Unified Probabilistic Models for Face Recognition from a Single Example Image per Person

Pin Liao and Li Shen

Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, P.R. China

E-mail: pliao@ict.ac.cn

Received August 6, 2002; revised November 19, 2002.

Abstract This paper presents a new technique of unified probabilistic models for face recognition from only one single example image per person. The unified models, trained on an obtained training set with multiple samples per person, are used to recognize facial images from another disjoint database with a single sample per person. Variations between facial images are modeled as two unified probabilistic models: within-class variations and between-class variations. Gaussian Mixture Models are used to approximate the distributions of the two variations and exploit a classifier combination method to improve the performance. Extensive experimental results on the ORL face database and the authors' database (the ICT-JDL database) including totally 1,750 facial images of 350 individuals demonstrate that the proposed technique, compared with traditional eigenface method and some well-known traditional algorithms, is a significantly more effective and robust approach for face recognition.

Keywords pattern recognition, face recognition, Gaussian mixture model, classifier combination, unified probabilistic model

1 Introduction

Over the past few years, automated face recognition has received significant attention and become one of the most active research areas in computer vision and pattern recognition. There are at least two reasons for this trend: the notable growth in the wide range of commercial and law enforcement applications, and the availability of many new feasible technologies after over 30 years of research.

Numerous methods have been proposed for face recognition within the last several years. Generally they can be classified into two categories according to feature-extraction techniques: analytic geometrical feature based and holistic template matching based.

The analytic geometrical-feature-based techniques are based on the computation of a set of explicit geometrical features from a facial image, such as the relative positions and other parameters of eyes, mouth, nose, chin and face-outline. Wiskott *et al.* proposed an elastic bunch graph matching method[1], which extracts concise face descriptions in the form of image graphs. Cootes *et al.* developed an Active Shape Model (ASM)[2] to model shape and local gray-level appearance and locate flexible objects in new images. Lanitis *et al.* used this approach to interpret face images and warp the face into a normalized frame[3]. Following that, Cootes *et al.* presented an Active Appearance Model (AAM)[4], which is a generalization of the ASM and uses all the information in the image region covered by the target object, rather than just that near modeled edges. Penev *et al.* developed a technique of Local Feature Analysis (LFA)[5] that builds a sparsely-distributed representation of faces in terms of flexible templates of local features.

The holistic template matching methods do not use any detailed biometric knowledge of the human face, but consider the global properties of facial images. Based on Principal Component Analysis (PCA), an effective technique for signal representation and dimensionality reduction, Turk *et al.* proposed an eigenface system, which projects face images onto a feature space that spans the significant variations among known face images[6]. The eigenface method, being a milestone work, has become a common performance benchmark for comparison in the area. As a classical pattern recognition technique, Linear Discriminant Analysis (LDA) has been applied for face recognition[7–9]. Unlike PCA that derives Most Expressive Features (MEF), LDA derives Most Discriminating Features (MDF)[8] and becomes an attractive choice for face recognition and verification tasks. However, LDA may suffer from poor generalization of new data,

*Regular Paper

This work is supported partly by the National Natural Science Foundation of China (Grant No.69789301), National Hi-Tech R&D 863 Program of China (Grant No.2001AA114190), and Sichuan Chengdu Yinchen Net. Co. (YCNC).

especially when the training dataset is small or not representative[10].

Furthermore, in very recent years quite a number of novel techniques have been exploited in face recognition, such as: Support Vector Machines (SVMs)[11–13], kernel methods[14–16], Independent Component Analysis (ICA)[17], Gaussian Mixture Models (GMMs)[18–20], Neural Networks[21–24], evolutionary pursuit[25], probabilistic reasoning models[26], optimal discriminant vectors[27,28], Fourier transform[29], nearest feature line[30], Hidden Markov Models (HMMs)[31], Hausdorff distance[32] and dual attribute graph[33].

Face recognition differs from many other traditional classification problems such as optical character recognition (OCR). In a traditional classification application, there are usually few classes but numerous samples for each class. With numerous samples per class, samples not seen before can be classified by interpolating among the training data points. In contrast, for a system of face recognition, there are popularly a large number, often over several hundreds, of subjects and only a few images for each person. And it is not uncommon to have only a single example image for each person in many applications. Hence face recognition becomes a problem of extrapolation from the single samples, and the simple Euclidean nearest-neighbor matching technique is often adopted[6], as most advanced classification techniques perform poorly with only one sample per class.

