PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition 38 (2005) 799–811

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Face recognition based on multi-class mapping of Fisher scores

Ling Chen[a,*], Hong Man[a], Ara V. Nefian[b]

[a]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030, USA
[b]Microprocessor Research Labs, Intel Corporation, SC12-303, 2200 Mission College Blvd., Santa Clara, CA 95052-8119, USA

Received 9 August 2004

Abstract

A new *hidden Markov model* (HMM) based feature generation scheme is proposed for *face recognition* (FR) in this paper. In this scheme, HMM method is used to model classes of face images. A set of Fisher scores is calculated through partial derivative analysis of the parameters estimated in each HMM. These Fisher scores are further combined with some traditional features such as log-likelihood and appearance based features to form feature vectors that exploit the strengths of both local and holistic features of human face. *Linear discriminant analysis* (LDA) is then applied to analyze these feature vectors for FR. Performance improvements are observed over stand-alone HMM method and Fisher face method which uses appearance based feature vectors. A further study reveals that, by reducing the number of models involved in the training and testing stages of LDA, the proposed feature generation scheme can maintain very high discriminative power at much lower computational complexity comparing to the traditional HMM based FR system. Experimental results on a public available face database are provided to demonstrate the viability of this scheme.

Keywords: Face recognition; Hidden Markov model; Fisher score; Linear discriminant analysis

1. Introduction

One of the most popular appearance based methods [1–3] for *face recognition* (FR) developed in recent years is the Fisherface method. The Fisherface method performs *linear discriminant analysis* (LDA) of feature vectors obtained as one-dimensional representation of a face image and retrieves the identity of person based on the nearest-neighbor classification criterion in the LDA space. This method is insensitive to large variation in lighting direction and facial expression [2].

Meanwhile, statistical model based methods such as *hidden Markov model* (HMM) have also been proposed for FR problems [4–8]. This method uses HMM to describe the statistical distribution of observation vector sequences which are generated from small sub-image blocks of face image. Classification is usually based on Bayesian decision rule, e.g., maximum a posteriori (MAP) criterion. Comparing with appearance based methods, HMM methods focus mainly on local characteristics of human faces. These methods have the flexibility to incorporate information from different instances of faces at different scales and orientations [5]. However, in these existing statistical model based methods, only the calculated *likelihood* of a particular observation on each established model is used as the measure of closeness of the observation towards the corresponding class.

In this work, we present a new feature vector generation scheme from HMMs. The scheme generates feature vectors which represent the influence of the model parameters of several competing HMMs on the generation of a particular observation vector sequence. Similar methods were proposed and used in biosequence analysis, speech

* Corresponding author. Tel.: +1 201 216 8019.
E-mail addresses: lchen@stevens.edu (L. Chen), hman@stevens.edu (H. Man), ara.nefian@intel.com (A.V. Nefian).

0031-3203/$30.00 © 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.11.003

recognition, and speaker identification [9,14,15]. Unlike previous schemes which are inherently two-class problem oriented, the proposed scheme in this work is multi-class problem oriented and the resulting feature vectors appear to be more effective. We also explore the strengths of both Fisherface method and HMM method by combining appearance based features (as seen in Fisherface approaches) and statistical model based features together to form new feature vectors, which may have greater discriminative power over those used separately. Furthermore, in a typical multi-class HMM method, one HMM is established for each class of object (e.g. faces of one person), and a test observation is compared to all the available classes in order to determine its identity. In this work we attempt to reduce the number of HMMs involved in this process and manage to achieve a comparable recognition performance as when all HMMs are used. Apparently the model reduction translates to a significant computational advantage, which effectively improves the scalability of such statistical model based methods.

The paper is organized as follows: Section 2 discusses model based methods for pattern recognition and the concept of *Fisher score* used in generation of statistical model based feature vectors; Section 3 introduces multi-class mapping to generalize previous schemes for multi-class classification problems; Section 4 presents the computation of Fisher scores in regard to our specific statistical model structure; Section 5 details the combination scheme for feature vector generation; Section 6 implements LDA on feature vectors and summarize the proposed system structure; Section 7 discusses the choice of the sampling scheme and HMM model structure in experiments, including experimental results and discussion. The paper is concluded with a summary and possible future research directions.

2. Model based methods and Fisher score

A common scenario of using model based methods for pattern recognition is that all training and testing observations are assumed to follow a predefined form of statistical distribution. The parameter estimation of the statistical model of each class, $\hat{\boldsymbol{\theta}}_i$, are found by maximizing the likelihoods of training observations labeled for that class.[1] The pdf of an observation $\mathbf{O}$ based on the estimated model parameters is $f(\mathbf{O}|\hat{\boldsymbol{\theta}}_i)$. For a $N_{\rm c}$-class problem, we have $\{\hat{\boldsymbol{\theta}}_i|\hat{\boldsymbol{\theta}}_i \in \boldsymbol{\Theta},\ i=1,\ldots,N_{\rm c}\}$, where $\boldsymbol{\Theta}$ is the space of model parameters. If the form of the statistical distribution and the parameters estimated are appropriate and precise enough in describing the distribution pattern of the training observations, the a posteriori of a testing observation on the trained model with the same class label should be higher than those from other trained models. The MAP criterion is then applied for classification. If the priors of all classes are equal, which is commonly assumed in FR problem, the MAP criterion equals the *maximum likelihood* (ML) criterion:

$$i' = \arg \max_{1 \leqslant i \leqslant N_{\rm c}} \log f(\mathbf{O}_{\rm q}|\hat{\boldsymbol{\theta}}_i), \tag{1}$$

where $\mathbf{O}_{\rm q}$ is a query (testing) observation. Therefore the model structure definition and the parameter estimation have heavy influence on the effectiveness of this method. The most successful HMM-based methods for FR include the 1-D HMM [4], the pseudo-2-D HMM [4,5,7], the low complexity 2-D HMM [5,8], and more recently, the embedded Bayesian networks [6]. In all these variations of HMM structures, the likelihood score remains the only measurement of the match between the observation and the model.

Recently *Fisher kernel method* was proposed by Jaakkola and Haussler [9] for protein sequence analysis. This approach is theoretically justified in the framework of maximum entropy discrimination [10]. It can be considered as an approximation of the mutual information kernel [11], or as a method of constructing a posterior probability model for the class labels [12]. The Fisher kernel method calculates the difference in generative processes between observations rather than the likelihood difference. The difference in the generative processes between observations is represented by the difference of the vectors composed of *Fisher scores* [9]. To elaborate, consider a class of statistical models $f(\mathbf{O}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Under certain conditions, this class of statistical models defines a Riemannian manifold [13]. The tangent space at point $\boldsymbol{\theta}$ of the manifold is composed of the tangent vectors of smooth curves passing through $\boldsymbol{\theta}$. Fisher scores are the gradients of the log-likelihood of an observation with respect to the parameters of a statistical model. That is, given the observation $\mathbf{O}$ and the model parameters $\boldsymbol{\theta}=\{\theta_i|i=1,2,\ldots,P\}$, Fisher score vector of the observation $\mathbf{O}$ with regard to the given model $\boldsymbol{\theta}$ is defined as

$$\nabla_{\boldsymbol{\theta}}(\mathbf{O}) = \left[\frac{\partial \log f(\mathbf{O}|\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \log f(\mathbf{O}|\boldsymbol{\theta})}{\partial\theta_P}\right]^{\rm T}. \tag{2}$$

