. The differentiation of J_B with respect to A is as follows:

$$\frac{\partial J_B}{\partial A} = \frac{(\mu_s - \mu_d)(\frac{\partial \mu_s}{\partial A} - \frac{\partial \mu_d}{\partial A}) + (\frac{\partial \Sigma_s}{\partial A} + \frac{\partial \Sigma_d}{\partial A})}{2(\Sigma_s + \Sigma_d)} - \frac{(\mu_s - \mu_d)^2 (\frac{\partial \Sigma_s}{\partial A} + \frac{\partial \Sigma_d}{\partial A})}{4(\Sigma_s + \Sigma_d)^2} - \frac{\frac{\partial \Sigma_s}{\partial A}}{2\Sigma_s} - \frac{\frac{\partial \Sigma_d}{\partial A}}{2\Sigma_d} \tag{14}$$

where

$$\frac{\partial \mu_s}{\partial A} = 2 M_s A \tag{15}$$

$$\frac{\partial \mu_d}{\partial A} = 2 M_d A \tag{16}$$

$$\frac{\partial \Sigma_s}{\partial A} = 4E(Tr(A^T x_{is} x_{is}^T A) x_{is} x_{is}^T A) - 4Tr(A^T M_s A) M_s A \tag{17}$$

$$\frac{\partial \Sigma_d}{\partial A} = 4E(Tr(A^T x_{id} x_{id}^T A) x_{id} x_{id}^T A) - 4Tr(A^T M_d A) M_d A \tag{18}$$

From the above description, we can see that the proposed method tries to find the embedding subspace during learning the distance metric inspired by the Bayesian classification error estimation. The transformation A does not change the *k*-nearest neighborhood relationship of the data, which is similar to the local preserving property of manifold learning, but it is different from popular manifold learning methods in that it aims to make each sample far from its inter-person set and close to its intra-person set. Although we use the Gaussian distribution to model the the variables of the intra-person

distances and inter-person distances in the subspace, they are based on the local neighbors, so we do not make assumption on the global distribution of the data compared to LDA. Compared with most distance metric learning methods, the proposed method uses the Bhattacharyya coefficient measurement, which has intuitionistic geometric interpretation and links to Bayesian classification error under the Bayesian framework. The proposed method can handle high dimensional data well.

4 Experiments

We test the proposed method on the two benchmarks, i.e., the CMU-PIE face database [21] and the extended YALE face database [15]. The data of the two face databases are available in [1]. In our experiments, we take the PCA as the baseline, where we keep 98% energy of eigenvalues. We compare the proposed method with related works, i.e., LDA, RCA, and NCA. The codes of RCA and NCA are downloaded from [2] and [3] respectively. For RCA, we use the prior label information to form the chunklets. In addition, we also compare the proposed method with the Bayesian face subspace (BFS) [6]. In the Bayesian face subspace, the face images are modeled by the intra-face and the inter-face subspaces, which are represented by PCA directly in the input data space. For the Bayesian face subspace, we construct the principal subspace with the 90% energy of the eigenvalues, and the complemental subspace with the rest of 10% energy. In the experiments, we set the number of neighbors k as the training numbers of each class minus 1.

4.1 CMU-PIE Face Database

The CMU PIE face database contains 68 subjects and 41368 images [25]. Each subject has 13 different poses, 43 different illuminations, and 4 different expressions. In this paper, our dataset is composed of all the images from five near frontal poses $(C05, C07, C09, C27, C29)$ including all the illumination and expression variations as in [7] [1]. There are 170 face images for each subject in our dataset. The images are cropped by fixing two eyes, and the cropped image size is 32×32. No image preprocessing is performed except normalizing the image into unit vector as in [7] [1]. Figure 2 shows some samples of one subject.

