LDA 혼합 모형을 이용한 얼굴 인식
(Face Recognition using LDA Mixture Model)

김 현 철 † 김 대 진 †† 방 승 양 †††
(Hyun-Chul Kim) (Daijin Kim) (Sung-Yang Bang)

요 약 LDA는 데이타를 잘 구분하게 하는 변환을 제공하고, 얼굴 인식에서 우수한 성능를 보였다. 그러나, LDA는 전체 데이타에 대해 단 하나의 변환 행렬만을 주므로 사람 얼굴과 같은 많은 클래스로 구성되어 있는 복잡한 데이타를 구분하기에 충분하지 않다. 이런 약점을 극복하기 위해 우리는 LDA 혼합 모형이라는 새로운 얼굴 인식 방법을 제안한다. LDA 혼합 모형에서는 모든 클래스가 여러 개의 군집으로 분할되고 각 군집에 대해서 하나의 변환 행렬을 얻는다. 이렇게 더 세세히 표현하는 방법은 분류 성능을 크게 향상시킬 것이다. 얼굴 인식 실험 결과, LDA 혼합 모형은 PCA, LDA, PCA 혼합 모형보다 더 우수한 분류 성능을 보여주었다.
키워드 : LDA, LDA 혼합 모형, PCA, PCA 혼합 모형, 얼굴인식

Abstract LDA (Linear Discriminant Analysis) provides the projection that discriminates the data well, and shows a very good performance for face recognition. However, since LDA provides only one transformation matrix over whole data, it is not sufficient to discriminate the complex data consisting of many classes like human faces. To overcome this weakness, we propose a new face recognition method, called LDA mixture model, that the set of all classes are partitioned into several clusters and we get a transformation matrix for each cluster. This detailed representation will improve the classification performance greatly. In the simulation of face recognition, LDA mixture model outperforms PCA, LDA, and PCA mixture model in terms of classification performance.
Key words : LDA, LDA mixture model, PCA, PCA mixture model, face recognition

1. Introduction

Face recognition is an active research area spanning several research fields such as image processing, pattern recognition, computer vision, and neural networks [1,2]. Face recognition has many applications, such as the biometrics system, the surveillance system, and the content-based video processing system. However, there are still many open problems, such as the recognition under illumination variations.

There are three main approaches for face recognition [1-3]. The first approach is the feature-based matching approach using the relationship between facial features such as eyes, mouth and nose. The second approach is the template matching approach using the holistic features of face image. The third approach is the hybrid approach combining the first and second approaches.

The eigneface method is a well-known template matching method [4]. Face recognition using the eigenface method is performed using feature values transformed by PCA. LDA has also been applied to face recognition. Face recognition method using LDA is called the fisherface method [5], and provides much better classification performance than the eigenface method due to its use of class information [5,6]. LDA is known as one of the best face recognition techniques. ICA has been also proposed for template matching face recognition method [7], and it could be used as filters for light-invariant recognition [8].

† 비 회 원 : 포항공과대학교 컴퓨터공학과
grass@postech.ac.kr
†† 비 회 원 : 포항공과대학교 컴퓨터공학과 교수
dkim@postech.ac.kr
††† 정 회 원 : 포항공과대학교 컴퓨터공학과 교수
sybang@postech.ac.kr
논문접수 : 2004년 12월 9일
심사완료 : 2005년 7월 12일

Although LDA usually gives a very good discrimination performance, LDA has a drawback in that LDA has only one transformation matrix for any data. For many-class complex data like face image data, only one transformation matrix is not be enough to provide a good discrimination performance. In such a case, more than one transformation matrixes are required.

To overcome this drawback of LDA, we propose LDA mixture model, which is using several transformation matrices. We partition the set of all classes into several clusters, and apply LDA technique for each cluster. In this way, we obtain several transformation matrixes. These transformation matrixes can be used cluster-wisely for recognition. For face image data, each cluster may correspond to a race or a group of similar faces. By using more than one transformation matrixes, we can get a better discrimination performance.