The P-dimensional space spanned by Fisher score vector at point $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$ is called the *l*-representation of the tangent space. Then the geometrical meaning of Fisher score vector can be interpreted as the tangent vector at point $\boldsymbol{\theta}$. In a Fisher score vector, the physical meaning of the value of each component $\nabla_{\theta_i}(\mathbf{O})$ can be interpreted as the significance of the influence of a particular model parameter in the generation of the observation. Obviously, the value of the Fisher score vector is influenced by the observation $\mathbf{O}$ and the model parameters $\boldsymbol{\theta}$. When the value of $\boldsymbol{\theta}$ is fixed, the similarity between two observations $\mathbf{O}_i$ and $\mathbf{O}_j$ can be calculated by an inner product between two corresponding Fisher score vectors $\nabla_{\boldsymbol{\theta}}(\mathbf{O}_i)$ and $\nabla_{\boldsymbol{\theta}}(\mathbf{O}_j)$, scalded by a local metric $I = E_{\mathbf{O}}(\nabla_{\boldsymbol{\theta}}(\mathbf{O})^{\rm T}\nabla_{\boldsymbol{\theta}}(\mathbf{O}))$, which is called *Fisher information matrix*. That is, the similarity between two observations given the model parameters is calculated as

$$K(\mathbf{O}_i, \mathbf{O}_j) = \nabla_{\boldsymbol{\theta}}(\mathbf{O}_i)^{\rm T} I^{-1} \nabla_{\boldsymbol{\theta}}(\mathbf{O}_j) \tag{3}$$

[1] Definitions of symbols can be found in Table 1.

which is called *Fisher kernel* [9]. In a binary classification problem, Fisher kernel method begins with the training of an HMM by using positive observation sequences from a given class. This HMM is used to map each positive or negative observation sequence $\mathbf{O}_i$ into a fixed length Fisher score vector. Fisher score vectors from positive and negative observation sequences are used to train a *support vector machine* (SVM) with the Fisher kernel function. Each query observation sequence $\mathbf{O}$ is mapped into a query Fisher score vector and classification is carried out via the trained SVM. The resulting discriminant function is:

$$\mathscr{L}(\mathbf{O}) = \sum_{i:\mathbf{O}_i \in H_1} \alpha_i K(\mathbf{O}, \mathbf{O}_i) - \sum_{i:\mathbf{O}_i \in H_0} \alpha_i K(\mathbf{O}, \mathbf{O}_i), \quad (4)$$

where the *Lagrange multipliers* α_i are estimated by positive and negative examples $\mathbf{O}_i$; H_1, H_0 represent observations of the two competing classes. In fact, besides Fisher scores obtained from the gradients (the first-order partial derivatives) of the log-likelihood, the zeroth-order partial derivative, i.e., the log-likelihood itself can be used independently or jointly with other features to form fixed dimensional feature vectors. This is discussed in the following section. Moreover, higher order derivatives can also be incorporated in the process of finding discriminative information for classification [14].

3. Multi-class mapping

In the frameworks proposed by Jaakkola and Haussler [9] and Fine et al. [15], Fisher scores are computed from the log-likelihood of a single statistical model representing one class [15] or both competing classes [9]. However, if two statistical models are established, and each representing one of the two competing classes, the feature vectors based on the Fisher scores from these two models may carry more discriminative information. Smith and Gales [14] proposed a method that uses the log ratio of two likelihoods calculated from two competing models for Fisher score generation. This scheme was justified for providing a solution to the wrap-around phenomenon and the realization of the optimal decision rule [14]. Although these approaches can be used to handle multi-class problems, the feature vectors in these approaches are designed intrinsically for binary classifiers such as the SVM for that the amount of statistical models involved in the computation of Fisher scores is at most two that corresponds to two competing classes.

Our approach treats the calculation of Fisher scores as a process of mapping (or projecting) the observations towards the derivative space of a particular statistical model. From this perspective, the Fisher scores used in the previous schemes can be thought as the results of single- or two-class mapping processes. It is expected that, through the mapping of the observations which share the same class membership, towards the derivative space of an arbitrary statistical model coming from one of the competing classes, their resulting Fisher scores will cluster together under a predefined similarity criterion, regardless which class this statistical model represents. Therefore the distribution patterns of the Fisher scores from the observations should be highly related in the derivative space. For binary or multi-class classification (such as FR), the feature vectors composed of Fisher scores extracted from the models of more than one competing classes are likely to carry more discriminative information than those from the single model. We call this procedure of mapping a particular observation towards the derivative spaces of multiple competing statistical models as multi-class mapping. By introducing multi-class mapping, not only Fisher scores (the first-order derivatives of log-likelihood on model parameters), but also log-likelihood itself (the zeroth-order derivatives of log-likelihood on model parameters) can be used to form feature vectors (see Fig. 1).[2]

4. Fisher scores for diagonal Gaussian HMM

The computation of the Fisher scores depends on the structure of the statistical model. The statistical model we choose for FR is a one-dimensional ergodic HMM which assumes the observation distribution density as Gaussian with diagonal covariance matrix (it will be discussed in Section 7.2.2). For a Gaussian HMM, the parameters needed to represent the model include three components, i.e., the state transition distribution A, the observation probability distribution B, (Table 1) and the initial states distribution π [16]. In order to completely represent the gradients, all three components should be considered. For each positive or negative observation sequence, the gradients of its log-likelihood with respect to the parameters of an HMM are defined as follows:

- The gradients with respect to the state transition distribution: $\nabla_{a_{\tilde{s}''|\tilde{s}'}}(\mathbf{O})$, for $1 \leqslant \tilde{s}',\ \tilde{s}'' \leqslant S$.
- The gradients with respect to the Gaussian observation probability distribution: $\nabla_{\mu_{\tilde{s},i}}(\mathbf{O})$ and $\nabla_{\sigma_{\tilde{s},i}}(\mathbf{O})$, for $1 \leqslant \tilde{s} \leqslant S,\ 1 \leqslant i \leqslant D$.
- The gradients with respect to the initial states distribution: $\nabla_{\pi_{\tilde{s}}}(\mathbf{O})$, for $1 \leqslant \tilde{s} \leqslant S$.

The above defined gradients are calculated as follows:

$$\nabla_{a_{\tilde{s}''|\tilde{s}'}}(\mathbf{O}) = \sum_{t=1}^{T} \xi_t(\tilde{s}', \tilde{s}'') \frac{1}{a_{\tilde{s}''|\tilde{s}'}},$$
$$1 \leqslant \tilde{s}', \tilde{s}'' \leqslant S, \quad (5)$$

[2] The observations used in this figure come from Georgia Tech Face Database (GTFD) which will be addressed in Section 7.1.

