Eigenfaces and Fisherfaces

Presenter: Harry Chao
aMMAI 2010

1

Outline

- 1. Introduction:

  Face recognition & Dimensionality reduction
- 2. Principal Component Analysis
- 3. Linear Discriminant Analysis
- 4. Other methods
- 5. Conclusion
- 6. Reference

2

I. Introduction

What & Why is face recognition?

The definition:

Indentify or verify a person based on face

The motivation:

- Remarkable face recognition capability of human visual system
- Numerous important application:

ex: Surveillance & face ID

Community involved:

Neuroscience, Psychology, pattern recognition, computer vision, machine learning……

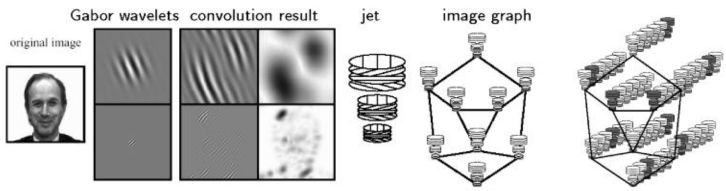
General techniques

Observation:

Images, videos, and 3D images

Based on images:

- Holistic-based methods (appearance)
- Feature-based methods (landmark)



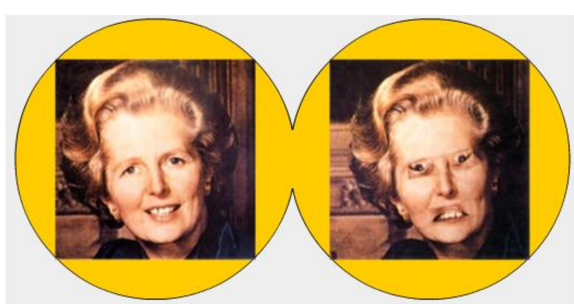
(a) Elastic Graph Representation (b) Bunch Graph

[5] Wiskott, 1997

5

Funny idea form psychology

- **Thatcher Illusion**



[13]

6

Challenge of face recognition

Challenge:
- Feature + Classifier
- The distance between different faces is not obvious!!

Distortion:
- illumination, pose, affine transform, expression, occlusion, noise


[9] Yang,2004


[7] He,2005

7

Dimensionality reduction

Why?
- The curse of dimensionality
- Intrinsic dimensionality may be smaller
- Some feature are not relevant

Idea:
- Reduce the feature dimension while preserving as much information as possible
- Decorrelation
- Extract the real distribution of the population

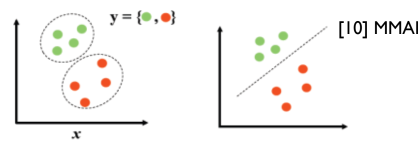
Methods:
- Feature selection & Feature reduction
- Supervised (LDA) & Unsupervised (PCA)

8

Meet with face recognition

Face recognition is a special case:

- Many classes but a few samples of each class
- KNN or other distance-based measurements may perform better than classifiers



For holistic-based:

Dimension reduction performs like data-driven features

For feature-based:

Dimension reduction is based on domain knowledge

2. PCA: Eigenfaces

General idea

Objective:

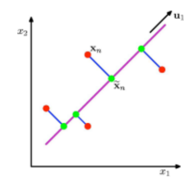
- Look for a few linear combinations, which can be used to summarize the data and loses in data as little as possible (want to preserve the variance)

For face recognition:

- A 256x256 face image is equivalent to a 665536-dim vector
- We want to reduce the dimension based on the database
- The new dimensionality depends on the number of images in the database

PCA is also known as:

- Karhunen-Loeve methods



[10] MMAI

11

Covariance matrix

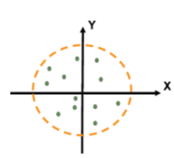
- The covariance matrix is symmetric with variances on the diagonal; assuming D dimensions (or variables)

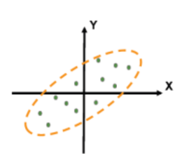
$$\Sigma = [\sigma_{ij}^2]_{D\times D} = \begin{bmatrix} \sigma_{0,0}^2 & \sigma_{0,1}^2 & & \sigma^2 0, D-1 \\ \sigma_{1,0}^2 & \sigma_{0,2}^2 & & \sigma_{1,D-1}^2 \\ & & \ddots & \\ \sigma_{D-1,0}^2 & \sigma_{D-1,1}^2 & & \sigma_{D-1,D-1}^2 \end{bmatrix}$$