However, face recognition has another very important and particular characteristic, which has hardly ever been pointed out explicitly in literature so far as we are aware of: human faces are very similar objects with similar geometrical shape and configuration; and as a result, the variations of each specific person's facial images, due to changes of pose, illumination, expression, age and so on, are rather similar to each other. Therefore, using a training set with multiple samples per class from a given known people group, which is usually easy to be obtained, we can build unified probabilistic models to model all the variations corresponding to each specific subject. Then the unified probabilistic models trained on samples from a known people group, for the significant similarity of the variations, can be generalized well to samples from another different unknown people group. Consequently, the considerable generalization of the unified models provides a promising solution to the problem of face recognition from a single example image for each target person. The special and interesting characteristic is demonstrated by the following extensive experimental results.

To some extent, the idea of unified models is originated from the techniques in [11, 34], but the excellent generalization of the unified models to unknown people's facial images was never mentioned explicitly. Moghaddam *et al.*[34] proposed a novel method for face recognition based on a Bayesian analysis of image differences, which performed consistently near the top in the 1996 FERET test[35]. In the approach, variations between facial images are modeled as two mutually exclusive classes: intra-personal (variations in appearance of the same individual, due to different expressions, poses or lighting) and extra-personal (variations in appearance due to a difference in identity). Afterwards, both the classes are assumed as normal distributions, and then two kinds of similarity measures, maximum *a posteriori* (MAP) and maximum likelihood (ML), are computed. The MAP similarity measure is based on both intra-personal differences (within-class variations) and extra-personal differences (between-class variations); however, the ML approach only uses intra-personal differences (within-class variations). It was reported that the former outperformed the latter with a minor (2–3%) increase in recognition rate according to the experimental results. In the same way as [34], Philips[11] also modeled dissimilarities between two facial images as two classes: dissimilarities between facial images of the same person, and dissimilarities between facial images of different people. Differing from [34], SVMs are directly applied to the two-class problem to produce a similarity measure between two facial images. Both the techniques achieved much better performances than the standard eigenface approach in the respective experiments.

There are inevitably some inaccuracies and limitations with the approach in [34], which simply uses normal density to model the distributions of the two kinds of variations, since it is usually difficult to provide a proper representation of a practical distribution by a common typical distribution form of a parametric function in statistics. Consequently, we propose a technique to model the variations using GMMs instead of normal density, since an important attribute of mixture models is that they can approximate any continuous density to arbitrary accuracy provided the model has a sufficiently large number of components, and provided the parameters of the model are chosen correctly[36]. Note that GMMs were also used in

[18–20], in a different manner, to learn the global distribution or a specific distribution but not the variations' distributions. Since the selection of the number of mixture components is still an open problem, we choose a method of classifier combination as a solution and make a notable improvement on performance. In our approach, the unified models are trained on an obtained training set with multiple samples per person. Then the trained models are used to recognize facial images from another disjoint database with a single sample per person. Extensive experimental results on the ORL face database and our own database (the ICT-JDL database) including totally 1,750 facial images of 350 individuals show that our technique outperforms remarkably the two above-mentioned methods and the standard eigenface approach, and demonstrate that the unified probabilistic models are an effective and robust technique for face recognition from only a single example facial image per person.

The remainder of this paper is organized as follows. In Section 2, the unified models of within-class variations and between-class variations are described. Section 3 gives a brief description of Gaussian Mixture Models. Section 4 introduces the process of combining multiple GMMs based on the mean rule. The experimental procedures and results are presented in detail in Section 5. In Section 6, we discuss some important issues concerning the proposed approach. Finally, we conclude and give future work in Section 7.

2 Within-Class Variations and Between-Class Variations

As shown in [34], the intensity difference between two facial images I_1 and I_2 is denoted by $\Delta = I_1 - I_2$. Then two mutually exclusive classes are defined as: within-class variations Ω_I and between-class variations Ω_E. Consequently, an M-ary classification problem for M individuals is reduced to a binary classification problem with Ω_I and Ω_E.

In terms of the within-class *a posteriori* probability as given by MAP rule, the similarity measure between two facial images can be directly defined as:

$$\begin{aligned} S(I_1, I_2) =& P(\Delta \in \Omega_I) = P(\Omega_I|\Delta) \\ =& \frac{p(\Delta|\Omega_I)P(\Omega_I)}{p(\Delta|\Omega_I)P(\Omega_I) + p(\Delta|\Omega_E)P(\Omega_E)} \end{aligned} \tag{1}$$

An alternative probabilistic similarity measure can be defined in simpler form using the ML rule instead of the MAP rule by only exploiting the within-class variations,

$$S(I_1, I_2) = P(\Delta|\Omega_I) \tag{2}$$

Therefore, when in identification there is a gallery $\{y_l\}$ of M individuals and a probe x is to be identified, the similarity score between x and each y_l is $S(x, y_l)$. Accordingly the probe is identified as person k with the maximum similarity score, namely

$$k = \arg\max_l S(x, y_l) \tag{3}$$

where $l = 1, \ldots, M$.