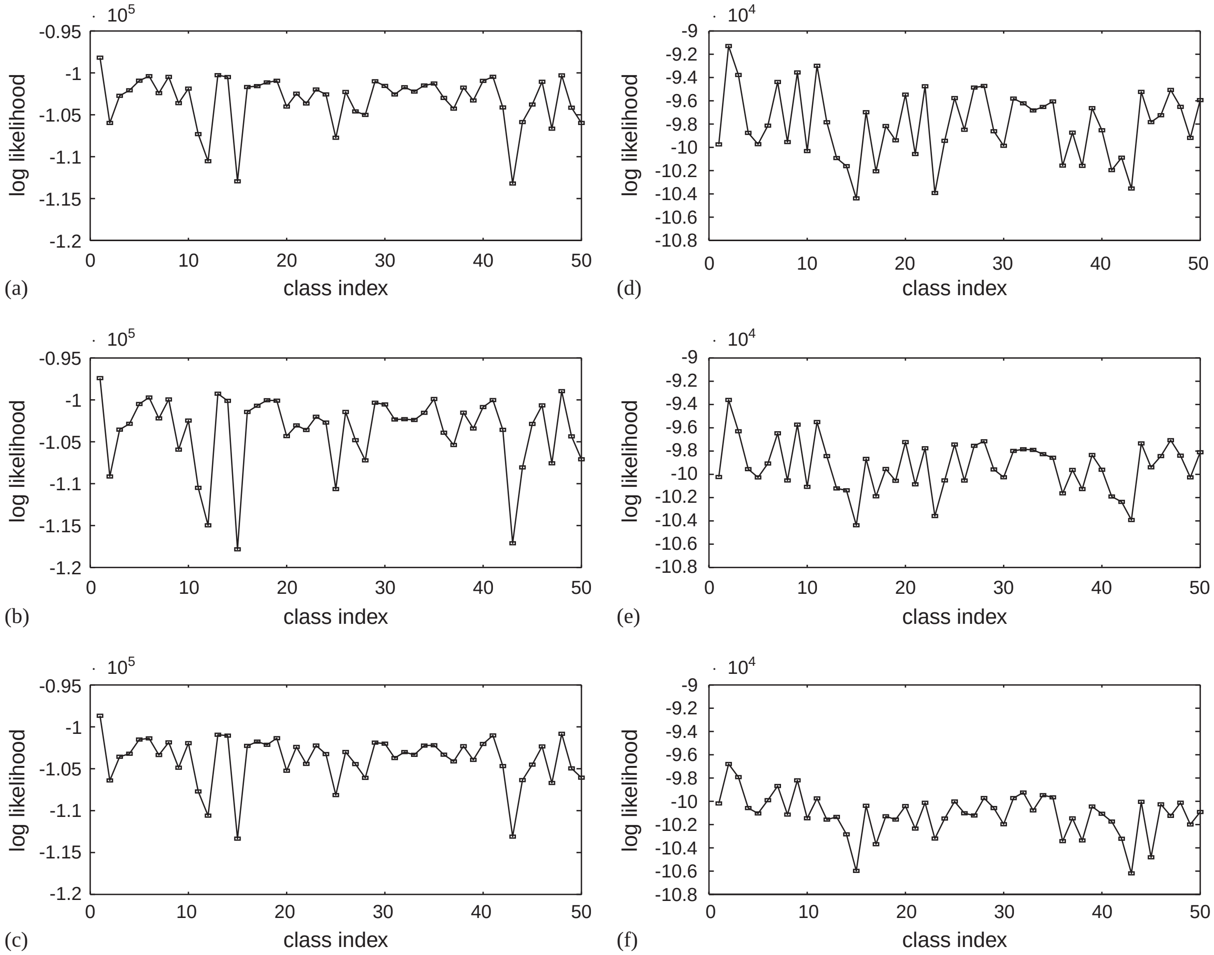

Fig. 1. Example of multi-class mapping. (a)–(c) are zeroth-order multi-class mappings of three randomly selected observations of subject 1 with respect to the parameters of log-likelihoods of 50 HMMs trained for all 50 subjects in Georgia Tech Face Database (GTFD). (d)–(f) are zeroth-order multi-class mappings of three randomly selected observations of subject 2 with respect to the parameters of log-likelihoods of 50 HMMs trained for all 50 subjects in GTFD. From (a)–(c) or (d)–(f), it can be seen that the mapping values of the multi-class mapping of a particular observation are highly different from each other. Whereas the overall distribution patterns of multi-class mapping value among observations which share the same class membership, are similar. Also from comparing (a)–(c) with (d)–(f), inter-class difference is obvious to be noticed.

where

$$\xi_t(\tilde{s}', \tilde{s}'') = P(s_t = \tilde{s}'', s_{t-1} = \tilde{s}' | \mathbf{O}, \lambda)$$

which is the probability of being in state $\tilde{s}'$ at time $t-1$, and in state $\tilde{s}''$ at time t, given the observation sequence $\mathbf{O}$, and the model λ. This probability can be obtained through forward–backward procedure [16]

$$\nabla_{\mu_{\tilde{s},i}}(\mathbf{O}) = \sum_{t=0}^{T} \gamma_t(\tilde{s}) \frac{o_{t,i} - \mu_{\tilde{s},i}}{\sigma_{\tilde{s},i}^2}, \quad 1 \leqslant \tilde{s} \leqslant S, \quad 1 \leqslant i \leqslant D, \tag{6}$$

where

$$\gamma_t(\tilde{s}) = P(s_t = \tilde{s} | \mathbf{O}, \lambda)$$

which is the probability of being in state $\tilde{s}$ at time t, given the observation sequence $\mathbf{O}$, and the model λ. Again, this probability can be obtained through forward–backward procedure [16].

$$\nabla_{\sigma_{\tilde{s},i}}(\mathbf{O}) = \sum_{t=0}^{T} \gamma_t(\tilde{s}) \left[\frac{(o_{t,i} - \mu_{\tilde{s},i})^2}{\sigma_{\tilde{s},i}^3} - \frac{1}{\sigma_{\tilde{s},i}} \right], \quad 1 \leqslant \tilde{s} \leqslant S, \quad 1 \leqslant i \leqslant D, \tag{7}$$

$$\nabla_{\pi_{\tilde{s}}}(\mathbf{O}) = \frac{\gamma_0(\tilde{s})}{\pi_{\tilde{s}}}, \quad 1 \leqslant \tilde{s} \leqslant S. \tag{8}$$

A set of typical examples of the Fisher scores are shown in Fig. 2.

Table 1
Notation conventions for a Gaussian HMM

t	Step t of observation: $t = 0, 1, \ldots, T$
s	States of HMM: $s = 1, 2, \ldots, S$
D	Dimension of observation vectors
$\mathbf{o}_t$	Observation vector at step t : $\mathbf{o}_t \in \mathbb{R}^D$
$\mathbf{O}$	Observation sequence: $\mathbf{O} = (\mathbf{o}_0, \mathbf{o}_1, \ldots, \mathbf{o}_T)$
$A = \{a_{i,j}\}$	The state transition probability distribution, $a_{i,j} = a_{\tilde{s}''\|\tilde{s}'}$ where $\tilde{s}'' = j, \tilde{s}' = i$ and $1 \leqslant i, j \leqslant S$
$\mathcal{N}_s(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$	Gaussian distribution of state s with mean vector $\boldsymbol{\mu}_s$ and diagonal covariance vector $\boldsymbol{\sigma}_s$
$B = \{b(\mathbf{o}\|s)\}$	The observation probability distribution, $b(\mathbf{o}\|s) \sim \mathcal{N}_s(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$
$\pi = \{\pi_s\}$	The initial state distribution
$\lambda = (A, B, \pi)$	Model parameters for Gaussian HMM
$\boldsymbol{\theta}$	Model parameters for a generic statistical model

5. Combination schemes of Fisher scores for feature generation

Based on the structure of the statistical model chosen in our system, we have four types of Fisher scores available to be used to form feature vectors. They are $\nabla_{a_{\tilde{s}''|\tilde{s}'}}$, $\nabla_{\mu_{\tilde{s},i}}$, $\nabla_{\sigma_{\tilde{s},i}}$, and $\nabla_{\pi_{\tilde{s}}}$. The formation of feature vectors depends on the types of Fisher scores chosen in the mapping procedure and the mapping procedure itself (i.e., single- or multi-class mapping). In addition to Fisher scores, the feature vectors can include other features such as the multi-class mapping of the zeroth-order partial derivative and appearance based features:

- *The multi-class mapping of the zeroth-order partial derivative*, *i.e.*, *log-likelihood*: From Fig. 1, we can see that multi-class mapping of log-likelihood displays strong intra-class relationship and inter-class difference. This suggests it should be exploited in the formation of feature vectors.
- *Appearance based features*, *i.e.*, *vectorization of face image*: Appearance based features are commonly used in LDA for FR. They are fundamentally different from statistical model based features, therefore they may have complemental effect. By combining appearance and statistical based features, it is possible to obtain feature vectors with increased discriminative power.

Combinations of various features discussed above are summarized in Table 2, where *holi* stands for holistic (appearance based) features. The dimensions of the feature vectors for all categories are listed in Table 3, where N_c denotes the number of classes in the database, and R and C are the number of rows and number of columns of face images, respectively (refer to Table 1 for definitions of other symbols).