- and covariance of two random variables (dimensions) $\mathbf{x}, \mathbf{y}$ is:

$$\sigma_{x,y}^2 = E[(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)] = \sum_{i=0}^{N-1} (x_i-\mu_x)(y_i-\mu_y)/N$$

  - Diagonal elements are individual variances in each dimension
  - Off-diagonal elements are covariance indicating data dependency between variables (dimensions in histogram)





[10] MMAI

12

Procedure for PCA

Linear projection:

- Originally N points in D-dim: $\{\mathbf{x}_i\}_{i=1}^{N} \in \mathcal{R}^D$
- A set of basis for projection: $\{\mathbf{u}_\mathrm{i}\}_{i=1}^{M} \in \mathcal{R}^D$
- These basis are orthonormal, and generally we have M<<D
- Preserve the reconstruction error as well as variance

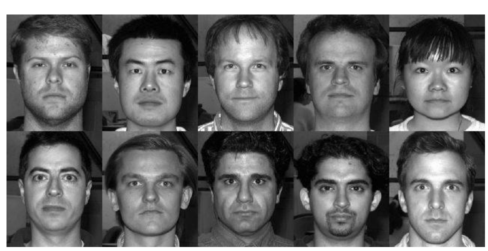
Procedure:

- Find the mean vector Ψ (D-by-1)
- Subtract each vector by Ψ and get Φ_i
- Calculate the covariance matrix Σ of Φ_i (*D*-by-*D*)
- Calculate the set of eigenvectors of Σ (*D*-by-*N* matrix)
- Preserve the *M* largest eigenvalues (*D*-by-*M* matrix *U*)
- U' Φ_i is the eigenfaces of the *i*th face (*M*-by-1)

13

Let's see an example

Face database



[11] Yale database

Eigenvectors

Mean face

14

Formula of PCA

- Assume we have Subtract each vector by Ψ and get *Φ*i, we want to find a projection vector *b* to minimize:

$$E[\| bb^{\mathrm{T}}\phi_i - \phi_i \|^2] = E[\| (bb^{\mathrm{T}} - I)\phi_i \|^2] = E[((bb^{\mathrm{T}} - I)\phi_i)^{\mathrm{T}}(bb^{\mathrm{T}} - I)\phi_i]$$

- Important tools

$$\boxed{\begin{aligned} &trace(scale) = scale \\ &trace(ABC) = trace(CAB) = trace(BCA) \end{aligned}}$$

$$\boxed{trace(E) = E(trace)}$$

$$b: \underset{1}{[\;\;]}^{D} \quad \phi_i : \underset{1}{[\;\;]}^{D} \quad bb^{\mathrm{T}} : \underset{D}{[\;\;]}^{D} \quad \phi_i\phi_i^{\mathrm{T}} : \underset{D}{[\;\;]}^{D}$$

$$\Sigma = E[\phi_i\phi_i^{\mathrm{T}}] : \underset{D}{[\;\;]}^{D}$$

- Then we can rewrite the formula

$$tr(E[((bb^{\mathrm{T}} - I)\phi_i)^{\mathrm{T}}(bb^{\mathrm{T}} - I)\phi_i]) = E[tr(\phi_i^{\mathrm{T}}(bb^{\mathrm{T}} - I)^{\mathrm{T}}(bb^{\mathrm{T}} - I)\phi_i)] = E[tr((bb^{\mathrm{T}} - I)^{\mathrm{T}}(bb^{\mathrm{T}} - I)\phi_i\phi_i^{\mathrm{T}})]$$

$$= E[tr((I - bb^{\mathrm{T}})\phi_i\phi_i^{\mathrm{T}})] = tr(E[\phi_i\phi_i^{\mathrm{T}}]) - tr(bb^{\mathrm{T}}E[\phi_i\phi_i^{\mathrm{T}}]) = tr(\Sigma) - tr(b^{\mathrm{T}}\Sigma b)$$

- Now we want to maximize: (Using Language multiplier)

$$tr(b^{\mathrm{T}}\Sigma b) = b^{\mathrm{T}}\Sigma b \text{ with } b^{\mathrm{T}}b = 1$$

Formula for eigenvectors

No we want to get the eigenvectors of Σ:

- Problem: Σ is of size 65536-by-65536 for 256-by-256 images

Solution:

$$\Sigma = E[\phi_i\phi_i^{\mathrm{T}}] = \text{constant} * \Phi\Phi^{\mathrm{T}} \;\; (\Phi \text{ is of } D\text{-by-}N)$$

$$\boxed{\text{We can first solve } \Phi^{\mathrm{T}}\Phi x = \lambda x}$$

$$\text{then do } \Phi\Phi^{\mathrm{T}}(\Phi x) = \lambda(\Phi x)$$

$$\text{where } \Phi\Phi^{\mathrm{T}} \text{ is of } D\text{-by-}D \text{ and } \Phi^{\mathrm{T}}\Phi \text{ of } N\text{-by-}N$$

$$\Phi : \underset{N}{[\;\;]}^{D} \quad \Phi\Phi^{\mathrm{T}} : \underset{D}{[\;\;]}^{D} \quad \Phi^{\mathrm{T}}\Phi : \underset{N}{[\;\;]}^{N} \quad \Sigma = E[\phi_i\phi_i^{\mathrm{T}}] : \underset{D}{[\;\;]}^{D}$$

Covariance matrix

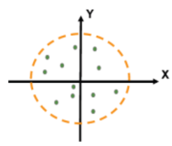
- The covariance matrix is symmetric with variances on the diagonal; assuming D dimensions (or variables)

$$\Sigma = [\sigma_{ij}^2]_{D\times D} = \begin{bmatrix} \sigma_{0,0}^2 & \sigma_{0,1}^2 & & \sigma^2 0, D-1 \\ \sigma_{1,0}^2 & \sigma_{0,2}^2 & & \sigma_{1,D-1}^2 \\ & & \ddots & \\ \sigma_{D-1,0}^2 & \sigma_{D-1,1}^2 & & \sigma_{D-1,D-1}^2 \end{bmatrix}$$

- and covariance of two random variables (dimensions) $\mathbf{x}, \mathbf{y}$ is:

$$\sigma_{x,y}^2 = E[(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)] = \sum_{i=0}^{N-1}(x_i-\mu_x)(y_i-\mu_y)/N$$

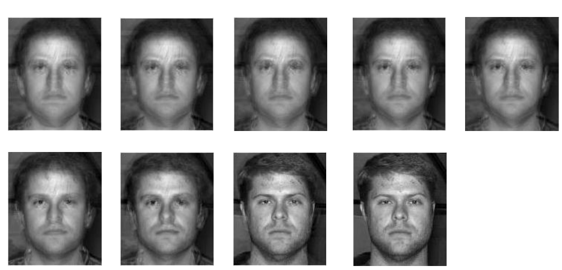
  - Diagonal elements are individual variances in each dimension
  - Off-diagonal elements are covariance indicating data dependency between variables (dimensions in histogram)