This paper is organized as follows. Section 2 describes the LDA. Section 3 describes PCA mixture model and LDA mixture model. Section 4 shows and discusses the simulation results for face recognition. Finally, we present our conclusions.

2. LDA

The goal of LDA is to find an orientation for which the projected samples are well separated [9]. Specifically, LDA seeks a transformation matrix $\boldsymbol{W}$ that in some sense maximizes the ratio of the between-class scatter to the within-class scatter. Initially, we consider a within-class scatter matrix for the within-class scatter. A within-class scatter matrix $\boldsymbol{S_W}$ is defined as

$$\boldsymbol{S_W} = \sum_{i=1}^{c}\sum_{x\in C_i}(\boldsymbol{x}-\boldsymbol{m_i})(\boldsymbol{x}-\boldsymbol{m_i})^t, \quad (1)$$

where c is the number of classes, C_i is a set of data belonging to the ith class, and $\boldsymbol{m_i}$ is the mean of the ith class. The within-class scatter matrix represents the degree of scatter within classes as a summation of covariance matrices of all classes. Next, we consider a between-class scatter matrix for between-class scatter. A between-class scatter matrix $\boldsymbol{S_B}$ is defined as

$$\boldsymbol{S_B} = \sum_{i=1}^{c} n_i(\boldsymbol{m_i}-\boldsymbol{m})(\boldsymbol{m_i}-\boldsymbol{m})^t \quad (2)$$

The between-class scatter matrix represents the degree of scatter between classes as a covariance matrix of means of all classes.

We seek a transformation matrix $\boldsymbol{W}$ that in some sense maximizes the ratio of the between-class scatter and the within-class scatter. The criterion function $J(\boldsymbol{W})$ can be defined as

$$J(\boldsymbol{W}) = \frac{|\boldsymbol{W}^t\boldsymbol{S_B}\boldsymbol{W}|}{|\boldsymbol{W}^t\boldsymbol{S_W}\boldsymbol{W}|}. \quad (3)$$

We can obtain the transformation matrix $\boldsymbol{W}$ as one that maximizes the criterion function $J(\boldsymbol{W})$. The columns of optimal $\boldsymbol{W}$ are the generalized eigenvectors $\boldsymbol{w_i}$ that correspond to the largest eigenvalues in

$$\boldsymbol{S_B}\boldsymbol{w_i} = \lambda_i \boldsymbol{S_W}\boldsymbol{w_i}. \quad (4)$$

3. LDA Mixture Model

As mentioned before, LDA mixture model partition the set of all classes to several clusters. Clustering can be performed by PCA mixture model, and the formulation of LDA mixture model becomes simpler if we use PCA mixture model. Therefore, we use PCA mixture model for dividing the samples into several clusters for LDA mixture model. Here, we explain PCA mixture model.

3.1 PCA Mixture Model

We consider a PCA mixture model of the data x by combining PCA and a mixture model in a way that the component density of the mixture model can be estimated onto the PCA transformed space as

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} P(\boldsymbol{x}|c_k,\theta_k)P(c_k). \quad (5)$$

$$P(\boldsymbol{x}|c_k,\theta_k) = P(\boldsymbol{s_k}|c_k,\theta_k), \quad (6)$$

where $\boldsymbol{s}_k = \boldsymbol{T_k}^T(\boldsymbol{x}-\boldsymbol{\mu_k})$. The PCA feature vectors $\boldsymbol{s_k}$ are decorrelated due to the orthogonality of the transform matrix $\boldsymbol{T_k}$. Thus, its covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{S_k}} = E[\boldsymbol{s}_k\boldsymbol{s}_k^T]$ is a diagonal matrix whose diagonal elements are corresponding to the principal eigenvalues. Next, the conditional density function $P(\boldsymbol{s}_k|k,\theta_k)$ of the PCA feature vectors in the kth cluster can be simplified as

$$P(\boldsymbol{s_k}|c_k,\theta_k)=\frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma_k^{s_k}|^{\frac{1}{2}}}exp(-\frac{1}{2}\boldsymbol{s_k}^T\Sigma_k^{-1s_k}\boldsymbol{s_k}) \tag{7}$$
$$=\prod_{j=1}^{m}\frac{1}{(2\pi)^{\frac{1}{2}}\lambda_{k,j}^{\frac{1}{2}}}exp(-\frac{s_j^2}{2\lambda_{k,j}}),$$