As previously mentioned, the similarity between two Fisher score vectors can be calculated by the Fisher kernel (refer Eq. (3)), which is the scaled inner product of two Fisher score vectors. Because that $E_{\mathbf{O}}(\nabla_{\boldsymbol{\theta}_i}(\mathbf{O})) = 0$, where $i = 1, \ldots, P$ [13], the scaling factor I is effectively the covariance matrix, which is symmetric and positive definite, of Fisher score vectors. Suppose $\boldsymbol{\Phi}$ is an $P \times P$ matrix, consisting of eigenvectors of I as

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_P]$$

and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of I as

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_i & & 0 \\ & \ddots & \\ 0 & & \lambda_P \end{bmatrix},$$

then by rewriting the definition of Fisher kernel as

$$K(\mathbf{O}_i, \mathbf{O}_j) = [(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}} \nabla_{\boldsymbol{\theta}}(\mathbf{O}_i)]^{\mathrm{T}} \times [(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}} \nabla_{\boldsymbol{\theta}}(\mathbf{O}_j)], \tag{9}$$

it is clear that *whitening transformation* [18] is applied to Fisher score vectors before the inner product is computed. After combining different categories of features (such as combining Fisher scores with appearance based features) to generate feature vectors, the dynamic ranges of different components of the generated feature vectors may vary significantly. Whereas large dynamic range does not necessarily mean greater discriminative power. Then whitening transformation is also needed to decorrelate different components and normalize their dynamic ranges before comparing the similarity between them. Unfortunately, because of the limited training data, the covariance matrix of feature vectors can hardly be obtained. A practical solution to this problem is assuming independency among different components of feature vectors and individually normalizing each component by its mean and variance.

6. LDA on feature vectors generated from multi-class and single-class mapping

LDA is a subspace analysis method that projects high-dimensional data to a lower dimensional subspace which maximize a predefined class separability criterion. Let the training samples (in our case the feature vectors generated under a particular combination scheme listed in Table 2) be $\mathbf{x} = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{N_i}^i | i = 1, \ldots, N_\mathrm{c}\}$, where $\mathbf{x}_1^i, \ldots, \mathbf{x}_{N_i}^i$ denotes N_i training samples of class i. In the training stage of LDA, training samples are used to find the *optimal projection matrix*:

$$\mathbf{W}_{\mathrm{opt}} = \mathbf{W}_{\mathrm{lda}} \mathbf{W}_{\mathrm{pca}} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D_{\mathrm{lda}}}],$$

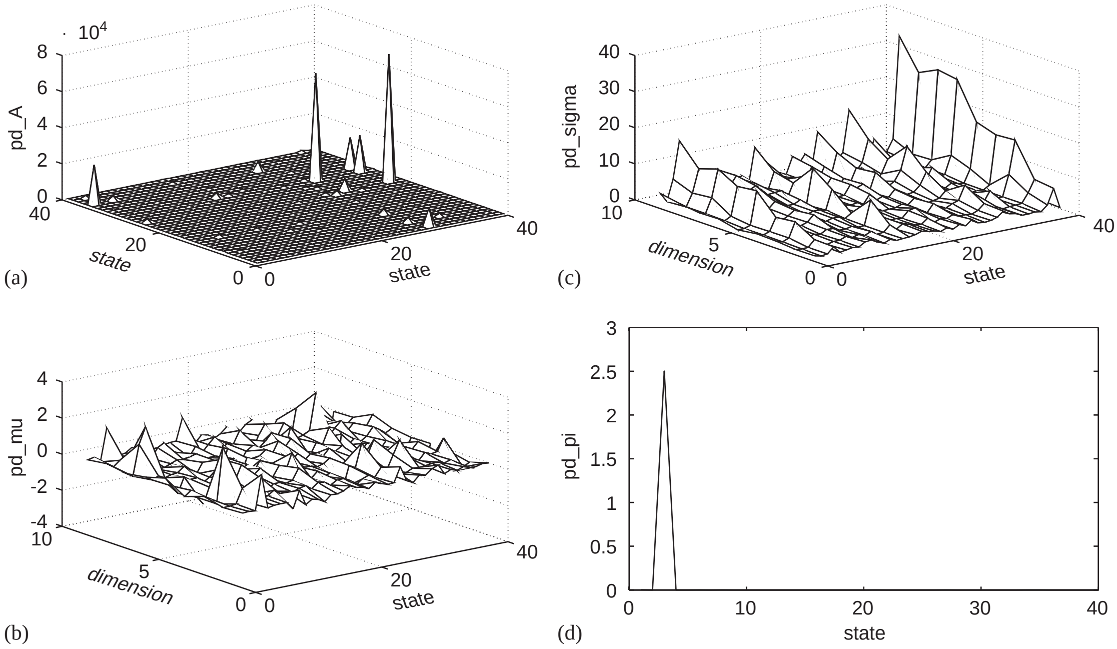

Fig. 2. Example of Fisher scores. (a)–(d) are partial derivatives (pd) of A, μ, σ, π, respectively.

Table 2
Feature generation schemes for LDA

Multi-class mappings	
m_loglik	Multi-class mapping of log-likelihood
m_a	Multi-class mapping of *A*
m_mu	Multi-class mapping of mu
m_sigma	Multi-class mapping of sigma
m_pi	Multi-class mapping of pi
m_mu_sigma	Multi-class mapping of mu and sigma
m_loglik_mu_sigma	Multi-class mapping of log-likelihood, mu, and sigma
Multi-class mappings with holistic (*appearance based*) *features*	
m_loglik_holi	Multi-class mapping of log-likelihood and holistic (appearance based) features
m_mu_holi	Multi-class mapping of mu and holistic features
m_sigma_holi	Multi-class mapping of sigma and holistic features
m_pi_holi	Multi-class mapping of pi and holistic features
m_mu_sigma_holi	Multi-class mapping of mu, sigma and holistic features
m_loglik_mu_sigma_holi	Multi-class mapping of mu, sigma, log-likelihood and holistic features
Single-class mappings	
s_mu	Single-class mapping of mu
s_sigma	Single-class mapping of sigma
s_mu_sigma	Single-class mapping of mu and sigma

where

$$\mathbf{W}_{\mathrm{pca}} = \arg\max_{\mathbf{W}} |\mathbf{W}^{\mathrm{T}}\mathbf{S}_{\mathrm{m}}\mathbf{W}|,$$

$$\mathbf{W}_{\mathrm{lda}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^{\mathrm{T}}\mathbf{W}_{\mathrm{pca}}^{\mathrm{T}}\mathbf{S}_{\mathrm{b}}\mathbf{W}_{\mathrm{pca}}\mathbf{W}|}{|\mathbf{W}^{\mathrm{T}}\mathbf{W}_{\mathrm{pca}}^{\mathrm{T}}\mathbf{S}_{\mathrm{w}}\mathbf{W}_{\mathrm{pca}}\mathbf{W}|}$$

and $\mathbf{S}_{\mathrm{m}}$, $\mathbf{S}_{\mathrm{b}}$, $\mathbf{S}_{\mathrm{w}}$ are the *mixture scatter matrix*, the *between-class scatter matrix* and the *within-class scatter matrix*, respectively [18]. To avoid the singularity problem of $\mathbf{S}_{\mathrm{w}}$ due to the high dimensions of the training samples, we first project the training samples to the subspace spanned by principal components of the mixture scatter matrix $\mathbf{S}_{\mathrm{m}}$. Let $\mathbf{y} = \{\mathbf{y}_1^i, \ldots, \mathbf{y}_{N_i}^i | i = 1, \ldots, N_{\mathrm{c}}\}$ and $\tilde{\mathbf{y}} = \{\tilde{\mathbf{y}}_1^i, \ldots, \tilde{\mathbf{y}}_{\tilde{N}_i}^i | i = 1, \ldots, N_{\mathrm{c}}\}$ be the projections of the training and testing samples on the optimal projection matrix $\mathbf{W}_{\mathrm{opt}}$, respectively. For any $\tilde{\mathbf{y}}_l^i$, where $i = 1, \ldots, N_{\mathrm{c}}$ and $l = 1, \ldots, \tilde{N}_i$, we define the distance from testing image l in class i towards the class j as

$$d(\tilde{\mathbf{y}}_l^i, j) = \min_{l'} \{\|\tilde{\mathbf{y}}_l^i - \mathbf{y}_{l'}^j\|_2\}, \quad l' = 1, \ldots, N_j,$$

where $\|\mathbf{y}\|_2$ is the L_2-norm of the vector $\mathbf{y}$. Then the identity of the testing image l in class i is assigned as j' when

$$j' = \arg\min_j \{d(\tilde{\mathbf{y}}_l^i, j)\}, \quad j = 1, \ldots, N_{\mathrm{c}}.$$

To summarize, Fig. 3 describes the use of LDA in the overall FR system presented in this paper.