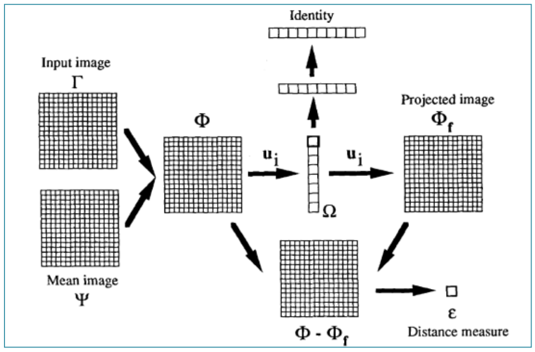
[10] MMAI

Example of face reconstruction

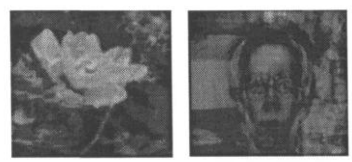
Reconstruction procedure



Eigenfaces for face recognition



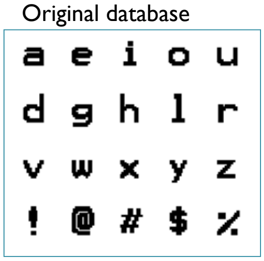
[1] Turk, 1991



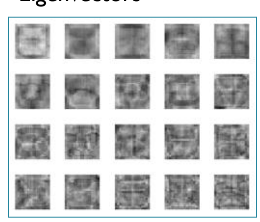
[1] Turk, 1991

Example of character recognition

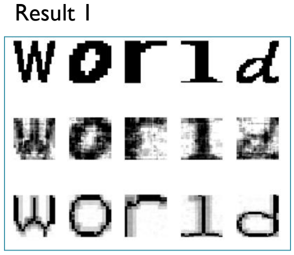
Original database



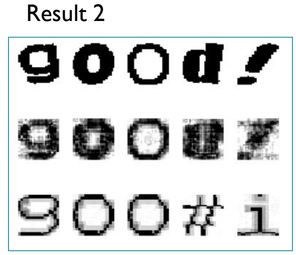
Eigenvectors



Result 1

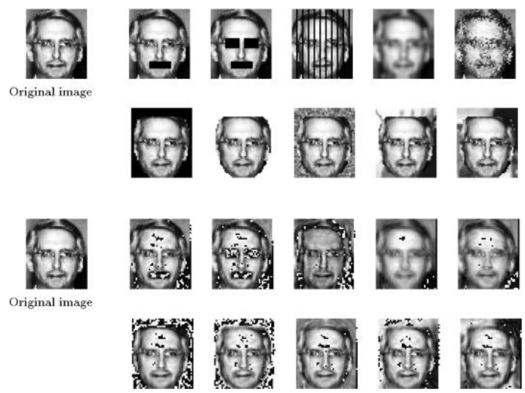


Result 2



Good properties of PCA

- Good for dealing with random noise, but not good for rotation-scaling-translation (RST) distortion. It could minimize the distance between projection space and data space, and really reduce the redundancy!



[3] Zhao, 2003

3. LDA: Fisherfaces

General idea (I)

Objective:

- Look for dimension reduction based on discrimination purpose

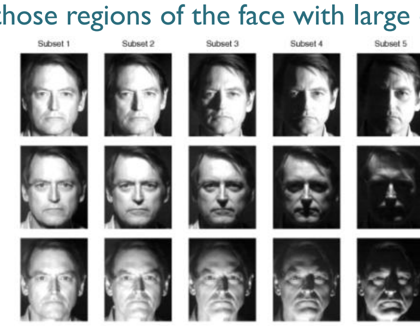
For face recognition:

- The variance among faces in the database may come from distortions such as illumination, facial expression, and pose variation. And sometimes, these variations are larger than variations among standard faces!!
- The images of a particular face, under varying illumination but fixed pose, lie in a 3D linear subspace of the high dimensional image space. (without shadowing)

General idea (II)

Idea:

- Try to find a basis for projection that minimize the intra-class variation but preserve the inter-class variation.
- Rather than explicitly modeling this deviation, we linearly project the image into a subspace in a manner which discount those regions of the face with large deviation



[3] Belhumeur, 1997

Fisher linear discriminant

inter-class: $|\tilde{m}_1 - \tilde{m}_2| = |w^{\mathrm{T}}(m_1 - m_2)|$

intra-class: $\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$

want to maximize: $J(w) = \dfrac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$

$y, \tilde{m}_1, \tilde{m}_2 : [\;]_1^1 \quad (w^{\mathrm{T}}x - w^{\mathrm{T}}m_i) : [\;]_1^1$

$x, w, m_1, m_2 : [\;]_1^{\mathrm{D}} \quad S_{\mathrm{B}}, S_{\mathrm{w}} : [\;]_{\mathrm{D}}^{\mathrm{D}}$

$$\tilde{s}_i^2 = \sum_{x \in D_i} (w^{\mathrm{T}}x - w^{\mathrm{T}}m_i)(w^{\mathrm{T}}x - w^{\mathrm{T}}m_i)^{\mathrm{T}} = \sum_{x \in D_i} w^{\mathrm{T}}(x - m_i)(x - m_i)^{\mathrm{T}} w = w^{\mathrm{T}} S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^{\mathrm{T}} S_1 w + w^{\mathrm{T}} S_2 w = w^{\mathrm{T}} S_{\mathrm{w}} w$$

$$|\tilde{m}_1 - \tilde{m}_2|^2 = (w^{\mathrm{T}}m_1 - w^{\mathrm{T}}m_2)^2 = w^{\mathrm{T}}(m_1 - m_2)(m_1 - m_2)^{\mathrm{T}} w = w^{\mathrm{T}} S_{\mathrm{B}} w$$

want to maximize: $J(w) = \dfrac{w^{\mathrm{T}} S_{\mathrm{B}} w}{w^{\mathrm{T}} S_{\mathrm{w}} w}$

