Analysis of Different Subspace Mixture Models in Handwriting Recognition

V.N. Manjunath Aradhya[1] & S.K. Niranjan [2]
[1]*Dept. of ISE, Dayananda Sagar College of Engineering, Bangalore, India*
[2]*Dept. of MCA, S.J. College of Engineering, Mysore, India*
Email: aradhya1980@yahoo.co.in, sriniranjan@yahoo.com

Abstract

In this paper we explore, analyze and propose the idea of subspace mixture models such as Principal Component Analysis (PCA), Fisher's Linear Discriminant Analysis (FLD) and Laplacian in handwriting recognition. Statistically, Gaussian Mixture Models (GMMs) are among the most suppurate methods for clustering (though they are also used intensively for density estimation). By modeling each class into a mixture of several components and by performing the classification in the compact and decorrelated feature space it may result in better performance. To do this, each character class is partitioned into several clusters and each cluster density is estimated by a Gaussian distribution function in the PCA, FLD and Laplacian transformed space. The analysis of different mixture models are experimented out on handwritten Kannada characters.

1. Introduction

In character recognition, a popular feature extraction methods include geometric features, structural features, and feature space transformations methods. Out of these, transformation methods usually used for reducing the dimensionality of features, and some of them can also improve the classification accuracy [4]. Feature extraction from data or a pattern is a necessary step in pattern recognition and can raise generalization of subsequent classification and avoid notorious curse of dimensionality. Appearance-based schemes, which uses the holistic texture features are successfully developed for face recognition. The objective of subspace based approach is to project the data of faces onto a dimensionally reduced space where the actual recognition will be carried out.

In face recognition, Turk and Pentland [13] first explored the PCA and used the PCA projected components as features. The PCA does not include the label information of the data due to unsupervised in nature. In order to employ the class label information of the data, Fisher's Linear Discriminant Analysis (FLD) was proposed [12, 2]. This is unlike the PCA method, which searches for basis vectors that best describes the data. The main objective of FLD is to maximize the between-class measure while minimizing the within-class measure. Due to large dimensions and singularity of within-class matrix,implementation of the LDA method becomes an stubborn task [1]. As an alternative to the PCA, the Locality Preserving Projections (LPP) also known as Laplacian faces, was proposed which optimally preserves the neighborhood structure of the data set [5]. The LPP shares many of the data representation properties of nonlinear techniques such as Laplacian Eigenmaps or Locally Linear Embedding.

The above said algorithms are the state-of-the-art subspace methods proposed for face recognition [3]. Many variants of these algorithms are devised to overcome specific anomaly such as storage burden, computational complexity, etc. To enhance the performance of PCA and LDA methods, instead of extracting single set of features, more than one set of features were extracted by using Gaussian Mixture Models [3, 7, 8]. In the similar lines, the LPP method was also extended by extracting more than one set of features using Gaussian Mixture Models [11].

Motivated from the above facts and due to inherent advantages of subspace methods with mixture models, in this work we study and explore the techniques of PCA, FLD and Laplacian Mixture models in handwriting recognition. The rest of the paper is organized as follows: Section 2 presents a brief review of PCA, FLD and Laplacian Mixture models. Section 3 presents the experiment results and comparative analysis. Finally conclusion and future work is drawn at the end.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.178

2. Subspace Mixture Models

In this section, we present method based on PCA, FLD and Laplacian mixture model. This is incited by the idea that the classification accuracy is improved by modeling each class into a mixture of several components and by performing the classification in the compact and decorrelated feature space. To take this, each character class is partitioned into several clusters and each cluster density is estimated by a Gaussian distribution function in the PCA [13], FLD [2] and Laplacian [5]transformed space. The parameter estimation is performed by an iterative EM algorithm. The following section briefly presents the above three mentioned models for the sake of completeness. Detail description regarding the same can be seen in [7, 8, 11]