where $\boldsymbol{s}=(\boldsymbol{s_1},\cdots,\boldsymbol{s_m})$, and $(\lambda_{k,1},\cdots,\lambda_{k,m})$ are the m dominant eigenvalues of the feature covariance matrix Σ^{S_k} in the kth cluster. The proposed model,which has no Gaussian error term, can be considered as a simplified form of Tipping and Bishop's model [10].

The parameters of a mixture model can be estimated by an EM algorithm, which maximizes the likelihood as follows [11].

(1) E-step

Given the feature data set $\boldsymbol{X}$ and the parameters $\Theta^{(t)}$ of the mixture model at the tth iteration, we estimate the posterior distribution $P(z|\boldsymbol{x},\Theta^{(t)})$ using

$$P(z|\boldsymbol{x},\Theta^{(t)})=\frac{P(\boldsymbol{x}|z,\Theta^{(t)})P(z)}{\sum_{k=1}^{K}P(\boldsymbol{x}|k,\Theta^{(t)})P(k)}, \tag{8}$$

where $P(\boldsymbol{x}|z,\Theta^{(t)})$ is calculated by Eq. 6 and 7.

(2) M-step

Next, The new means $\boldsymbol{\mu}_k^{X(t+1)}$ and the new covariance matrixes $\Sigma_k^{X^{(t+1)}}$ of the kth mixture component are obtained by the following update formula.

$$\boldsymbol{\mu}_k^{X(t+1)}=\frac{\sum_{p=1}^{N}P(k|\boldsymbol{x}_{p,\Theta^{(t)}})\boldsymbol{x}_p}{\sum_{p=1}^{N}P(k|\boldsymbol{x}_{p,\Theta^{(t)}})} \tag{9}$$

$$\Sigma_k^{(t+1)X}=\frac{\sum_{p=1}^{N}P(k|\boldsymbol{x}_p,\Theta^{(t)})(\boldsymbol{x}_p-\boldsymbol{\mu}_k^{(t)X})^T(\boldsymbol{x}_p-\boldsymbol{\mu}_k^{(t)X})}{\sum_{p=1}^{N}P(\boldsymbol{x}_p|\Theta^{(t)})} \tag{10}$$

The new eigenvalue parameters $\lambda_{k,j}^{(t+1)}$ and the new eigenvetor (PCA basis) parameters $\boldsymbol{t}_{k,j}$ are obtained by selecting the largest m eigenvalues in the eigenvector computation (PCA computation) as

$$\Sigma_k^{(t+1)X}\boldsymbol{t}_{k,j}=\lambda_{k,j}^{(t+1)}\boldsymbol{t}_{k,j}. \tag{11}$$

PCA transformation matrixes $\boldsymbol{T_k}$ is obtained as $[\boldsymbol{t_{k,1}}\ \boldsymbol{t_{k,2}}\ldots\boldsymbol{t_{k,m}}]$. The mixing parameters $P(z_p)$ can be estimated as follows.

$$P(k)=\frac{1}{N}\sum_{p=1}^{N}P(k|\boldsymbol{x}_p,\Theta^{(t)}). \tag{12}$$

The above two steps will be repeated until a stopping condition is satisfied, where the three parameters will not be changed any further.

3.2 LDA Mixture Model

LDA gets one transformation matrix over all classes. This property degrades the performance of LDA because only one transformation matrix is not enough for the classification of complex data with many classes. To overcome this limitation, we propose to use LDA mixture model that uses several transformation matrices over all classes. LDA mixture model partitions the set of all classes into an appropriate number of clusters and applies LDA to each cluster, independently.