$S_{\mathrm{B}} w = \lambda S_{\mathrm{w}} w$

Multiple discriminant analysis

$$S_{\mathrm{B}} = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^{\mathrm{T}} \quad \text{c-1}$$

$$S_{\mathrm{w}} = \sum_{i=1}^{c} \sum_{x \in D_i} (x - m_i)(x - m_i)^{\mathrm{T}} \quad \text{N-c}$$

want to maximize: $J(W) = \dfrac{|W^{\mathrm{T}} S_{\mathrm{B}} W|}{|W^{\mathrm{T}} S_{\mathrm{w}} W|}$

with $W = [w_1 \; w_2 \ldots\ldots w_{\mathrm{m}}]$

$S_{\mathrm{B}} w_i = \lambda_i S_{\mathrm{w}} w_i$

$m \le c - 1$

$S_{\mathrm{B}}, S_{\mathrm{w}} : [\;]_{\mathrm{D}}^{\mathrm{D}} \quad W : [\;]_{\mathrm{m}}^{\mathrm{D}} \quad W_{\mathrm{PCA}} : [\;]_{\mathrm{N\text{-}c}}^{\mathrm{D}}$

Problem: S_{w} is always singular

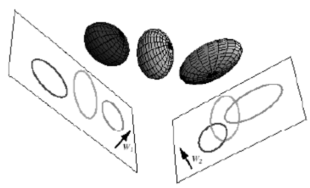
Fisherface solution:

$$W_{\mathrm{PCA}} = \arg \max_W |W^{\mathrm{T}} S_{\mathrm{T}} W| \quad \text{where } S_{\mathrm{T}} = \sum_x (x - m)(x - m)^{\mathrm{T}}$$

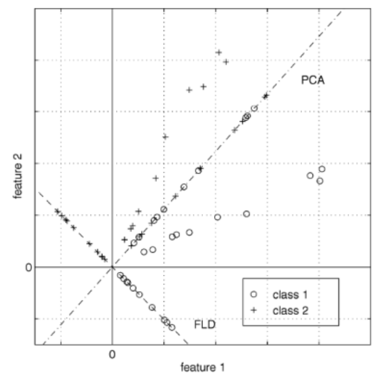
$$W_{\mathrm{FLD}} = \arg \max_W \frac{|W^{\mathrm{T}} W_{\mathrm{PCA}}{}^{\mathrm{T}} S_{\mathrm{B}} W_{\mathrm{PCA}} W|}{|W^{\mathrm{T}} W_{\mathrm{PCA}}{}^{\mathrm{T}} S_{\mathrm{w}} W_{\mathrm{PCA}} W|}$$

S_{T} is called the total scatter matrix

PCA vs. LDA (I)



[3] Duda, 2000



[3] Belhumeur, 1997

27

PCA vs. LDA (II)

PCA:

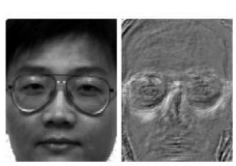
- The performance is weaker than correlation

LDA:

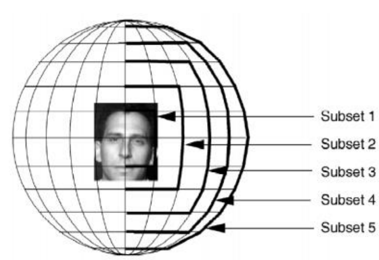
- LDA can be used for any kinds of classification problems
- Ex. Glasses recognition

Experimental types:

- Extrapolation & Interpolation
- Leaving-one-out



[3] Belhumeur, 1997



[3] Belhumeur, 1997

28

4.Other methods

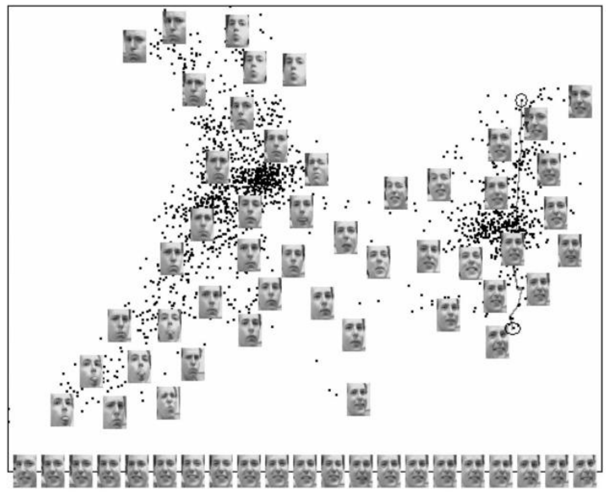
Other methods

- The combination of PCA & LDA: [Zhao,1998]
  - Use PCA for noise cleaning and generalization when only a few samples in each class
- The use of 2-D PCA: [Yang, 2004]

$$\boxed{\Sigma = E[(A-E[A])^{\mathrm{T}}(A-E[A])]} \qquad A:\underset{n}{[\;]}^{m}$$