To implement LDA method on feature vectors generated from multi-class mapping when all competing models are involved in the mapping process, the training

Table 3
Dimension of feature vectors

Scheme	Dimension	Scheme	Dimension
m_loglik	N_C	m_loglik_holi	$N_C + RC$
m_mu	SDN_C	m_mu_holi	$SDN_C + RC$
m_sigma	SDN_C	m_sigma_holi	$SDN_C + RC$
m_pi	SN_C	m_pi_holi	$SN_C + RC$
m_mu_sigma	$2SDN_C$	m_mu_sigma_holi	$2SDN_C + RC$
m_loglik_mu_sigma	$2SDN_C + N_C$	m_loglik_mu_sigma_holi	$2SDN_C + N_C + RC$
m_a	S^2N_C	s_mu	SD
s_sigma	SD	s_mu_sigma	$2SD$

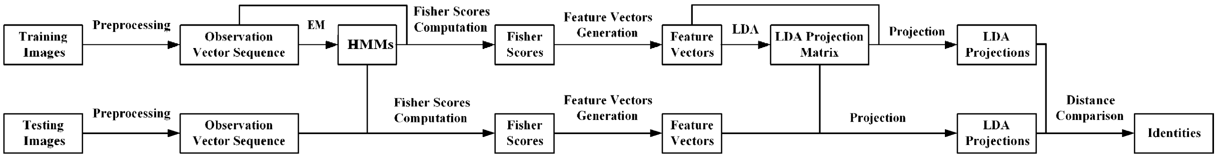

Fig. 3. Thumbnail of the proposed system.

samples $\{\mathbf{x}_1^i, \ldots, \mathbf{x}_{N_i}^i | i = 1, \ldots, N_C\}$ and the testing samples $\{\tilde{\mathbf{x}}_1^i, \ldots, \tilde{\mathbf{x}}_{\tilde{N}_i}^i | i = 1, \ldots, N_C\}$ are composed of Fisher scores of all competing models under a particular combination scheme (see Fig. 4). The computation of this approach can be quite intensive, especially when the total number of all classes N_C is very large. One solution to this problem is to reduce the number of models used in the multi-class mapping. If $N' \ll N_C$ models are used, the saving of computation and memory requirement can be very significant. In this work we examine an approach in which N' HMMs are selected randomly from the total of N_C HMMs. Once the N' HMMs have been chosen, the process of feature vector generation is the same as depicted in Fig. 4. With this approach there is some randomness being introduced in the performance of the whole FR system since different selections of HMMs can result in different recognition rates (RR).

The computational complexity of $\nabla_{a_{\tilde{s}''|\tilde{s}'}}$, and $\nabla_{\pi_{\tilde{s}}}$ are $O(S^4T)$, and $O(D+S)$ (refer Table 1 for the definitions of S, T, and D). The computational complexity of $\nabla_{\mu_{\tilde{s},i}}$ and $\nabla_{\sigma_{\tilde{s},i}}$ are $O(S^2T + STD)$. Because usually D is much smaller than S, then it can be simplified as $O(S^2T)$. Then to find out the identity of a test image, for multi-class mapping with feature vector composed of $\nabla_{a_{\tilde{s}''|\tilde{s}'}}$, $\nabla_{\mu_{\tilde{s},i}}$, $\nabla_{\sigma_{\tilde{s},i}}$, and $\nabla_{\pi_{\tilde{s}}}$, the computational complexities are $O(N'S^4T)$, $O(N'S^2T)$, $O(N'S^2T)$, and $O(N'(D+S))$, where N' is the number of HMMs involved in the process of multi-class mapping. The computational complexity for the computing of the log-likelihood is $O(S^2T)$. Then in traditional HMM based FR system, suppose there are N_C subjects (HMMs) under consideration, the computational complexity is $O(N_C S^2 T)$. Then depends on the value of N', the computational complexity of our proposed system can be much less than and at most equal to the traditional HMM based system. One exception is $\nabla_{a_{\tilde{s}''|\tilde{s}'}}$. Due to its overly high computational complexity, we excluded feature vector of category m_a from our experiments.

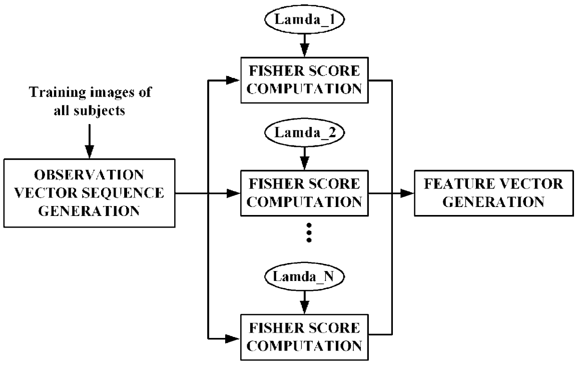
Lamda_1
FISHER SCORE COMPUTATION
Training images of all subjects
Lamda_2
OBSERVATION VECTOR SEQUENCE GENERATION
FISHER SCORE COMPUTATION
FEATURE VECTOR GENERATION
Lamda_N
FISHER SCORE COMPUTATION

Fig. 4. Feature vector generation for the training (testing) of LDA when all competing models are involved in multi-class mapping.

The LDA method can also be used with feature vectors generated from single-class mapping method. One possible scheme is that the training samples of each class are composed of Fisher scores from single-class mapping where the statistical model selected is the corresponding HMM of that class (see Fig. 5). Then the optimal projection matrix is found based on the training samples and the training projections are computed based on the optimal projection matrix. In the testing stage, to compute the distance of a testing image towards class i, the testing feature vector is composed

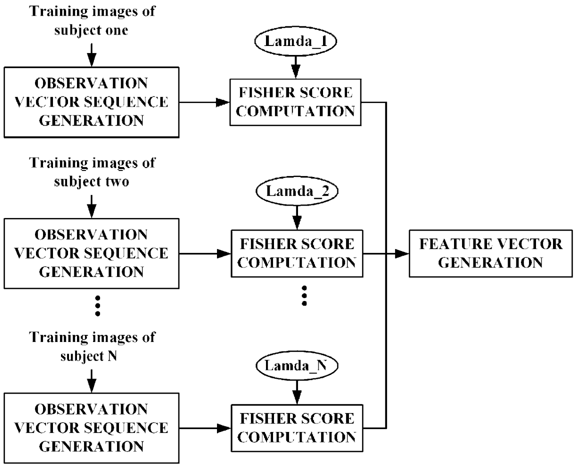
Training images of subject one
Lamda_1
OBSERVATION VECTOR SEQUENCE GENERATION
FISHER SCORE COMPUTATION
Training images of subject two
Lamda_2
OBSERVATION VECTOR SEQUENCE GENERATION
FISHER SCORE COMPUTATION
FEATURE VECTOR GENERATION
Training images of subject N
Lamda_N
OBSERVATION VECTOR SEQUENCE GENERATION
FISHER SCORE COMPUTATION

Fig. 5. Feature vector generation for the training of LDA in single-class mapping.

of Fisher scores from single-class mapping of the testing image to the HMM of class *i*. Then the distance of the testing image towards class *i* is computed. The same kind of process is repeated to find all distances of the testing image towards all competing classes, and each testing image is assigned to the class which has the smallest distance towards the testing image. Beside this scheme, we can also treat the single-class mapping as a special case of multi-class mapping in order to reduce the number of models involved in the training and testing stages. Then above mentioned random selection of HMM can be used. To differentiate these two schemes on single-class mapping in our following experiments, we denote s_mu_1, s_simga_1, and s_mu_sigma_1 as feature vectors to be used in the first scheme. And we denote s_mu_2, s_sigma_2, and s_mu_sigma_2 as feature vectors to be used in the second scheme (random selection).