We randomly select 30 images from each subject for training, and the other 140 images of each subject for testing. The experiments are randomly run 50 times, and all the results reported in Figure 3 are the average of 50 times experiments. Because there are 68 classes, the maximum feature dimension of LDA is 68-1 = 67. From Figure 3, we can see that MBC is better than PCA, LDA, RCA, NCA, and BFC. The minimum classification error of MBC is 5.46%, while those of PCA, LDA, RCA, NCA, and BSF are 29.4%, 7.84%, 14.62%, 6.76%, and 6.76% respectively. The performance of MBC is still better than the modified LPP [7]. In [7] [1] , the modified LPP obtained the minimum average classification error of 7.5% over 50 times experiments under the same testing protocol, i.e., 30 images are randomly selected from each subject, and the rest images of each subject are used for testing.

Fig. 2. Samples of the CMU-PIE database

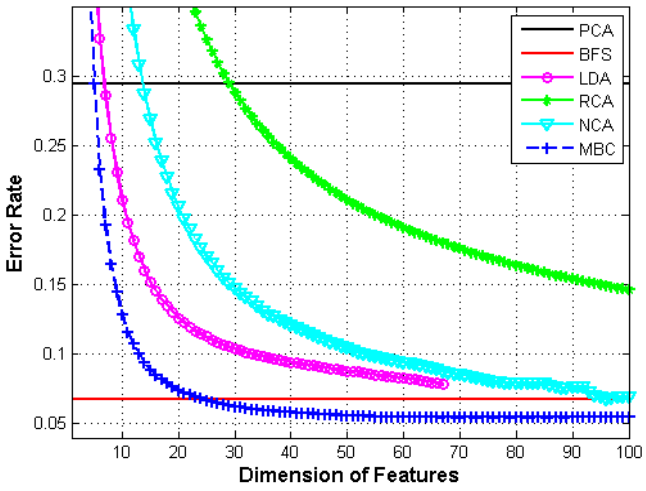

Fig. 3. Testing error rate on the CMU-PIE database

4.2 Extended YALE Face Database

The extended YALE face database has 38 subjects, each subjects has 64 near frontal view images under different illuminations [1] [15]. The images are cropped to 32 $\times$ 32, and images are normalized into unit vectors as in [7] [1]. Figure 4 shows some image samples. Same as the experiments on the CMU-PIE database, we randomly select 30 images from each individual for training, and the rest 34 images per subject are used for testing. The experiments are run 50 times, and Figure 5 reports their average results. Because the training data has 38 classes, the maximum feature dimensions of LDA is 38-1 = 37. The minimum classification error of MBC is 2.5%, while those of PCA, LDA, RCA, NCA, and BSF are 25.59%, 13.34%, 10.88%, 4.93%, and 3.93% respectively. The performance of MBC is still better than the modified LPP [7], for the minimum classification error of the latter reported is 7.5% under a similar testing in [7] [1].