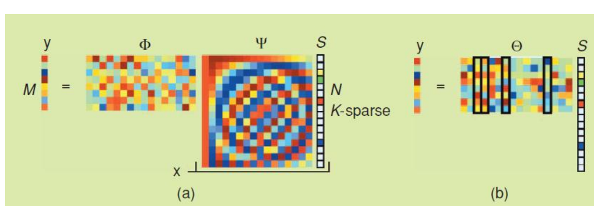
- Laplacianfaces: [He ,2005]
  - Extract the low-dimensional manifold structure
- Robust face recognition: [Wright, 2007]
  - Involved compressive sensing, sparse representation, and L1 minimization
  - Feature extraction is no longer important

Laplacianfaces

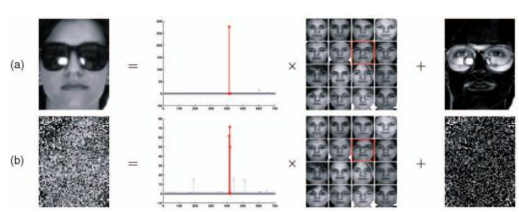


Robust face recognition

- Robust for occlusion, and the feature extraction is no longer important!



[Baraniuk, 2007]



[Wright, 2007]

5. Conclusion

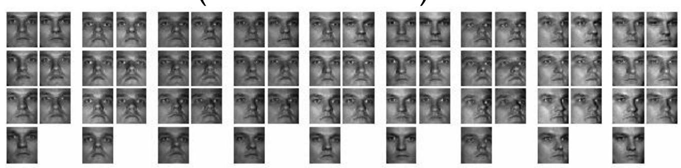
PCA vs. LDA

- PCA is an unsupervised dimension reduction algorithm, while LDA is supervised
- PCA is good at outlier cleaning, and LDA could learn the within-class deviation
- These two methods only extract 1st and 2nd statistical moments
- The combination of PCA & LDA could enhance the performance
- PCA serves as the first-step processing of several kinds of face recognition technique
- Techniques of dimension reduction are frequently used in face recognition

Database

- FERET database

- Yale database (suitable for LDA)



- More resources http://www.face-rec.org/

Reference

[1] M.Turk and A. Pentland,"Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, 1991.

[2] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE PAMI*, 1997.

[3] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, "Face Recognition: A Literature Survey, " *ACM Computing Surveys*, 2003, pp. 399-458

[4] W. Zhao, A. Krishnaswamy, R. Chellappa, "Discriminant Analysis of Principal Components for Face Recognition," In *Proceedings, International Conference on Automatic Face and Gesture Recognition. 336–341.*

[5] L. Wiskott, J.-M. Fellous, C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Patt. Anal. Mach. Intell. 19, 775–779,*1997.

[6] Richard Duda, et. al., Pattern Classification, 2nd Edition, Wiley-Interscience, 2000

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face Recognition Using Laplacianfaces," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 328-340, Mar. 2005.

[8] J. Wright, A. Ganesh, A. Yang, and Y. Ma, "Robust face recognition via sparse representation," Technical Report, University of Illinois, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[9] J. Yang, D. Zhang, A. Frangi, and J. Yang, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 1, pp. 131-137, Jan. 2004.

[10] W. Hsu, "Multimedia Analysis and Indexing – Course Website," 2009. [online] Available: http://www.csie.ntu.edu.tw/~winston/courses/mm.ana.idx/index.html. [Accessed Oct. 21, 2009].

[11] Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," IEEE Trans. Pattern Anal. Mach. Intelligence, 2001

[12] R. Baraniuk, "Compressive Sensing," IEEE Signal Processing Magazine, 2007

[13] http://www.michaelbach.de/ot/fcs_thompson-thatcher/index.html