A New Approach to Bayesian Method for Face Recognition

Len Bui, Dat Tran, Xu Huang and Girija Chetty
Faculty of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia

Abstract - *In this paper, we propose a new approach to Bayesian subspace method for face recognition. We review this method and point out some weak points in its assumptions and propose a practical solution to overcome those weak points. In addition, we present an efficient way to estimate the high dimensional Gaussian density function.*

Keywords: Face recognition, Subspace, Bayesian, Gaussian density function

1 Introduction

Face recognition is one of the successful applications of computer vision. There have been recently advances in accuracy and time performance due to new algorithms and computer development.

A popular method used in face recognition is linear subspace due to its high stability and performance compared with other methods. The first subspace method is Principal Component Analysis (PCA) proposed by Turk and Pentland [1]. This method is based on the work done by Kirby and Sirovich on face representation and analysis [2]. Since then, it has become the de factor method in face recognition. However, it is completely based on distances and there is no clear explanation for choosing metrics; therefore, there is no strong mathematical foundation to support this method.

To overcome this problem, a variety of statistical based methods was proposed. For example, Belhumeur et al. [3] proposed Linear Discriminant Analysis (LDA). This method aims at finding a subspace that can maximize variances between classes and maximize variances in each class compared with PCA where a subspace can only capture the most variance of data set. The PCA method reduces only the dimensions of space but there is no strong evidence to show that it can improve recognition rate, whereas the LDA method could show an improvement in practice. The Bayesian method proposed by Moghaddam et al. [4, 5] is based on Bayes theorem and is the best method in independent tests performed on 1996 FERET dataset [6, 7]. The idea of the Bayesian method is to use two probability density distributions of intra-variation and extra-variation subspaces to classify an unknown object. In 2003, Wang and Tang [8] attempted to unify PCA, LDA and Bayesian subspaces and presented a good unified model to explain three kinds of subspaces. However, they did not include the extra- variation subspace in their solution to improve the recognition rate although this information is also very important according to Bayes theorem. Other researchers [9] attempted to discover the structure of that extra-variation and reported better recognition results, however, there is no independent evaluation about this method and the most important factor is that it needs a lot of computing resources.

In this paper we present a new approach to the Bayesian method. The challenge in face recognition is that there are many classes but only a few samples are available for each class. To deal with this challenge, two kinds of variations have been investigated: intra-variation for one person and extra-variation between two people. However, there are not sufficient data to estimate distributions for those variations and hence equal intra-variations and equal extra-variations are assumed to have sufficient data to estimate their distributions.

On the other hand, based on experimental evaluations for the Bayesian method, it is shown that this method does not always provide better performance for face recognition [?]. We have faced on this issue and found a new similarity score that can improve the performance. This score is called soft similarity score and its best value can be found through a cross validation process.

The remainder of the paper is organized as follows. In the next section, we will present the Bayesian method and our proposed approach. We then present in Section 3 experiments to evaluate our proposed approach on two public face data sets which are FERET and AT&T. Finally, Section 4 gives a brief review of our proposed approach and future work.

2 Bayesian Method for Face Recognition

In this section, we present the conventional Bayesian method and its distribution estimation for face recognition.

2.1 Distribution estimation

It is hard to determine density distribution for random variables in multi-dimensional space. A mixture of Gaussian distributions is assumed as their mean vector μ and covariance matrix Σ can be estimated from a training set [10-13].

$$p(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (1)$$

where $p(.|.)$ is the density distribution function and D is the dimension. To efficiently compute the value of distribution, the following equation is used [14]

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{V},\rho) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{\frac{M}{2}}\prod_{i=1}^{M}\lambda_i^{\frac{1}{2}}} \frac{\exp\left(-\frac{\epsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{\frac{D-M}{2}}} = p_{span(\mathbf{V})}(\mathbf{x})\, p_{span(\bar{\mathbf{V}})}(\mathbf{x}) \quad (2)$$

where

- $\boldsymbol{\Lambda}$ be eigenvalues of $\boldsymbol{\Sigma}$
- $\mathbf{V}$ be eigenvectors of $\boldsymbol{\Sigma}$
- ρ be an average of remainder eigenvalues
- $\mathbf{y} = \mathbf{V}(\mathbf{x}-\boldsymbol{\mu})$ and $\epsilon^2(\mathbf{x}) = \|\mathbf{x}-\boldsymbol{\mu}\|^2 - \|\mathbf{y}\|^2$

In Equation 2, the density distribution is represented as a product of two distributions in subspace $\mathbf{V}$ and $\bar{\mathbf{V}}$, respectively. As mentioned in the previous section, the second term for subspace $\bar{\mathbf{V}}$ was ignored in most of publications. Recent study such as [15] has included this second term. Moghaddam et al. [4] uses the following to estimate ρ

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{D}\lambda_i \quad (3)$$

In practice computing exactly high eigenvalues is intractable due to numerical computing errors. Therefore, fitting techniques are used to estimate these eigenvalues. Bishop [10] suggests the use of an expectation-maximization (EM) algorithm, however the computation cost is very high.