2.1 The PCA Mixture Model

In a mixture model [6], a character class can be partitioned into a number of clusters and its density function of the n-dimensional observed data $x = x_1, ..., x_n$ is represented by a linear combination of component density of the partitioned clusters as [7]

$$P(x) = \sum_{k=1}^{K} P(x|c_k, \theta_k)P(c_k), \qquad (1)$$

where $P(x|c_k, \theta_k)$, $P(c_k)$ and θ_k are the conditional density, prior probability and unknown model parameter of the kth cluster respectively. The conditional density function $P(x|c_k, \theta_k)$ is frequently modeled by a Gaussian function as

$$P(x|c_k, \theta_k) = \frac{1}{(2\pi)^{n/2}|\sum_k|^{1/2}} \times exp\left\{M\right\}, \qquad (2)$$

where M is $\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)$, μ_k and $\sum_k$ are the sample mean and covariance of the kth cluster, respectively. In order to reduce the dimensionality of feature space of the data, in this work we have used PCA technique. In PCA, a set of observed n-dimensional data vector $X = x_p, p \in 1, ..., N$ is reduced to a set of m-dimensional feature vector $S = s_p, p \in 1, ..., N$ by a transformation matrix W as

$$s_p = W^T(x_p - \delta[x]), \qquad (3)$$

where $W = (w_1, ..., w_m)$ and the vector w_d is the eigenvector corresponding to the d^{th} largest eigenvalue of the sample covariance matrix C such that $Cw_k = \lambda_k w_k$. Now we consider a PCA mixture model of the PCA transformed data $s = s_1, ..., s_m$, which combines the above two models (Eqs(1 and 3)) in a way that the mixture model is mapped onto the PCA transformed space as

$$P(s) = \sum_{k=1}^{K} P(s|c_k, \theta_k)P(c_k). \qquad (4)$$

then the conditional density function $P(s|c_k, \theta_k)$ of the PCA feature vectors in the kth cluster can be simplified as

$$P(s|c_k, \theta_k) = \frac{1}{(2\pi)^{m/2}|\Sigma_k^s|^{1/2}} exp\left\{-\frac{1}{2}s^T(\Sigma_k^s)^{-1}s\right\} \qquad (5)$$

$$= \prod_{j=1}^{m} \frac{1}{(2\pi)^{1/2}\lambda_{kj}^{1/4}} exp\left\{-\frac{s_j^2}{2\lambda_{kj}}\right\} \qquad (6)$$

where $\lambda_{k,1}, ..., \lambda_{k,m}$ are the m dominant eigenvalues of the feature covariance matrix $\sum_k^s$ in the k^{th} cluster.

To perform both the appropriate partitioning of the class and the estimation of model parameters of the partitioned clusters, the transformation matrix W is linear, and the log-likelihood function with respect to the transformed PCA feature vectors $S = s_1, ..., s_n$ can be represented as

$$\ell(S|\ominus) = \sum_{p=1}^{N} lnP(s_p) \qquad (7)$$

$$= \sum_{p=1}^{N} ln\left\{P(s_p|c_k, \theta_k)P(c_k)\right\}. \qquad (8)$$

This formulation will allow to determine both the appropriate partitioning of the class and the estimation of the model parameters simultaneously when the log-likelihood is maximized. We solve this log-likelihood maximization problem using EM iterative algorithm [3]. Each iteration consists of two-steps: an expectation step (E-step) followed by a maximization step (M-step). Each step is run for each mixture component. The EM algorithm starts its run after the parameters are initialized, and stops when the density undergoes no further changes.