To make a theoretical analysis of the variations, we write the conditional density $P(\Delta|\Omega_I)$ as a linear combination of the densities of the with-class variations of each class, and $P(\Delta|\Omega_E)$ as a linear combination of the densities of the between-class variations of each two different classes respectively.

$$p(\Delta|\Omega_I) = \sum_{i=1}^{N} p(\Delta|\Omega_{ii})P(\Omega_{ii}|\Omega_I) \tag{4}$$

$$p(\Delta|\Omega_E) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} p(\Delta|\Omega_{ij})P(\Omega_{ij}|\Omega_E) \tag{5}$$

where Ω_{ij} is denoted as the variations from class i to class j, and N the total number of all human beings as we consider an ideal condition (which can also be denoted as the number of individuals in the training set for another consideration).

Generally, all individuals are treated evenly. Uniform "flat" priors are adopted: $P(\Omega_{ii}|\Omega_I) = 1/N$ in (4) and $P(\Omega_{ij}|\Omega_E) = 1/(N \times (N-1))$ in (5). As a result, we obtain $P(\Omega_I) = N/N^2 = 1/N$ and $P(\Omega_E) = N \times (N-1)/N^2 = (N-1)/N$.

We denote each specific class-conditional probability density function as $f_i(x)$. And then $p(\Delta|\Omega_{ij})$, the pdf of the difference between two independent random variants, can be written as

$$p(\Delta|\Omega_{ij}) = \int_{-\infty}^{+\infty} f_i(\Delta + z) f_j(z) \mathrm{d}z. \tag{6}$$

For simplification, $p(\Delta|\Omega_{ij})$ can be denoted as $p_{ij}(\Delta)$. Hence we obtain

$$\begin{aligned} p(\Delta|\Omega_I) =& \frac{1}{N} \sum_{i=1}^{N} p_{ii}(\Delta) \\ =& \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} f_i(\Delta + z) f_i(z) \mathrm{d}z \end{aligned} \tag{7}$$

$$\begin{aligned} p(\Delta|\Omega_E) &= \frac{1}{N \times (N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} p_{ij}(\Delta) \\ &= \frac{1}{N \times (N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \\ &\quad \cdot \int_{-\infty}^{+\infty} f_i(\Delta + z) f_j(z) \mathrm{d}z. \end{aligned} \tag{8}$$

Consequently, we can write the unified decision rule of the MAP method as

$$\begin{aligned} k &= \arg\max\nolimits_l \frac{\sum_{i=1}^{N} p_{ii}(x - y_l)}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}(x - y_l)} \\ &= \arg\max\nolimits_l \frac{\sum_{i=1}^{N} \int_{-\infty}^{+\infty} f_i((x - y_l) + z) f_i(z) \mathrm{d}z}{\sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{\infty} f_i((x - y_l) + z) f_j(z) \mathrm{d}z} \end{aligned} \tag{9}$$

and the unified decision rule of the ML method as

$$\begin{aligned} k &= \arg\max\nolimits_l \sum_{i=1}^{N} p_{ii}(x - y_l) \\ &= \arg\max\nolimits_l \sum_{i=1}^{N} \int_{-\infty}^{+\infty} f_i((x - y_l) + z) f_i(z) \mathrm{d}z \end{aligned} \tag{10}$$

It is easy to find that, if all the specific densities are normal densities with equal covariance matrices and different mean vectors, both decision rules are equivalent to the Bayes optimal classifier.

However, it is really troublesome to compute the theoretical errors of the two decision rules, as (9) and (10) demonstrate. In [34] the experimental results show that the unified MAP always makes a minor improvement over the unified ML, but all of our simulation data, on the contrary, demonstrate that the latter moderately outperforms the former. In Section 6, the problem is discussed in more detail.

3 Gaussian Mixture Models

GMMs are a semi-parametric approach to density estimation, combining the advantages of both parametric and non-parametric methods. GMMs are not restricted to specific functional forms, where the size of the model only grows with the complexity of the problem being solved, and not simply with the size of the data set[36].

GMMs are defined as a linear combination of K component normal densities $p(x|i)$:

$$p(x) = \sum_{i=1}^{K} p(x|i) P(i) \tag{11}$$

where the number K of components is typically much less than the size T of the training data set $\{x^t\}$. Such a representation is called a mixture distribution[36] and the coefficients $P(i)$ are called the mixing parameters.

$P(i)$ can be regarded as the prior probability of the data point generated from component i of the mixture. These priors are chosen to satisfy the constraints: $\sum_{i=1}^{K} P(i) = 1$ and $0 \leqslant P(i) \leqslant 1$.

The component density functions $p(x|i)$ are normalized so that $\int p(x|i) \mathrm{d}x = 1$, and hence can be regarded as class-conditional densities.