Fig. 4. Samples of the extended YALE database

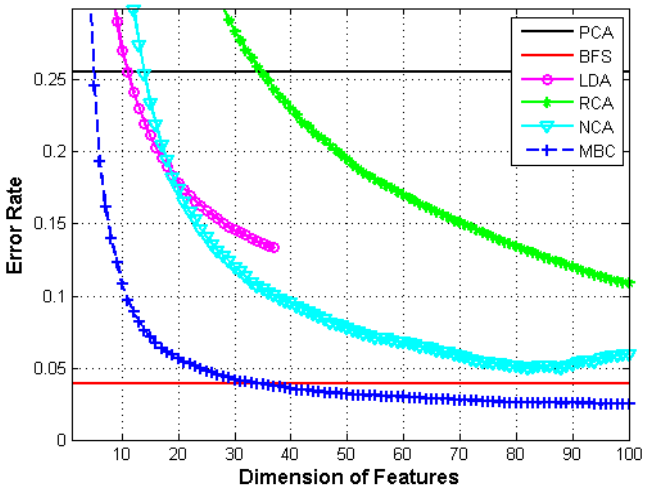

Fig. 5. Results on the extended YALE database

5 Conclusions

In this paper, we presented a unified scheme of subspace and distance metric learning under the Bayesian framework for face recognition. We divided the k-nearest neighbors of each sample into the intra-person set and the inter-person set according to the local distribution of the data, and we attempted to learn a distance metric in the embedding subspace, which made the distances between the sample and its intra-person set smaller than the distances between it and its inter-person set in the embedding subspace. To reach this goal, we defined two variables in the subspace, i.e., the intra-person distance and the inter-person distance, and modeled them with two different probabilistic distributions. Then we converted our problem to that of minimizing the overlap between these two distributions. Inspired by Bayesian classification error estimation, Our goal was equivalent to minimizing their Bhattachyrra coefficient measurement. The pro-

posed framework made no assumption on the global distribution of the data. Moreover, it links to Bayesian error. We proved the power of the proposed approach on the CMU-PIE face database and the extended YALE face database.

6 Acknowledgements

We would like to thanks the reviewers' comments. The first author would also like to thank for the support of the NSFC (No. 60405005 and 60675003). This work is done at Rutgers University.

References

1. http://ews.uiuc.edu/ dengcai2/data/data.html. 7, 8
2. http://www.cs.huji.ac.il/ aharonbh/. 7
3. http://www.eng.biu.ac.il/ goldbej/papers.html. 7
4. A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constrains. *Journal of Machine Learning Research*, 2005. 3
5. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 2
6. B.Moghaddam, T.Jebara, and A.Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000. 7
7. D. Cai, X. He, and J. Han. Using graph model for face analysis. *Tech Report UIUCDCS-R-2636, University of UIUC*, 2005. 7, 8
8. H. T. Chen, H. W. Chang, and T. L. Liu. Local discriminant embedding and its variants. In *Proc. of Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005. 3
9. D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003. 5
10. K. Fukunaga. Introduction to statistical pattern recognition. *Academic Press, New York*, 1990. 5, 6
11. A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems (NIPS)*, 2005. 3
12. J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood component analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2004. 3
13. X. F. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems (NIPS)*, 2003. 3
14. S. C. Hoi, W. Liu, M. R. Lyu, and W. Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *Proc. of Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006. 3
15. K.-C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):1–15, 2005. 2, 7, 8
16. S. Mika, G. Ratsch, and J. Weston. Fisher discriminant analysis with kernels. In *Proc. of Neural Networks for Signal Processing Workshop*, 1999. 2
17. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Sciences*, 290(5500):2323–2326, 2000. 2
18. R. Salakhutdinov and S. T. Roweis. Adaptive over- relaxed bound optimization methods. In *Proc. of Int. Conf. Machine Learning (ICML)*, 2003. 3
19. B. Scholkopf, A. Smola, and K. R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. 2

20. N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In *Europen Conf. on Computer Vision (ECCV)*, 2003. 3
21. T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Trans. on PAMI*, 25(12):1615–1618, 2003. 2, 7
22. M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proc. of Int. Conf. Machine Learning (ICML)*, 2006. 3
23. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Sciences*, 290(5500):2319–2323, 2000. 2
24. L. Torresani and K. C. Lee. Large margin component analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2006. 3, 4
25. M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):72–86, 1991. 1, 2
26. K. Q. Weinberger, J. Blitzer, and L. K. Saul. Metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2005. 3, 4
27. E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems (NIPS)*, 2004. 3
28. S. C. Yan, D. Xu, B. Y. Zhang, and H. J. Zhang. Graph embedding: A general framework for dimensionality reduction. In *Proc. of Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005. 3
29. L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI*, 2006. 3
30. W. Zhao, R. Chellappa, and P. J. Phillips. Subspace linear discriminant analysis for face recognition. *Tech Report CAR-TR-914, University of Maryland*, 1999. 2
31. W. Zhao, R. Chellappa, A. Rosenfeld, and P.J.Phillips. Face recognition: A literature survey. *CS-Tech Report-4167, University of Maryland*, 2000. 1