7. Experiments

7.1. Database

The face image database used in our experiments is the Georgia Tech Face Database (GTFD) [17], which consists of 50 subjects with 15 face images available for each subject. These face images varies in size, facial expression, illumination, and rotation both in the image plane and perpendicular to the image plane. In our experiments, all images in the database were manually cropped and resized to 112 × 92. After the image cropping, most of the complex background has been excluded. Also, in-plane rotation was partially eliminated but the out-of-plane rotation was left untouched. They are further converted to gray level images for both training and testing purposes (see Fig. 6). In our experiments, the training set consists of five randomly selected images of each subject, and the testing set consists of the remaining 10 images of each subject.

Fig. 6. Sample images of Georgia Tech Face Database.

7.2. Training of HMMs

7.2.1. Sampling scheme

To generate the observation vector sequences from the face image, an 8 × 8 sized sliding window is used to scan a face image with 75% overlap between consecutive steps from left to right and from top to bottom. The windowed sub-image blocks are normalized to zero mean and further transformed by an 8 × 8 DCT. Only the 3 × 3 lowest frequency coefficients in the DCT domain are used to form the 9-dimensional observation vectors. The size of the observation vectors is very small comparing to the size of the face image. The choice of small dimension observation vectors is appropriate because of the limited training data problem which is rampant in FR problems. Observation vectors with high dimension will be problematic in the training of HMM and the trained models are prone to be over-fitted.

7.2.2. Statistical model structure

As mentioned in Section 4, the model structure chosen for our system is one-dimensional ergodic HMM with observation density as Gaussian with diagonal covariance matrix. The reason for this choice is two-fold: First, because the out-of-plane rotation was not eliminated from the cropped images, the performance of the HMMs should be robust towards pose variations. Secondly, for the simplicity of the Fisher score formulation, the model structure should not be too complex.

Specifically speaking, the small size of the observation vector means it does not represent any physical part (such as eyes, nose, mouth, chin, etc.) of a human face. A particular observation vector (e.g., a region with flat skin texture), can appear in many location in the observation vector sequence with certain probability. Therefore the HMM is ergodic. We choose Gaussian HMM rather than mixture Gaussian HMM to make sure that the transitional statistics embedded among the observation sequences will be exploited as much as possible. In an extreme example, if we use only mixture models with one state, the transition relationship among clusters will be lost. It can also be justified by the size of the observation vectors. The mixture models are usually more appropriate for large observation vectors. The small observation vectors used in our system can come from any part of a human face. The training algorithm tries to group similar observation vectors into clusters described by Gaussian and then find transition relationships among different clusters. Meanwhile, comparing with more sophisticated model structures, one-dimensional HMM and diagonal covariance matrix of Gaussian can reduce the computation for Fisher scores. The formulae derived on our configuration of HMM is relatively simple (refer Eqs. (5)–(8)) and can be easily computed by the standard forward–backward procedure.

7.3. Experimental facts and discussion

7.3.1. Experimental facts

Experiments in our work include three parts. In the first part, we measure class separability for all categories of feature vectors on the training data set and testing data set. In the second part, we test the effectiveness of all those feature vectors by directly testing the RR. In the third part, we test the computational efficiency of our proposed FR system over the traditional HMM based FR systems.

One way to test the effectiveness of our proposed feature vector generation schemes is to find the class separability of each scheme. To measure class separability, we adopt the trace ratio of between-class scatter matrix and within-class scatter matrix as the criterion [18]

$$J = \frac{\mathrm{tr}\mathbf{S}_\mathrm{b}}{\mathrm{tr}\mathbf{S}_\mathrm{w}}. \qquad (10)$$

There are totally four trace ratios to be generated for each feature vector generation scheme, i.e., trace ratios of training set before LDA and after LDA, trace ratios of testing set before LDA and after LDA. Testing results are listed in Table 4. For comparison purposes, trace ratios of holistic (appearance based) features used by the Fisherface method are also listed. In each row of the table, the before-LDA trace ratio and the after-LDA trace ratio are used to represent the effectiveness of LDA to the improvement of class separability on both the training set and the testing set. In each column of the table, trace ratios of different feature generation schemes are used to exemplify their individual discriminative power. Because of the randomness of the secondary scheme of LDA on single-class mapping, the trace ratios of s_mu_2, s_sigma_2, and s_mu_sigma_2 are the mean value of trace ratios got from 50 different selections of HMM used in the single-class mapping.

The other way to test the effectiveness of our proposed feature vector generation schemes is to directly compute the RR of each of them. Table 5 lists the recognition results. The RR of s_mu_2, s_sigma_2, and s_mu_sigma_2 are the mean value of RR got from 50 different selections of HMM which was used in the single-class mapping. For comparison purposes, the RR of Fisherface method and that of stand alone HMMs are listed also.

To implement our proposed method for FR, one important issue to be considered is the number of HMMs to be used in the process of multi-class mapping as discussed in Section 6. Obviously, the less number of selected HMMs are involved in the multi-class mapping, the lower computational complexity the whole FR system will have. Then in the third part of our experiments, we select feature vectors m_mu and m_sigma to test their RR when different amount of HMMs are involved in the multi-class mapping process. For $1 < N' < N_\mathrm{c}$, we randomly select N' HMMs and test the recognition rate and this procedure is repeated for 100 possible selections of N' HMMs. When $N' = 1$, i.e., single-class mapping, we use the second LDA scheme on single-class mapping. Then each of all the HMMs under consideration is selected once and the RR of the system based on the model is found. The boxplot (also known as “box and whisker'' plot) of the RR on feature vectors m_mu and m_sigma is illustrated in Fig. 7.

7.3.2. Discussion

From Table 4 we can see that in most cases, the class separability of various feature generation schemes will improve after LDA. It is also apparent that the improvements on training set are always much higher than those on testing set. This is obviously true because the eigenvectors for the subspace projection are calculated based on the training set.

The last column of Table 4 deserves more attention, because the after-LDA trace ratio of testing set reveals the real effectiveness of feature vectors in term of class separability. For feature vectors from multi-class mappings, most of their trace ratios surpass that of holistic (appearance based) features. This means feature vectors from multi-class mappings are expected to have higher discriminative power than that of holistic features. There is one exception, i.e., m_pi. This type of feature vectors have a relatively very small trace ratio, which indicates the discriminative power of these feature vectors is very limited. This is also exemplified by its recognition rate: 16.60%. Ironically, the trace ratio of the training set of this category is the highest among all the different schemes. Possible reasons contributing to this phenomenon are:

- After the sampling scheme, the distribution pattern of the first observation vector in the observation vector sequence is highly unpredictable because it represents the first 8×8 image block in the cropped images and it could come from the hair, the skin, or even the background. Then the

