Enhanced maximum likelihood face recognition

X.D. Jiang, B. Mandal and A. Kot

A method to enhance maximum likelihood face recognition is presented. It selects a more robust weighting parameter and discards unreliable dimensions to circumvent problems of the unreliable small and zero eigenvalues. This alleviates the over-fitting problem in face recognition, where the high dimensionality and limited number of training samples are critical issues. The proposed method gives superior experimental results.

Introduction: The maximum likelihood (ML) method [1] is one of the best performing face recognition approaches. It decomposes the face image space into a principal subspace $\mathbb{F}$ and a residual subspace $\bar{\mathbb{F}}$. A main contribution of ML is the replacement of the unreliable eigenvalues in $\bar{\mathbb{F}}$ by a constant. This solves the singularity problem of the covariance matrix. However, ML does not well solve the over-fitting problem. The constant used to replace the eigenvalues in $\bar{\mathbb{F}}$ is estimated as the average eigenvalue over $\bar{\mathbb{F}}$. The high-dimensional face image and the limited number of training samples result in over-fitting problem because eigenvalues in $\bar{\mathbb{F}}$ are unreliable and so is their arithmetic average. We propose an approach that selects a more robust constant and discards some unreliable dimensions to circumvent this problem. It alleviates the over-fitting problem and hence boosts the accuracy of the ML face recognition approach.

Problems of ML method in face recognition: The difference image $\Delta = I_1 - I_2$, $I \in \mathbb{R}^n$, of two face images I_1 and I_2 in an n-dimensional space falls into an intrapersonal class Φ^I if I_1 and I_2 originate from the same person or into an extrapersonal class Φ^E if I_1 and I_2 originate from different persons. The likelihood measure $P(\Delta|\Phi^I)$ is modelled as an n-dimensional Gaussian density [1]. As the dimensionality n is very high compared to the number of the available training samples, $P(\Delta|\Phi^I)$ is estimated as the product of two independent marginal Gaussian densities, respectively, in $\mathbb{F}$ and $\bar{\mathbb{F}}$ as

$$P(\Delta|\Phi^I) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{m} y_k^2/\lambda_k\right)}{(2\pi)^{m/2}\prod_{k=1}^{m}\lambda_k^{\frac{1}{2}}}\right]\cdot\left[\frac{\exp\left(-\frac{1}{2\rho}\varepsilon^2(\Delta)\right)}{(2\pi\rho)^{(n-m)/2}}\right] \quad (1)$$

where λ_k is the kth largest eigenvalue of the intrapersonal covariance matrix, y_k is the kth leading principal component and

$$\varepsilon^2(\Delta) = \|\Delta\|^2 - \sum_{k=1}^{m} y_k^2 \quad (2)$$

The value of ρ is estimated by averaging the eigenvalues in $\bar{\mathbb{F}}$ [1] as

$$\rho = \frac{1}{n-m}\sum_{k=m+1}^{n}\lambda_k \quad (3)$$

The sufficient statistic for characterising the likelihood (1) is the Mahalanobis distance

$$d(\Delta) = \sum_{k=1}^{m}\frac{y_k^2}{\lambda_k} + \frac{\varepsilon^2(\Delta)}{\rho} \quad (4)$$

The first term is called distance-in-feature-space (DIFS) and the second term is called distance-from-feature-space (DFFS). Δ is classified into either Φ^I or Φ^E by evaluating the sum of these two distances (4).

The above ML approach for face recognition decomposes a high-dimensional image space into a reliable subspace $\mathbb{F}$ and an unreliable subspace $\bar{\mathbb{F}}$, and replaces the erratic eigenvalues, λ_k, $k > m$, in $\bar{\mathbb{F}}$ by a constant ρ. It makes the classifier less sensitive to noise to some extent. However, the high-dimensional face image and the limited number of the training samples in practice result in a large number of zeros and very small eigenvalues in $\bar{\mathbb{F}}$. This leads to a very small constant ρ by (3) compared to the eigenvalues in $\mathbb{F}$, i.e. $\rho \ll \lambda_k$, $k < m$. As a result, DFFS is much more heavily weighted than DIFS. Therefore, problems of over-fitting and noise sensitivity are still not well solved by this ML approach (3), (4).

Proposed approach: DFFS plays a critical role in classification [2, 3]. To look into the inside of the DFFS, we rewrite (4) as

$$d(\Delta) = \sum_{k=1}^{m}\frac{y_k^2}{\lambda_k} + \sum_{k=m+1}^{n}\frac{y_k^2}{\rho} = \sum_{k=1}^{n} w_k y_k^2 \quad (5)$$

where

$$w_k = \begin{cases} 1/\lambda_k, & k \le m \\ 1/\rho, & m < k \le n \end{cases} \quad (6)$$

Equations (5) and (4) are equivalent as it is not difficult to see

$$\varepsilon^2(\Delta) = \sum_{k=m+1}^{n} y_k^2 \quad (7)$$

Thus (5) is a weighted distance with a weighting function w_k given by (6).

If the training dataset consists of l images from ρ persons, the rank of the intrapersonal covariance matrix is r where $r \le \min(n, l-p)$. In the practical application of face recognition, $l-p$ is usually much smaller than n, which leads to $\rho \ll \lambda_k$, $k < m$, or a big jump of the weighting function w_k at $k = m+1$. As a result, the distance components in $\bar{\mathbb{F}}$ are over-weighted. For a clear illustration we plot λ_k of a typical real face training dataset in Fig. 1. The constant ρ of (3) and the weighting function w_k of (6) are also shown in Fig. 1. We see an undue big jump of the weighting function in $\bar{\mathbb{F}}$ from $\mathbb{F}$. The zero and very small eigenvalues in $\bar{\mathbb{F}}$ caused by the limited number of training samples are the culprits of the undue overemphasis in this subspace.