Specifically, we apply PCA mixture model to the set of means $\boldsymbol{m_i}$ of each class with K mixture components. Then, for the kth mixture component, we obtain a cluster mean $\boldsymbol{c_k}$, a transformation matrix $\boldsymbol{T_k}$, and a diagonal matrix $\boldsymbol{V_k}$ with eigenvalues, where $\boldsymbol{V_k}$ is a diagonal matrix whose diagonal element is eigenvalues $\boldsymbol{\lambda_{k,j}}$ which is the j th largest eigenvalue of the covariance matrix . In this case, the probabilistic covariance matrix for the kth mixture component is $\boldsymbol{T_k V_k T_k^t}$. Using this result, we get the between-class scatter matrix and the within-class scatter matrix for the k mixture component (which is corresponding to the kth cluster) as

$$\boldsymbol{S_{Bk}}=\boldsymbol{T_k V_k T_k^t} \tag{13}$$

$$\boldsymbol{S_{Wk}}=\sum_{l\in L_k}\frac{1}{n_l}\sum_{\boldsymbol{x}\in C_l}(\boldsymbol{x}-\boldsymbol{m_l})(\boldsymbol{x}-\boldsymbol{m_l})^t, \tag{14}$$

where $k=1,2,\ldots,K, l=1,2,\ldots,c, L_k$ is a set of class labels belonging to the kth mixture component that is determined by $L_k=j|j=\mathrm{argmin}_i(\boldsymbol{m_i}-\boldsymbol{c_k})S_{Bj}^{-1}(\boldsymbol{m_i}-\boldsymbol{c_k})^t$.

Based on $\boldsymbol{S_W}$ and $\boldsymbol{S_B}$, the transformation matrix $\boldsymbol{W_k}$ for the $\boldsymbol{k}$th mixture component is determined so as to maximize the criterion function

$$J_k(\boldsymbol{W})=\frac{|\boldsymbol{W^t S_B W}|}{|\boldsymbol{W^t S_{Wk} W}|}. \tag{15}$$

The columns of optimal $\boldsymbol{W_k}$ are the generalized eigenvectors $\boldsymbol{w_{k_j}}$ that correspond to the largest eigenvalues in

$$\boldsymbol{S_B w_{jk}}=\lambda_{jk}\boldsymbol{S_{Wk} w_{k_j}} \tag{16}$$

4. Simulation Results and Discussion

We took a partial set of PSL database obtained from the MPEG7 community [12]. PSL database consists of normalized images of 271 people, where images of some people have lighting variations and images of the other people have pose variations.

Among all images, We selected images of 133 people with lighting variations. In this partial set, there are five images for each person, having normal, smiling, angry expression and images of left lighting and right lighting. Since the number of images for each class is too small, we used 5-fold cross-validation to test the classification performances. Figure 1 shows some images used in the simulation.

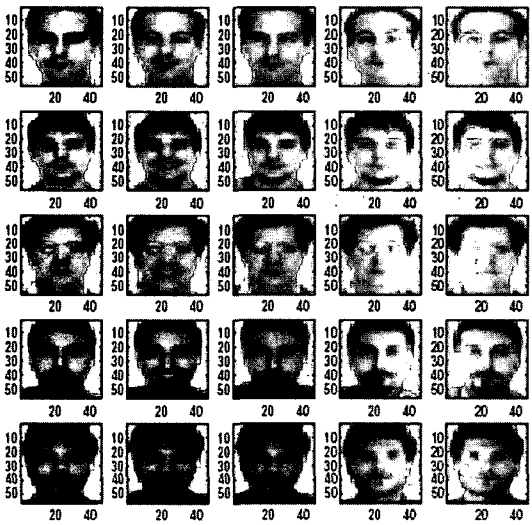

Figure 1 Some face images used in the simulation

We applied PCA, PCA mixture model, LDA, LDA mixture model and ICA to a training set. We used the Fast ICA algorithm for ICA [13]. We used the following recognition method for each method. For PCA, LDA and ICA, we transform the training data $\boldsymbol{x}_r$ and the test data $\boldsymbol{x}$ by a transformation matrix ($\boldsymbol{T}$ for PCA, $\boldsymbol{W}$ for LDA, $\boldsymbol{Z}$ for ICA), and assign the test data $\boldsymbol{x}$ to the class $C_{PCA(LDA)}$ of the transformed training data that is nearest to the transformed test data as