Each component density $p(x|i)$ is assumed to be a normal density with a covariance matrix Σ_i and a mean vector $\boldsymbol{\mu}_i$. Firstly, the mean vector $\boldsymbol{\mu}_i$ and the covariance Σ_i for each Gaussian component is initialized using the K-means algorithm, and $P(i)$ is initialized as $1/K$. And then, the parameters of GMMs can be achieved in a maximum likelihood framework by the well-known iterative expectation-maximization (EM) algorithm[37].

4 Combination of Multiple Gaussian Mixture Models

Although GMMs are a popular tool for density estimation, choosing the number of mixture components is still an open problem. There are many approaches proposed recently to this problem. But our intention here is to investigate the basic effectiveness of the GMMs approach. So we adopt a simple classifier combination approach based on the mean rule, for that a combination of many different classifiers can lead to notable improvements in the predictions on new data and outperform the best single classifier used in isolation[36], and that the classifier combination scheme of mean rule can usually outperform many other schemes such as the product rule, min rule, max rule, median rule, and majority voting[38,39].

By setting various numbers K of GMMs' components, different GMMs are produced to approximate the densities of the two kinds of variations between facial images, and then different classifiers are yielded.

It is supposed that we have obtained a set of L various classifiers C_l, where $l = 1, \ldots, L$ based on different GMMs. Then for a gallery $\{y_i\}$ of M known individuals, the similarity score between a probe x and each y_i presented by a classifier C_l is denoted as $S_l(x, y_i)$.

The similarity scores are firstly normalized as

$$S'_l(x, y_i) = \frac{S_l(x, y_i)}{\sum_{i=1}^{M} S_l(x, y_i)} \tag{12}$$

and then the mean combination rule is introduced to obtain the combined similarity score

$$S_{\mathrm{com}}(x, y_i) = \frac{1}{L} \sum_{l=1}^{L} S'_l(x, y_i) \tag{13}$$

As a result, the probe is identified as person k with

$$k = \arg\max{}_i S_{\mathrm{com}}(x, y_i) \tag{14}$$

5 Experiments

In order to evaluate the performance of the technique, extensive experiments are conducted on two face databases: the ORL database and a larger database, our own ICT-JDL database, consisting of 1,750 facial images of 350 individuals.

We approximate the two distributions using GMMs with different numbers of components and obtain two kinds of similarity measures based on the unified MAP rule and the unified ML rule respectively. Finally, combined similarity measures are obtained based on the classifier combination scheme of mean rule. In all our experiments, the latter outperforms the former with a minor advantage in the recognition rate, contrary to the experimental results of [34]. We therefore show only the results of our approach based on the unified ML rule in this paper. For comparison, the standard eigenface approach, SVM[11], the unified MAP[34] and the unified ML[34] are also evaluated with the two databases. As to the unified MAP, the unified ML and our method, the principal subspace dimensions for Ω_I and Ω_E are set as $M_I = 50$ and $M_E = 50$, respectively. The SVM-based algorithm, using a radial basis kernel, is conducted on the principal subspace consisting of the first 50 eigenfeatures.

5.1 ORL Database

The ORL database[40] is from the Olivetti Research Laboratory in Cambridge, U.K. There are ten different images for each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (with glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position, with tolerance for some tilting and rotation of up to about 20 degrees. There is some variation in scale of up to about 10%. The images are of grayscale with a resolution of 92 × 112. Some examples from the database are shown in Fig.1.

The 400 images are divided into non-overlapping training and testing sets. Each set consists of 10 images of 20 people. Accordingly 1,800 ($P_{10}^2 \times 20$) within-class difference samples and 38,000 ($P_{20}^2 \times 10 \times 10$) between-class difference samples corresponding to the classes Ω_I and Ω_E respectively are created from the training set.

A gallery is built by randomly selecting 20 images from the testing set, with one image per person, and the remaining 180 images make up a probe set. This random selection has been repeated for 100 times to create 100 different galleries and 100 different probe sets accordingly. All recognition techniques to be evaluated below are assessed with these 100 testing sets.

As the standard eigenface approach, SVM, the unified MAP and the unified ML obtain average recognition rates of 71.38%, 77.78%, 79.87%, and

Fig.1. Some examples from the ORL database.

81.68% respectively. We combine GMMs with 1 to 8 Gaussian components using the mean combination method and achieve an average recognition rate of 84.25%. Table 1 shows that our method outperforms all the others as a robust technique.

Table 1. Results for ORL Database by Different Methods

Approach	Average recognition rate (%)	Variance
Eigenface[6]	71.38	0.0434
SVM[11]	77.78	0.0384
Unified MAP[34]	79.87	0.0237
Unified ML[34]	81.68	0.0323
Combination of GMMs	84.25	0.0275

Fig.2 illustrates the average recognition rates according to GMMs with 1 to 8 components and the combinations of them. It is notable that the combinations demonstrate a remarkable robustness and outperform the best of the individual GMMs when the number of the combined GMMs is not too small.