Table 4
Class separability

Scheme	Training set		Testing set	
	Bf LDA	Af LDA	Bf LDA	Af LDA
Fisherface (holistic)	1.0389	140.1155	0.8506	1.7165
Multi-class mappings				
m_loglik	2.1197	7.0007	1.2020	3.1653
m_mu	0.9165	355.2917	0.7395	5.9189
m_sigma	1.2122	301.9155	1.0562	6.1072
m_pi	0.3023	4.3231e3	0.1098	0.1098
m_mu_sigma	1.0537	369.2718	0.8846	7.3170
m_loglik_mu_sigma	1.0551	368.4242	0.8859	7.3182
Multi-class mappings with holistic (*appearance based*) *features*				
m_loglik_holi	1.0438	74.6465	0.8552	2.4675
m_mu_holi	0.9593	452.4430	0.7788	8.6503
m_sigma_holi	1.1458	515.7146	0.9754	5.4968
m_pi_holi	0.8672	930.6172	0.1098	0.1098
m_mu_sigma_holi	1.0504	432.1935	0.8768	9.1218
m_loglik_mu_sigma_holi	1.0515	432.4920	0.8778	9.1229
Single-class mappings				
s_mu_1	0.0267	0.6738	0.3492	0.4253
s_sigma_1	1.7094	58.9247	1.2308	1.5347
s_mu_sigma_1	0.4891	77.9655	0.6116	3.6006
s_mu_2	0.9214	116.4136	0.8638	2.3136
s_sigma_2	1.2123	87.2399	1.1399	2.4139
s_mu_sigma_2	1.0739	149.4664	1.1378	3.5251

Table 5
Recognition rates

Scheme	RR	Scheme	RR
Fisherface (holistic)	70.20%	HMM	90.80%
Multi-class mappings			
Scheme	RR	Scheme	RR
m_loglik	93.60%	m_mu	95.80%
m_sigma	95.40%	m_pi	16.60%
m_mu_sigma	95.40%	m_loglik_mu_sigma	95.40%
Multi-class mappings with holistic (*appearance based*) *features*			
Scheme	RR	Scheme	RR
m_loglik_holi	70.20%	m_mu_holi	96.40%
m_sigma_holi	96.00%	m_pi_holi	60.40%
m_mu_sigma_holi	96.60%	m_loglik_mu_sigma_holi	96.60%
Single-class mappings			
Scheme	RR	Scheme	RR
s_mu_1	45.20%	s_mu_2	86.40%
s_sigma_1	47.80%	s_sigma_2	89.80%
s_mu_sigma_1	50.80%	s_mu_sigma_2	92.52%

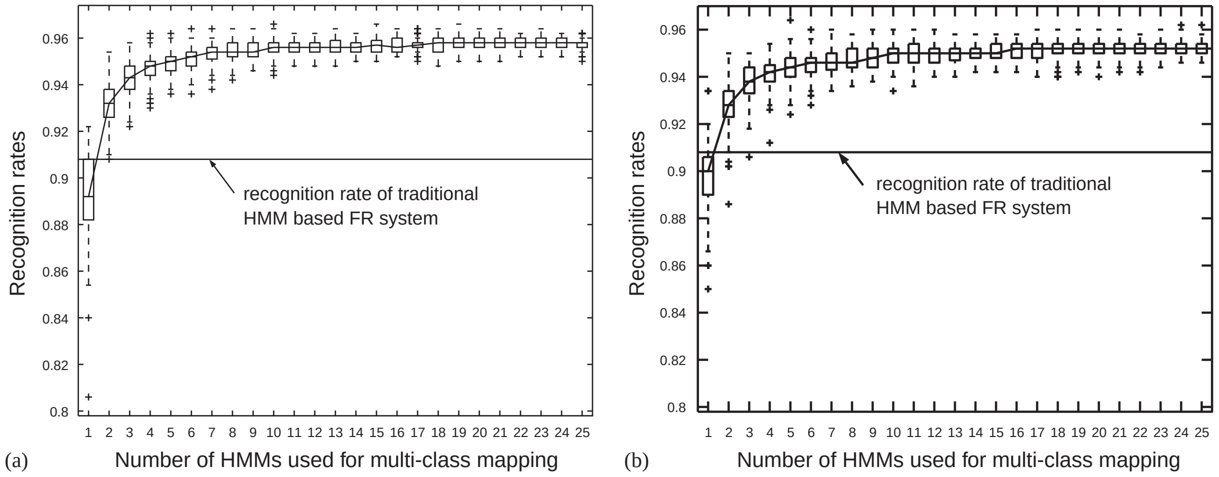

Fig. 7. Boxplot of the RR when various number of HMMs are used for multi-class mapping. Feature vectors used in (a) and (b) are feature vectors m_mu and m_sigma, respectively.

discriminative power of the initial state distribution itself is limited.

- Comparing to other parameters of HMM, the amount of information utilized by the re-estimation of π is limited. For example, this can be easily seen from the re-estimation formulae of the HMM, (refer Eqs. (11)–(14)), where $\bar{\Sigma}_i$ means the re-estimation of the covariance matrix of state i

$$\bar{\pi}_{\tilde{s}} = \gamma_1(\tilde{s}), \tag{11}$$

$$\bar{a}_{\tilde{s}''|\tilde{s}'} = \frac{\sum_{t=1}^{T-1} \xi_t(\tilde{s}', \tilde{s}'')}{\sum_{t=1}^{T-1} \gamma_t(\tilde{s}')}, \tag{12}$$

$$\bar{\mu}_{\tilde{s}} = \frac{\sum_{t=1}^{T} \gamma_t(\tilde{s}) \cdot \mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(\tilde{s})}, \tag{13}$$

$$\bar{\Sigma}_{\tilde{s}} = \frac{\sum_{t=1}^{T} \gamma_t(\tilde{s}) \cdot (\mathbf{o}_t - \mu_{\tilde{s}})(\mathbf{o}_t - \mu_{\tilde{s}})'}{\sum_{t=1}^{T} \gamma_t(\tilde{s})}. \tag{14}$$

- Due to the limited training data (in our experiments, for each subject in the database, there are only five randomly selected images used for training), the estimation of π is even more unreliable than the other parameters of HMM.

The RR in Table 5 also demonstrate the effectiveness of multi-class mapping. From the table we can see that, with the exception of m_pi, the performances of all multi-class mapping schemes (including m_loglik, which is in fact the multi-class mapping of the zeroth-order derivatives of the log-likelihood) surpass both the stand alone HMM and Fisherface method. Although both Fisherface method and our proposed system use subspace method such as LDA, our proposed method displays substantial improvement in recognition rate because of the difference between the feature generation methods. The Fisherface FR system in our experiments is a little bit naive in the sense that images used for the training and testing were taken directly by the Fisherface method without the adjustment on the precise location of important facial parts such as eyes, nose, and mouth (see Fig. 6), which is normally solicited by holistic feature based method such as Fisherface method e.g., [19,20]. Nevertheless, stringent preprocessing for precisely located facial parts are not solicited in our proposed methods. The testing environments of Fisherface method and our proposed methods are the same, yet the RR of our methods are clearly better than the Fisherface method. This indicates the flexibility of the statistical model system on FR.

Moreover, from the last column of Table 4, we can see that all trace ratios for feature vectors from Fisher scores of multi-class mappings are much higher than their counterparts in single-class mapping schemes. Also, the same tendency occurs on their RR. This strongly indicates that feature vectors generated from multi-class mapping indeed have more discriminative power over that of single-class mapping. For various combination schemes of different kinds of features, we can see from the Table 4 that in many cases, after combining two or more different categories of features together, the trace ratio will grow higher after the combination. For instance, the after-LDA trace ratio of the testing set of m_mu_sigma is higher than those of m_mu and m_sigma. This is also exemplified by the RR. For example, the highest recognition rate (96.60%) among all categories of feature vectors is obtained by m_loglik_mu_sigma_holi, i.e., the combination of the multi-class mappings of log-likelihood, $\nabla_{\mu_{\tilde{s},i}}$, $\nabla_{\sigma_{\tilde{s},i}}$, and holistic (appearance based) features. But there are some exceptions, e.g., m_loglik_holi and m_pi_holi. This may be explained by the fact that directions of projection vectors of the optimal projection matrix $\mathbf{W}_{\text{opt}}$ is largely determined by training features which have higher after-LDA trace ratio of the training set over that of other features in the combination. For example, while the after-LDA trace ratio of the training set of m_loglik is 7.0007, the after-LDA trace ratio of the training set of holistic features (Fisherface method) is 140.1155. This indicates that when the training set of the whitened combination of these two kinds of features is used for LDA, features of m_holi will play much more important roles in the process of determining the optimal projection matrix $\mathbf{W}_{\text{opt}}$ than those of m_loglik. However features of m_holi are inferior to features of m_loglik in term of discriminative power, noting that the RR of LDA of m_loglik and Fisherface method are 93.60% and 70.20%, respectively. Then the benefits of the combination scheme are impaired by the unbalanced contribution of the inferior features, which, in this case, are those features of m_holi.