$$C_{PCA} = L(\arg\min_{\boldsymbol{x}_r} |(\boldsymbol{x}-\boldsymbol{x}_r)\boldsymbol{T}|), \quad (17)$$

$$C_{LDA} = L(\arg\min_{\boldsymbol{x}_r} |(\boldsymbol{x}-\boldsymbol{x}_r)\boldsymbol{W}|), \quad (18)$$

and

$$C_{ICA} = L(\arg\min_{\boldsymbol{x}_r} |(\boldsymbol{x}-\boldsymbol{x}_r)\boldsymbol{Z}|), \quad (19)$$

where $\boldsymbol{x}_r$ is a sample in the training set and $L(\boldsymbol{x}_r)$ indicates the class label of a sample $\boldsymbol{x}_r$. For PCA mixture model and LDA mixture model, we transform the test data $\boldsymbol{x}$ and the training data $\boldsymbol{x}_r$ by a transformation matrix of a corresponding mixture component ($\boldsymbol{T}_{I(\boldsymbol{x}_r)}$ for PCA mixture model, $\boldsymbol{W}_{I(\boldsymbol{x}_r)}$ for LDA mixture model), and assign the data to the class $C_{PM(LM)}$ of corresponding transformed training data that is nearest to the transformed test data as

$$C_{PM} = L(\arg\min_{\boldsymbol{x}_r} |(\boldsymbol{x}-\boldsymbol{x}_r)\boldsymbol{T}_{I(\boldsymbol{x}_r)}|) \quad (19)$$

and

$$C_{LM} = L(\arg\min_{\boldsymbol{x}_r} |(\boldsymbol{x}-\boldsymbol{x}_r)\boldsymbol{W}_{I(\boldsymbol{x}_r)}|), \quad (20)$$

where $\boldsymbol{x}_r$ is a sample in the training set, $L(\boldsymbol{x}_r)$ indicates the class label of a sample $\boldsymbol{x}_r$, and $I(\boldsymbol{x}_r)$ indicates the mixture component label, to which the training sample $\boldsymbol{x}_r$ belongs, determined by $I(\boldsymbol{x}_r) = \arg\min_i (\boldsymbol{m}_i - \vec{c}_k)\boldsymbol{S}_{\boldsymbol{B}j}^{-1}(\boldsymbol{m}_i - \boldsymbol{c}_k)^t$.

Even though more than 2 mixture components have been taken, it does not show any significant improvement in classification performance. So, we used only two mixture components for learning the PCA mixture model for each class in PCA mixture model and LDA mixture model. Figure 2 plots the classification error rates according to a different number of features for the test data when each method is applied. The number of features and classification error rates in the best case are shown in Table 1. ICA does not have any order of its components and so its result is shown only in Table 1. The classification error rates in Figure 2 and Table 1 are averages of error rates in 5-fold cases. From Figure 2, we note that the LDA mixture model provides the best classification performance. For a small number of features, PCA mixture model outperforms LDA mixture model, but LDA mixture model outperforms PCA mixture model for more than 35 features. We also note that the classification performance of LDA is better than that of PCA, and both PCA mixture model and LDA mixture model outperform PCA and LDA.

From Table 1, ICA is quite good as it is known to be good filters for light-invariant face recognition, but it is outperformed by LDA mixture model.

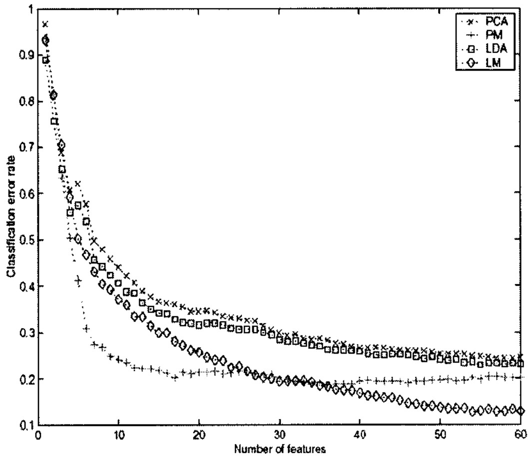

Figure 2 Classification error rates vs. the number of features

Table 1 The best performance and the corresponding number of features

Methods	Number of features	test error
PCA	92	22.41%
PCA mixture	36	18.65%
LDA	84	21.50%
LDA mixture	54	12.78%
ICA	100	17.44%