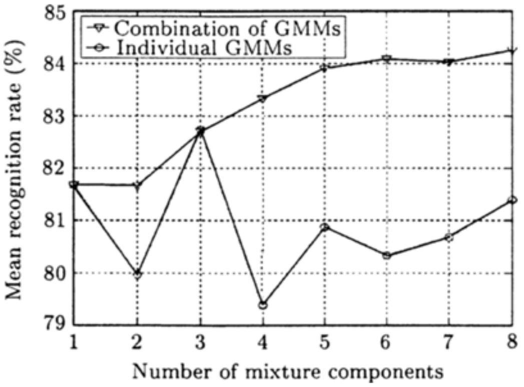


Fig.2. Results for ORL database according to GMMs with 1 to 8 components and the combinations of them.

5.2 ICT-JDL Database

The ICT-JDL database consists of 1,750 facial images of 350 persons, with 5 images per subject. All the images were taken with a general USB camera, which contain quite a high degree of variability in expression and pose. For some subjects, the images were taken with variation of illumination and facial details (with glasses/no glasses). Fig.3 shows some examples from the database.

All the images are partitioned into disjoint training set and testing set. 250 images of 50 individuals are contained in the training set, and all the rest constitute the testing set.

In the training set, tens of synthetic samples are derived from each original image with slight geometric transforms of translation, scaling, rotation and mirror-reflection. There are two reasons for exploiting the technique of making synthetic samples: firstly, it can make the exploited statistical classification algorithm perform more robust which usually needs a sufficiently large training set; furthermore, it can compensate alignment error to some extent, as none of the existing alignment algorithms can obtain a pixel precision. The technique has not been used in the experiments on the ORL database, since many of the ORL images do not leave enough margin for the transforms. At last, 60,000 randomly-selected within-class variation samples and 60,000 randomly-selected between-class variation samples are obtained from the expanded training set.

In the testing set, a normal facial image (taken with nearly frontal pose, neutral expression and ambient lighting condition) for each of the 300 individuals is chosen as a gallery example. And all the remaining 1,200 images constitute the probe set.

In Table 2, it is demonstrated that our method outperforms all the others with a recognition rate of 94.67%. Fig.4 shows the experimental results for the ICT-JDL database according to GMMs with 1 to 8 components and the combinations of them,

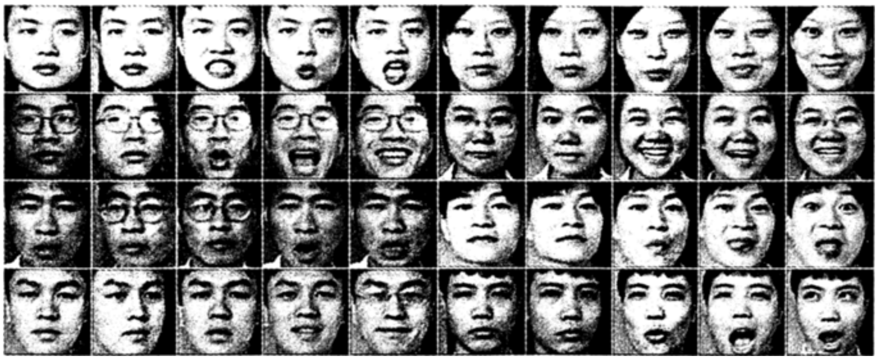


Fig.3. Some examples from the ICT-JDL database.

and indicates that the combination method provides a solution to the problem of choosing the number of mixture components.

Table 2. Results for our Database by Different Methods

Approach	Recognition rate (%)
Eigenface[6]	76.08
SVM[11]	90.42
Unified MAP[34]	89.75
Unified ML[34]	92.25
Combination of GMMs	94.67

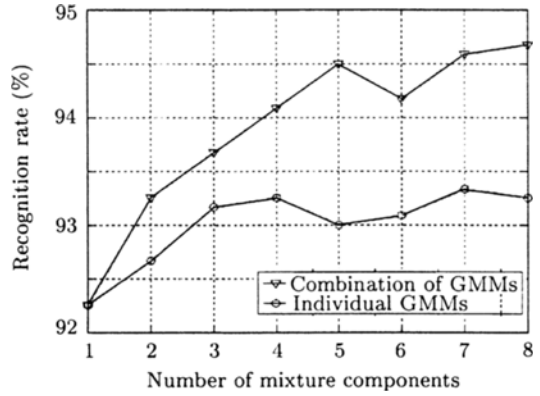


Fig.4. Results for ICT-JDL database according to GMMs with 1 to 8 components and the combinations of them.