For a particular feature combination scheme i, we denote J^i_{train} and J^i_{test} as the training set's and testing set's after-LDA trace ratio. One heuristic to evaluate the potential of various combination schemes is that, if $J^i_{\text{train}}/J^j_{\text{train}}$ and $J^i_{\text{test}}/J^j_{\text{test}}$ are roughly at the same scale, the combination of features from scheme i and j is very likely to have higher after-LDA trace ratio of testing set, e.g., $J^{ij}_{\text{test}} \geqslant J^i_{\text{test}}$ and $J^{ij}_{\text{test}} \geqslant J^j_{\text{test}}$. This means more discriminative power is expected.

From Fig. 7, we can see that as the number of HMMs involved in the multi-class mapping goes up, the RR go up rapidly and surpass the recognition rate of traditional HMM based FR system. For example, we can see from the figure that when there are only five HMMs involved in the multi-class mapping, the lowest recognition rates for m_mu and m_sigma are above 92%, which are higher than that of the RR of traditional HMM based FR system (90.80%). In Fig. 7, we can see that even the computational complexity is lowered by a factor of 10 ($N' = 5,\ N_{\text{c}} = 50$), the RR of the proposed FR system based on only five randomly selected HMMs are always higher than that of traditional HMM based FR system. This validates the effectiveness of the proposed FR system in the perspective of computational complexity.

In summary, our experimental results suggest:

- Multi-class mapping of log-likelihood can be used to form feature vectors and have higher recognition rate over those of Fisherface method and stand alone HMMs.
- Multi-class mappings of Fisher scores such as $\nabla_{\mu_{\tilde{s},i}}$, $\nabla_{\sigma_{\tilde{s},i}}$ and their combinations have higher RR than Fisherface method and stand alone HMMs. Also, they are superior over their counterparts of single-class mappings.
- Combinations of holistic features with Fisher scores $\nabla_{\mu_{\tilde{s},i}}$, $\nabla_{\sigma_{\tilde{s},i}}$ can result in even higher RR.
- While obtaining higher RR, the computational complexity of LDA on multi-class mapping of Fisher score method can be much lower than that of the traditional HMM based FR system.

8. Conclusion

In this paper, we developed a new feature vector generation scheme based on multi-class mapping of Fisher scores. We presented the derivation of Fisher scores for one-dimensional ergodic HMM with diagonal Gaussian observation density. Conventional appearance based features were combined with statistical model based features and overall performance improvement is observed. The effectiveness of the proposed feature vector generation scheme is testified by higher RR and lower computational complexity. The proposed method is quite generic and we expect it can be further explored in areas of pattern recognition where statistical models are involved such as speech recognition and texture analysis.

Acknowledgements

The authors thank W. Meissner for the valuable discussion and suggestions during the preparation of the paper.

References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1991) 71–86.

[2] P. Bellhumeur, J. Hespanha, D.J. Kriegmand, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 711–720.

[3] K. Etemad, R. Chellappa, Face recognition using discriminant eigenvectors, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 2148–2151.

[4] F. Samaria, Face recognition using hidden Markov model, Ph.D. Thesis, University of Cambridge, 1995.

[5] A. Nefian, A hidden Markov model-based approach for face detection and recognition, Ph.D. Thesis, Georgia Institute of Technology, 1999.

[6] A. Nefian, Embedded Bayesian networks for face recognition, ICME 2002—IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 2002, pp. 133–136.

[7] S. Eickeler, S. Müller, G. Rigoll, High performance face recognition using pseudo-2-D hidden Markov models, European Control Conference (ECC), August 1999.

[8] H. Othman, T. Aboulnasr, Low-complexity 2-D hidden Markov model face recognition, ISCA 2000—IEEE International Symposium on Circuits and Systems, Geneva, Switzerland, May, 2000, pp. V33–V36.

[9] T. Jaakkola, D. Haussler, Exploiting generative models in discriminative Classifiers, in: S.A. Solla, T.K. Leen, K.R. Müller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 2000.

[10] T. Jaakkola, D. Haussler, Maximum entropy discrimination, Technical Report AITR-1668 MIT, 1999.

[11] M. Seeger, Covariance kernels from Bayesian generative models, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge, MA, 2002.

[12] K. Tsuda, S. Akaho, M. Kawanabe, K.R. Müller, Asymptotic properties of the Fisher kernel, Neural Comput. 16 (1) (2004) 115–137.

[13] S. Amari, Differential-Geometrical Methods in Statistics, Lecture Notes in Statistics, vol. 28, Springer, New York, 1985.

[14] N. Smith, M. Gales, Using SVMs to classify variable length speech patterns, Technical Report CUED/F-INFENG/TR.412, Cambridge University Engineering Department, June 2001.

[15] S. Fine, J. Navrátil, R. Gopinath, A Hybrid GMM/SVM approach to speaker identification, ICASSP 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Utah, USA, May, 2001, pp. 417–420.

[16] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989).

[17] Georgia Tech Face Database, ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/.

[18] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Boston, 1990.

[19] W. Zhao, R. Chellappa, A. Krishnaswamy, Discriminant analysis of principal components for face recognition, AFGR 1998—IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April, 1998, pp. 336–341.

[20] A. Martínez, A. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 228–233.

About the Author—LING CHEN received the B.S. degree in Northwestern Polytechnical University, China, in 1996, the M.S. degree in University of Electronic Science and Technology of China, China, in 1999, both in electrical engineering. Currently, he is a Ph.D. candidate in the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, New Jersey, USA. His research interests include biometrics, machine learning, statistical pattern recognition and neural networks.

About the Author—HONG MAN received the B.S. degree from Soochow University, China, in 1988, the M.S. degree from Gonzaga University in 1994, and the Ph.D. degree from Georgia Institute of Technology in 1999, all in Electrical Engineering. He joined Stevens Institute of Technology in 2000, and currently he is an assistant professor in the Department of Electrical and Computer Engineering. He is serving as the director for Computer Engineering undergraduate program in the ECE department, and coordinator for NSA Center of Academic Excellence in Information Assurance in the School of Engineering. He is a member of the IEEE and ACM. He served as member of organizing committee for IEEE International Workshop on Multimedia and Signal Processing (MMSP) 2002 and 2005, member of technical program committee for IEEE Vehicular Technology Conference (VTC) Fall 2003, and IEEE/ACM International Conference on E-Business and Telecommunication Networks (ICETE) 2004. He is a committee member on IEEE SPS TC for Education. His current research interests include image analysis, medical imaging and multimedia networking.

About the Author—ARA V. NEFIAN is a Senior Researcher at Intel Corporation, Microprocessor Research Labs in Santa Clara, California. Ara received the engineering diploma degree in Electrical Engineering in 1994 from the “Politehnica’’ University of Bucharest, Romania. In 1995, he received the MSEE degree and in 1999, the Ph.D. degree, all in Electrical Engineering from Georgia Tech, Atlanta. Current research interests include the study of graphical models for face and gesture recognition and audio-visual signal processing.