5. Conclusion

We proposed LDA mixture model to overcome a limitation of LDA, which provided several transformation matrices by taking one transformation matrix for each cluster. This modification tries to cluster whole data by PCA mixture model and to make the data discrimination better by the use of several cluster-wise transformation matrixes. The simulation results showed that LDA mixture model outperformed PCA, LDA, ICA, PCA mixture model in terms of classification errors. For the complex data consisting of many classes in the problem of face recognition, LDA mixture model can be used for an alternative of LDA.

References

[1] R. Chellappa, C. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," Proc. IEEE, 83(5), 705-740, 1995.

[2] W. Zhao, R. Chellappa, P.J. Phillips and A. Rosenfeld, "Face Recognition: A Literature Survery," UMD CfAR Technical Report CAR-TR-948, 2000.

[3] R. Brunelli, T. Poggio, "Face recognition : feature versus templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 10, 1042-1052, 1993.

[4] M. Turk and A. Pentland, "Eigenfaces for Recognition," J. Cognitive Neuroscience, 3(1), 1991.

[5] P. Belhumeur, J. Hespanha, D. Kriegman, "Eigenfaces vs. Fisherfaces: class specific linear projection," IEEE Transactions on PAMI, 19(7), 711-720, 1997.

[6] K. Etemad, R. Chellappa, "Discriminant analysis for recognition of human face images," Journal of the Optical Society of America A, 14(8), 1724-1733, 1997.

[7] M.S. Bartlett, J.R. Movellan and T.J. Sejnowski, "Face recognition by independent component analysis," IEEE Transactions on Neural Networks 13(6), 2002.

[8] J. Fortuna and D. Capson, "ICA Filters For Lighting Invariant Face Recognition," Proceedings of the 17th International Conference on Pattern Recognition, 2004.

[9] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, New York, 2001.

[10] M. Tipping, C. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation, 11, 443-482, 1999.

[11] P. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society: series-B, 39(4), 1-38, 1977.

[12] L. Wang and T. K. Tan, "Experimental results of face description based on the 2nd-order eigenface method," ISO/MPEG m6001, Geneva, May, 2000.

[13] Hyvrinen A. and Oja E., "A Fast Fixed-Point Algorithm for Independent Component Analysis, Neural Computation, vol. 9, no. 7, pp. 1483-1492, 1997.

김 현 철

1999년 포항공대 컴퓨터공학 학사, 수학 학사(복수전공). 2001년 포항공대 컴퓨터 공학 석사. 2001년 이후 포항공대 컴퓨터 공학 박사과정 재학. 관심분야는 패턴인식, 기계학습 등

김 대 진

1981년 연세대학교 전자공학과 학사. 1983년 KAIST 전기 및 전자공학과에서 석사. 1991년 미국 Syracuse University 컴퓨터공학과에서 박사를 받았음. 한국방송공사 기술연구소, 동아대학교 컴퓨터공학과에 근무하다가 1999년 7월 이후 포항공과대학교 컴퓨터공학과에서 근무. 관심분야는 패턴인식, 지능형 시스템, 생체 인식 등

방 승 양

1966년 일본 Kyoto대학 전기공학에서 학사. 1969년 서울대학교 전기공학에서 석사. 1974년 미국 University of Texas 전산학에서 박사를 받았음. 미국 Wayne State University, NCR, Bell 연구소 등에서 근무하다가 1981년 귀국. 한국전자기술연구소 시스템부 실장, 부장 역임, (주)유니온시스템 전무. 1986년부터 포항공대 컴퓨터공학과 교수. 현재 뇌연구센터 소장. 관심분야는 패턴인식, 신경회로망