6 Discussions

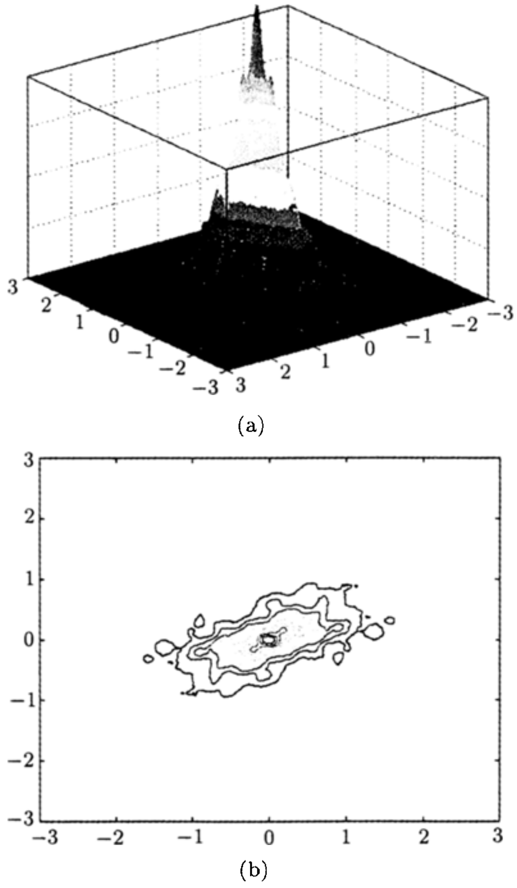
In [34] it is claimed that the two variation classes appear to be two enmeshed distributions, differing primarily in the amount of scatter, with Ω_1 displaying smaller differences as expected. In this paper the two distributions are investigated further. Visualizations of the distributions are shown in Fig.5 and Fig.6, which are the approximative 3-D plots and contour plots of the distributions of the two classes respectively in the first two principal components based on the training data of the ORL database. From the plots it can be found that the relative orientation and scatter of the distributions are considerably different. And we can also observe that it seems not quite precise to simply assume either the within-class variations or the between-classes variations as a Gaussian distribution, whereas mixture models provide a promising solution.

As pointed out in Section 2, it is difficult to compare theoretically the unified ML rule with the unified MAP rule. Here we give a simple example of a 3-class classification problem to demonstrate that the unified ML can outperform the unified MAP sometimes. The three classes are represented by three normal random vectors: $\boldsymbol{X}_1 \sim N(M_1, \Sigma_1)$, $\boldsymbol{X}_2 \sim N(M_2, \Sigma_2)$ and $\boldsymbol{X}_3 \sim N(M_3, \Sigma_3)$, where

$$M_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix},$$

$$M_3 = \begin{bmatrix} -4 \\ 2 \end{bmatrix}, \ \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}.$$



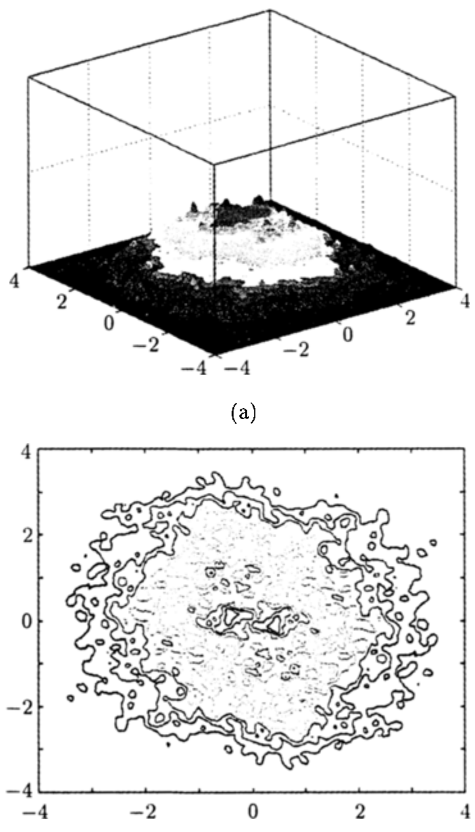
(a)

(b)

Fig.5. Approximative 3-D plot and contour plot of the distribution of the within-class variations class Ω_1 in the first two principal components.

100,000 random data points are generated for each class respectively as probe examples, in order to compare the classifiers. The gallery $\{y_1, y_2, y_3\}$ is set as $\{M_1, M_2, M_3\}$, and the results in classification obtained from the Bayes optimal classifier, the unified ML rule, and the unified MAP rule are shown in Table 3. The plot of the data points and the three decision boundaries are given in Fig.7.

The example can be regarded as an explanation for the fact that in all our experiments the unified



(a)

(b)

Fig.6. Approximative 3-D plot and contour plot of the distribution of the between-class variations class Ω_E in the first two principal components.