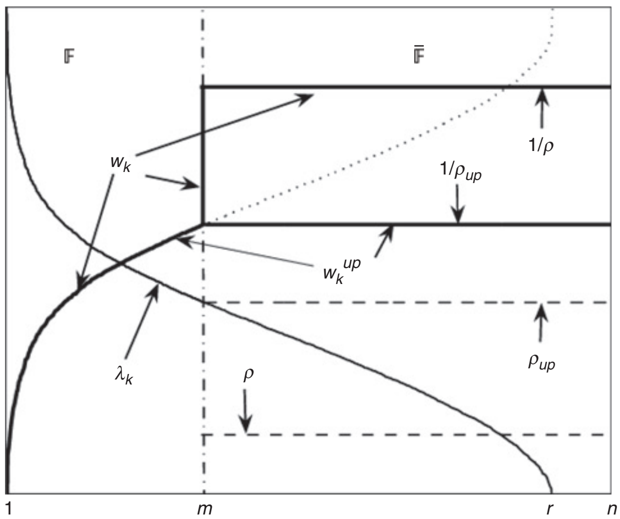

Fig. 1 *Logarithm scale plot of eigenvalues, the constants (3) and (8), and the weighting functions (6) and (9) in the principal and residual subspaces*

It is supposed that the statistics obtained in $\mathbb{F}$ are reliable and most of the statistics in $\bar{\mathbb{F}}$ are unreliable so that the eigenvalues in $\bar{\mathbb{F}}$ are replaced by a constant ρ. As most eigenvalues in $\bar{\mathbb{F}}$ are unreliable and so is their arithmetic average, we choose the upper bound of eigenvalues in $\bar{\mathbb{F}}$ as the constant ρ_{up}:

$$\rho_{up} = \max\{\lambda_k | k > m\} \quad (8)$$

Fig. 1 shows ρ_{up} and the resulting weighting function w_k^{up} given by

$$w_k^{up} = \begin{cases} 1/\lambda_k, & k \le m \\ 1/\rho_{up}, & m < k \le n \end{cases} \quad (9)$$

There is no undue big jump in w_k^{up}. The new weighting function w_k^{up} suppresses the contribution of the residual subspace to the distance. This alleviates the over-fitting problem caused by the small number of training samples relative to the high dimensionality of the face images. With this weighting function, the distance measure (4) of ML is modified as

$$d(\Delta) = \sum_{k=1}^{m}\frac{y_k^2}{\lambda_k} + \frac{\varepsilon^2(\Delta)}{\rho_{up}} \quad (10)$$

To decouple the unreliable statistics in some dimensions from the image space, we further propose to apply the principal component analysis (PCA) to reduce the data dimensionality before applying the above approach. The PCA works on the total scatter matrix of the training samples. It extracts a low-dimensional subspace corresponding to the $l-p$ largest eigenvalues from the high-dimensional image space, $\mathbb{R}^n\mathbb{R}^{l-p}$, $l-p \ll n$. The proposed enhanced ML approach is then applied in this reduced space $\mathbb{R}^{l-p}$.

Experimental results: 2388 face images comprising 1194 persons (two images per person) were picked from the FERET database [4]. Images were cropped into the size of 33×38 and pre-processed following the CSU evaluation system [5]. Three experiments were conducted using 500, 1000 and 1400 training samples, respectively. The remaining images constitute the testing sets. The recognition rate is the percentage of the correct top 1 match on the testing set. We compute the average over the top 10 recognition rates of the 20 different m values tested. Table 1 records the results. It shows that the proposed ML approach with the upper bound ρ_{up} consistently outperforms the conventional ML method. The proposed ML approach working in the reduced space further boosts the recognition performance.

Table 1: Average recognition rate (per cent)

	Number of training images		
Method	500	1000	1400
ML with ρ in $\mathbb{R}^n$	89.76	90.44	89.70
ML with ρ_{up} in $\mathbb{R}^n$	91.03	93.03	93.64
ML with ρ_{up} in $\mathbb{R}^{l-p}$	92.90	95.94	96.46

Conclusions: Owing to the high image dimensionality and the limited number of training data, the conventional ML algorithm is sensitive to noise and over-fits the training data. The replacement of the average eigenvalue by the upper-bound of the eigenvalues in the residual space alleviates this problem and hence boosts the recognition accuracy. The proposed approach working in the reduced space further improves the generalisation. The higher accuracy of the proposed approach is substantiated by the experiments.

30 June 2006
Electronics Letters online no: 20062035
doi: 10.1049/el:20062035

X.D. Jiang, B. Mandal and A. Kot (*School of EEE, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798*)

E-mail: exdjiang@ntu.edu.sg

References

1 Moghaddam, B., Jebara, T., and Pentland, A.: 'Bayesian face recognition', *Pattern Recognit.*, 2000, **33**, (11), pp. 1771–1782
2 Cevikalp, H., Neamtu, M., Wilkes, M., and Barkana, A.: 'Discriminative common vectors for face recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (1), pp. 4–13
3 Moghaddam, B.: 'Principal manifolds and probabilistic subspace for visual recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (6), pp. 780–788
4 Phillips, P.J., Moon, H., Rizvi, S., and Rauss, P.: 'The FERET evaluation methodology for face recognition algorithms', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (10), pp. 1090–1104
5 Beveridge, R., Bolme, D., Teixeira, M., and Draper, B.: 'The CSU face identification evaluation system users guide: Version 5.0'. Technical Report: http://www.cs.colostate.edu/evalfacerec/data/normalization.html, 2003