(1) E-step: Given the feature data set X and the parameters $\ominus^{(t)}$ of the mixture model at the t-th iteration, we estimate the posterior distribution using

$$P(c|s, \ominus^{(t)}) = \frac{P(s|c, \ominus^{(t)})P(c)}{\sum_{c=1}^{K} P(s|c, \ominus^{(t)})P(c)} \qquad (9)$$

(2) M-step: Next, the new mean $\mu_k^{s(t+1)}$ of the kth cluster are obtained by the following equation:

$$\mu_k^{s(t+1)} = \frac{\sum_{p=1}^{N} P(c_p|s_p, \ominus^t)s_p}{\sum_{p=1}^{N} P(C_p|s_p, \ominus^{(t)})} \qquad (10)$$

The new variance parameters $\lambda_{kj}^{(t+1)}$ are obtained by selecting the largest m eigenvalues in the eigenvector computation as

$$\sum_{k}^{s(t+1)} w_j = \lambda_j^{(t+1)} w_j, \tag{11}$$

where the new covariance matrix $\lambda_j^{(t+1)}$ is computed by

$$\sum_{k}^{s(t+1)} = \frac{\sum_{p=1}^{N} P(c_p|s_p, \ominus^{(t)})(s_p - \mu - k^{s(t+1)})^T(S)}{\sum_{p=1}^{N} P(c_p|s_p, \ominus(t))} \tag{12}$$

where S is $s_p - \mu_k^{s(t+1)}$ The above two steps will be repeated until a stopping condition that three parameters will not be changed any further is satisfied.

2.2 The FLD mixture model

FLD has one transformation matrix among over all classes. This property degrades the performance of FLD because only one transformation matrix is not enough for the classification of complex data with many classes with high variations. To overcome this drawback, we use FLD mixture model that uses several transformation matrices among over all classes. Specifically we apply the PCA mixture model to the set of mean m_i of each character class with K mixture component, we obtain a cluster mean c_k, a transformation matrix T_k, and a diagonal matrix V_k with eigenvalues, where U_k is a diagonal matrix whose diagonal element are the eigenvalues λ_{kd} which is the dth largest eigenvalue of the covariance matrix. The probabilistic covariance matrix for the kth mixture component is $T_k U_k T_k^T$. Using this result, we can get the between-class scatter matrix and within-class scatter matrix for the kth mixture component as

$$S_{B_k} = T_k U_k T_k^T, \tag{13}$$

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

Based on Eqs 13 and 14, the transformation matrix W_k for the kth mixture component is determined so as to maximize the criterion function

$$SJ_k(U) = \frac{|U^T S_B U|}{|U^T S_{W_k} U|} \tag{15}$$

The columns of optimal W_k are the generalized eigenvectors w_{k_d} that correspond to the largest eigenvalues of $S_B u_{kd} = \lambda_{kd} S_{w_k} u_{kd}$.

2.3 The Laplacian Mixture Model

Before describing the concept of Laplacian Mixture model, we will explain the concept of laplacian (LPP) model for the sake of completness. Unlike from Principal Component Analysis (PCA), main objective of LPP [5] is to preserve the local structure of the input vector space by explicitly considering the manifold structure. Because it preserves the neighborhood information, its classification performance is much finer than other subspace approach like PCA. The generic problem of linear dimensionality reduction is the following. Let there be N number of input data points $(d_1, d_2, \cdots, d_N)$, which are in $\Re^M$. In the first step of this algorithm is to construct an adjacency graph **G** of N nodes, such that node i and j are linked if d_i and d_j are *close* with respect to each other in any of the following two conditions:

1. k-nearest neighbors: Nodes i and j are linked by an edge, if i is among k-nearest neighbors of j or vice-versa.

2. ϵ-neighbors: Nodes i and j are linked by an edge if $\|d_i - d_j\|^2 < \epsilon$, where $\|\cdot\|$ is the usual Euclidean norm.

Next step is to construct the weight matrix **Wt**, which is a sparse symmetric $N \times N$ matrix with weights Wt_{ij} if there is an edge between nodes i and j, and 0 if there is no edge. Two alternative criterion to construct the weight matrix:

1. Heat-Kernel: $Wt_{ij} = e^{\frac{-\|d_i - d_j\|^2}{t}}$, if i and j are linked.

2. $Wt_{ij} = 1$, iff nodes i and j are linked by an edge.