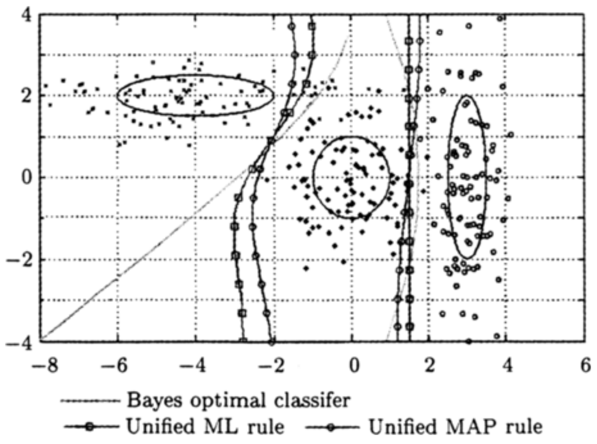


Fig.7. Comparison among the Bayes optimal classifier, the unified ML rule, and the unified MAP rule for a 3-class classification problem.

ML outperforms the unified MAP. Although when choosing which rule is still a difficult problem, it is noticeable that the computational cost of the unified ML is only half of the unified MAP's while their performances are almost the same.

Table 3. Results of Classification Obtained from the Bayes Optimal Classifier, the Unified ML Rule, and the Unified MAP Rule using the Synthetic Dataset

Approach	Recognition rate (%)
Bayes optimal classifier	95.50
Unified ML rule	94.05
Unified MAP rule	93.24

7 Conclusions and Future Work

In this paper we propose a novel technique of unified models as a promising solution to the problem of face recognition from a single example image for each target person. The unified models, trained by using an obtained training set with multiple samples per person, are used to recognize facial images from another disjoint database with a single sample per person. GMMs are applied to approximate the distributions of the variations between facial images, and a classifier combination method is exploited to improve the performance. The excellent performance of our method demonstrated by extensive experimental results confirms the following.

- The variations of each specific person's facial images, due to changes of pose, illumination, expression and so on, are rather similar to each other; therefore, the unified probabilistic models trained with samples from a group of known people for the significant similarity of the variations, can be generalized well to samples from another different groups of unknown people.
- GMMs, as a flexible approach for density estimation, can make efficient use of the information in face subspace.

It was pointed out that GMMs suffer from a drawback that each component is assumed as Gaussian, which is often violated in many natural clustering problems[41,42]. And accordingly a mixture of Independent Component Analysis (ICA) model[41−43] was proposed to be a substitution for GMMs. Therefore, an obvious extension of our work would employ ICA mixture models to model the unified variations of facial images.

References

[1] Wiskott Laurenz, Fellous Jean-Marc, Norbert Kruger *et al.* Face recognition by elastic bunch graph match-

ing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, 19(7): 775–779.

[2] Cootes T F, Taylor C J. Active shape model search using local grey-level models: A quantitative evaluation. *British Machine Vision Conference*, 1993.

[3] Lanitis A, Taylor C J, Cootes T F. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, 19(7): 743–756.

[4] Cootes T F, Edwards G J, Taylor C J. Active appearance models. *IEEE European Conf. Computer Vision*, 1998.

[5] Penev P, Atick J. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 1996, 7(3): 477–500.

[6] Turk M, Pentland A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991, 3(1): 71–86.

[7] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1997, 19(7): 711–720.

[8] Swets D L, Weng J. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1996, 18(8): 831–836.

[9] Etemad K, Chellappa R. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America*, 1997, 14: 1723–1733.

[10] Martinez A M, Kak A C. PCA versus LDA. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001, 23(2): 228–233.

[11] Philips P Jonathon. Support Vector Machines Applied to Face Recognition. In *Advances in Neural Information Processing Systems II*, Kearns M J, Solla S A, Cohn D A (eds.), MIT Press, 1999.

[12] Li Y, Gong S, Liddell H. Support vector regression and classification based multi-view face detection and recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[13] Johnsson K, Kittler J, Li Y, Matas J. Support vector machines for face authentication. *British Machine Vision Conference*, 1999.

[14] Liu Qingshan, Huang Rui, Lu Hanqing *et al.* Face recognition using kernel based Fisher discriminant analysis. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[15] Yang Ming-Hsuan, Ahuja Narendra, Kriegman David. Face recognition using kernel eigenfaces. *IEEE International Conference on Image Processing*, 2000.

[16] Yang Ming-Hsuan. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[17] Bartlett M, Sejnowski T. Independent components of face images: A representation for face recognition. *Annual Joint Symposium on Neural Computation*, 1997.

[18] Lawrence Steve, Yianilos Peter, Cox Ingemar. Face recognition using mixture-distance and raw images. *IEEE Int. Conf. Systems, Man, and Cybernetics*, 1997.

[19] Moghaddam Baback, Pentland Alex. Probabilistic visual learning for object detection. *IEEE International Conference on Computer Vision*, 1995.

[20] Gross Ralph, Yang Jie, Waibel Alex. Growing Gaussian mixture models for pose invariant face recognition. *International Conference on Pattern Recognition*, 2000.

[21] Er Meng Joo, Wu Shiqian, Lu Juwei *et al.* Face recognition with radial basis function (RBF) neural networks. *IEEE Trans. Neural Networks*, 2002, 13(3): 697–710.

[22] Dai Ying, Nakano Yasuaki. Recognition of facial images with low resolution using a Hopfield memory model. *Pattern Recognition*, 1998, 31(2): 159–167.

[23] Huang Fu Jie, Zhou Zhihua, Zhang Hong-Jiang *et al.* Pose invariant face recognition. *IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2000.

[24] Zhou Zhihua, Huang Fu Jie, Zhang Hong-Jiang *et al.* View-invariant face recognition based on neural network ensemble. *Journal of Computer Research and Development*, 2001, 38(10): 1204–1210. (in Chinese)

[25] Liu Chengjun, Wechsler Harry. Evolutionary pursuit and its application to face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000, 22(6): 570–582.

[26] Liu Chengjun, Wechsler Harry. Probabilistic reasoning models for face recognition. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1998.

[27] Wu Xiaojun, Yang Jingyu, Wang Shitong *et al.* A new algorithm for generalized optimal discriminant vectors. *J. Computer Science and Technology*, 2002, 17(3): 324–330.

[28] Guo Yuefei, Yang Jingyu. An iterative algorithm for the generalized optimal set of discriminant vectors and its application to face recognition. *Chinese Journal of Computer*, 2000, 23(11): 1189–1195. (in Chinese)

[29] Lai Jian Huang, Yuen P C, Feng G C. Face recognition using holistic Fourier invariant features. *Pattern Recognition*, 2001, 34(1): 95–109.

[30] Li S Z, Lu J. Face recognition using the nearest feature line method. *IEEE Trans. Neural Networks*, 1999, 10(2): 439–443.

[31] Samaria F, Fallside F. Face identification and feature extraction using Hidden Markov Models. *Image Processing: Theory and Applications*, Vernazza G, Elsevier, San Remo (eds.), Italy, June 1993.

[32] Liu Yi-Guang, Shen Li. Face image location using Hausdorff distance. *Journal of Computer Research and Development*, 2001, 38(4): 1204–1210. (in Chinese)

[33] Xiong Zhi-Yong, Shen Li. A general face image recognition system based on dual attribute graph. *Chinese J. Computer*, 2001, 24(7): 764–769. (in Chinese)

[34] Moghaddam Baback, Jebara Tony, Pentland Alex. Bayesian face recognition. *Pattern Recognition*, 2000, 33(11): 1771–1782.

[35] Philips P J, Wechsler H, Huang J S *et al.* The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 1998, 16(5): 295–306.

[36] Bishop C M. Neural Networks for Pattern Recognition. New York: Oxford University Press, 1995.

[37] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, B 39(1): 1–38.

[38] Kittler J, Hatef M, Duin R P W, Matas J. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1998, 20(3): 226–239.

[39] Tax D M J, Breukelen M, Duin R P W, Kittler J. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 2000, 33(9): 1475–1485.

[40] Olivetti & Oracle Research Laboratory. The Olivetti & Oracle Research Laboratory Face Database, http://www. uk.research.att.com/facedatabase.html.

[41] Lee Te-Won, Lewicki M S, Sejnowski T J. ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000, 22(10): 1078–1089.

[42] Roberts S J, Penny W D. Mixtures of independent component analysers. *Int. Conf. Artificial Neural Networks*, 2001.

[43] Choudrey R A, Robers S J. Variation mixture of Bayesian independent component analysers. *Technical Report*, 2001, http://www.robots.ox.ac.uk/~parg/publications.html.

Pin Liao received the B.S. degree in computer science from Nanchang University, Nanchang, P.R. China, in 1996 and the M.S. degree in pattern recognition and intelligent system from Beijing Institute of Technology, Beijing, P.R. China, in 1999. He is currently a Ph.D. candidate in Institute of Computing Technology, the Chinese Academy of Sciences. His current research interests include pattern recognition, computer vision, and neural networks.

Li Shen was born in Zhejiang, China, 1937. He graduated from the Department of Electrical Engineering, Zhejiang University, China, in 1959. Since then, he joined the staff of Institute of Computing Technology, the Chinese Academy of Sciences, where he is currently a professor. He is now an IEEE senior member. His research interests include soft computing, ASIC design, design for testability, and fault testing.