The objective function of LPP model is to solve the following generalized eigenvalue-eigenvector problem:

$$XLX^T a = \lambda XDX^T a \tag{16}$$

Where **D** is the diagonal matrix with entries as $D_{ii} = \sum_j w_{ji}$ and $L = D - W$ is the laplacian matrix.

The transformation matrix **W** is formed by arranging the eigenvectors of Eq.(16) ordered according to their eigenvalues, $\lambda_1 < \lambda_2, \ldots, < \lambda_l$. Thus, the feature vector y_i of input d_i is obtained as follows:

$$d_i \rightarrow y_i = A^T d_i \;\; \forall i = 1, 2, \ldots, N \tag{17}$$

Note: The XDX^T matrix is always singular because of high-dimensional nature of the image space. To alleviate this problem, PCA is used as the preprocessing step to reduce the dimensionality of the input vector space.

Since we obtained K number of PCA transformation matrices using PCA mixture model, a feature set for each mixture is obtained in the LPP mixture model. The objective function of the proposed method now becomes as follows:

$$X_k L X_k^T a_k = X_k D X_k^T a_k \qquad \forall k = 1, 2, 3, ..., K \tag{18}$$

Where X_k represents $p \times N$ feature matrix of training samples obtained after the transformation through k^{th} PCA mixture. The D and L matrix are obtained as mentioned above. The transformation matrices A_k = $(a_1^k, a_2^k, a_3^k ....., a_p^k)$ of LPP mixture model are formed by arranging p eigenvectors of k^{th} LPP mixture corresponding to p largest eigenvalues $\lambda_1^k < \lambda_2^k < \lambda_3^k < ...., \lambda_p^k \quad \forall k = 1, 2, 3, ..., K$. Using the A_ks, features for a training sample x can be obtained as follows:

$$f_i^k = A_k^T x_i \quad \forall i = 1, 2, ..., N \quad and \;\; \forall k = 1, 2, ..., K \tag{19}$$

Since there are K mixtures, K number of features are obtained for a unknown sample I. To combine K classification results of I from all the mixtures, a distance matrix is constructed and denoted by $D(I) = (d_{ij})_{NK}$ where d_{ij} is set to 1 if I is matched to i^{th} training sample after transformation through j^{th} mixture, else it is set to 0. Consequently, the total confidence value that the sample I belongs to the i^{th} class is $TC_i(I) = \sum_{j=1}^{K} d_{ij} \forall i = 1, 2, ..., N$. Finally, identity of the test sample I is computed as follows:

$$Identity(I) = argmax_i(TC(I)) 1 \leq i \leq N \tag{20}$$

3 Experiment Results and Comparative Analysis

This section presents the results carried out on handwritten Kannada characters. The experiment was conducted on the database comprising unconstrained handwritten isolated Kannada characters of 5,000 samples [9]. The dataset holds 50 classes where each class intern contains 100 samples written by individual writers. We train the system by varying the training sample number by 25, 50, and 75 and remaining samples of each character class are used during testing. Our preliminary experiments suggest that mixture of four Gaussians could be an optimal choice for competitive results. Hence, all our experiments were carried out on four mixture of Gaussains. Figure 1 shows the performance accuracy of different subspace and mixture models methods. We also compared the results of standard PCA, FLD and Laplacian methods with their Mixture models. From the Figure 1 it is noticed that the performance of the GMM combined with FLD achieved better results compared to other methods. The next best method is GMM combined with PCA.

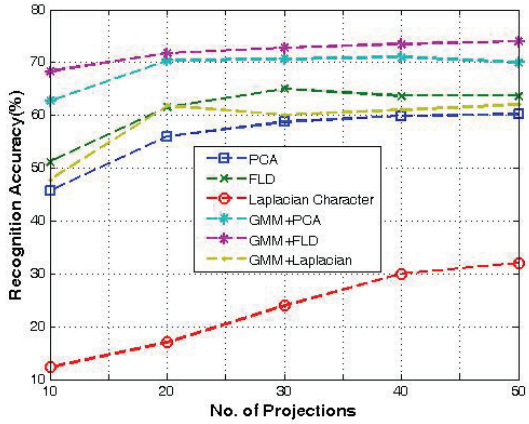

Figure 1. Recognition Accuracy of Subspace and Subspace Mixture Models Methods

3.1 Experiment Analysis

The following are the observations made from the experiment:

- The performance of the standard FLD method is better when compared to other standard subspace methods.
- The application of subspace mixture models showed improvement in recognition accuracy.
- Though the idea of GMM + Laplacian has showed better performance in face recognition, the same idea has not worked out in character recognition.
- GMM+FLD has outperformed all the other existing methods in terms of recognition accuracy.

Unlike the PCA method, the objective of FLD is to maximize the between-class measure while minimizing the within-class measure and the above observation on FLD technique has made clear. As an alternative to the PCA, the Laplacian characters optimally preserves the neighborhood structure of the dataset. But from the above observation it is noticed that the performance of the Laplacian method has not shown better results. It is also worth to note that the performance of the same

Laplacian is improved a lot when it is combined with the mixture models. From the above observation, we can notice that the application of mixture models have greater advantages in improving the recognition accuracy.

4 Discussion and Conclusion

Though there are vast number of feature extraction techniques available in literature for handwriting recognition character recognition, feature space transformation methods have gained lot of importance in pattern recognition research. Feature representation of such patterns are used for reducing the dimensionality of features and some of them can also improve the classification accuracy [4]. In addition, from the survey of literature it is quite evident that subspace based algorithms for recognition are widely admired and adapted in current object/face recognition research. Subspace methods can give superior performance for (i) Font-independent OCRs (ii) Noisy characters (iii) Degraded and Broken characters (iv) Easy adaptation across languages (v) Unconstrained handwritten characters.

Due to above facts, in this work we explored, analyzed and proposed the concept of subspace mixture models for handwritten Kannada character recognition. Some of the following exploration are still needed, which may be the far beyond the scope of this paper.

- The choice of exact mixture of Gaussians is highly subjective in nature, study on the same is interesting and challenging.
- Recently affect of different distance measure techniques such as Angle, Correlation has shown improvements in character recognition [10]
- Method based on Laplcian has shown greater improvement in face recognition, it is now required to analyze the technique and at the same time it is also necessary to analyze the character dataset.

References

[1] Aleix.M.Martinez and Avinash C Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:228–233, 2001.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] M. Cheriet, N. Kharma, C-L. Liu, and C. Y. Suen. *Character Recognition Systems*. Wiley, 2007.

[5] Xiaofei He, Shuicheng Yan, Yuxiao Hu, and Partha Niyogi. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:328–340, 2005.

[6] M. Jordon and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *neural Computing*, 6:181–214, 1994.

[7] H.C. Kim, D. Kim, and S.Y. Bang. Face recognition using the mixture-of-eigenfaces method. *Pattern Recognition Letters*, 23:1549–1558, 2002.

[8] H.C. Kim, D. Kim, and S.Y. Bang. Face recognition using LDA mixture model. *Pattern Recognition Letters*, 24:2815–2821, 2003.

[9] C. Naveena and V. N. Manjunath Aradhya. An impact of ridgelet transform in handwritten recognition: A study on very large dataset of kannada script. In *: proceedings of International Conference on World Congress on Information and Communication Technologies (WICT-2011)*, pages 622 – 625, 2011.

[10] S. K. Niranjan, V. Kumar, G. H. Kumar, and V. N. M. Aradhya. FLD based unconstrained handwritten kannada character recognition. In *: Proceedings of IEEE International Conference on Future Generation Communication and Networking Symposia*, pages 7–10, 2008.

[11] Noushath.S, Ashok Rao, and Hemantha Kumar.G. Mixture-of-laplacianfaces and its application to face recognition. In *2nd International Conference on Pattern Recognition and Machine Intelligence (PReMI-07)*, pages 568–575, 2007.

[12] R.A.Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:7186, 1991.