2.2 Similarity score

Due to insufficient data to estimate each variation, some assumptions have been made in the Bayesian method.

Assumption 1: Density distributions of variations for all individuals are the same.

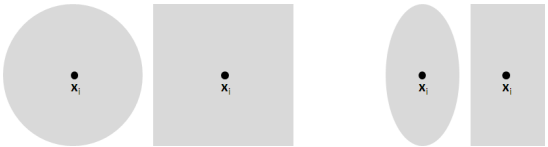

Figure 1 a) L_2, L_1 metrics b) Mahalanobis metrics

Figure 2 a) correct classification using L_2 b) incorrect classification using L_2

In statistical view, each metric is corresponding to an assumption ab out the form of variation distribution. For example, metric L_2 assumes it be a hyper-sphere "shape"; metric L_1 assumes it be hyper-cube "shape" (see Fig. 1a). In recent studies, Mahalanobis metric is used for face recognition (see Fig. 1b).

Obviously, the distribution of variation will decide which metric is used for classification. For example, as seen in Fig. 2a, the distribution has a hypersphere "shape" and hence L_2 metric can provide correct classification to $\mathbf{x}$. However the distribution in Fig. 4b is not a hyper-sphere "shape", L_2 cannot provide correct classification to $\mathbf{x}$.

Previous studies performed empirical tests to compare metrics and did not mention how a metric was theoretically chosen. In the Bayesian method, Moghaddam et al. [?] pointed out some weak points in previous studies. A variation between two images $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $\Delta = \mathbf{x}_i - \mathbf{x}_j$. Two classes which are intra-variation class Ω_I and extra-variation class Ω_E were defined and their density distributions were estimated if a training data set is provided. To simplify the distribution estimation process, the following assumptions on two density distributions are made.

Assumption 2: Intra-variation has Gaussian distribution

Assumption 3: Extra-variation has Gaussian distribution

With a variation Δ of two images, maximum likelihood value $p(\Delta|\Omega_I)$ or maximum posterior value $P(\Omega_I|\Delta)$ can be used as similarity scores to measure the similarity between two images. In practice, $\ln p(\Delta|\Omega_I)$ and $\ln P(\Omega_I|\Delta)/P(\Omega_E|\Delta)$ are used instead.

Maximum likelihood score

$$s_{ML}(\Delta) = \sum_{i=1}^{k}\frac{y_{I,i}^2}{\lambda_{I,i}} \quad (4)$$

or

$$s_{ML}(\Delta) = \sum_{i=1}^{k}\frac{y_{I,i}^2}{\lambda_{I,i}} + \frac{\epsilon^2(\Delta)}{\rho_I} \quad (5)$$

Maximum posterior score

$$s_{MAP}(\Delta) = \left[\sum_{i=1}^{k}\frac{y_{I,i}^2}{\lambda_{I,i}} + \frac{\epsilon^2(\Delta)}{\rho_I}\right] - \left[\sum_{i=1}^{k}\frac{y_{E,i}^2}{\lambda_{E,i}} + \frac{\epsilon^2(\Delta)}{\rho_E}\right] \quad (6)$$

3 The Proposed Approach to Bayesian Method

3.1 Determine ρ

We suggest a simple and closed-form way to compute the average of eigenvalues ρ.

Lemma 1

Given data set $\mathbf{X} = \{\mathbf{x}_i; i = 1,..., N\}$

If it has Gaussian distribution $\mathbf{\Sigma} = \rho\mathbf{I}$, where **I** is the unity matrix

Then

$$\rho = \frac{1}{ND}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu}) \quad (7)$$

Proof

Compute log likelihood for the data set

$$\ln p(\mathbf{X} \mid \boldsymbol{\mu}, \rho) = -\frac{ND}{2}\ln 2\pi - \frac{ND}{2}\ln\rho - \frac{1}{2\rho}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu})$$

Take derivative and set it to 0

$$\frac{\partial}{\partial\rho}p(\mathbf{X} \mid \boldsymbol{\mu}, \rho) = -\frac{ND}{2\rho} + \frac{1}{2\rho^2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu}) = 0$$

We have

$$\rho = \frac{1}{ND}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T(\mathbf{x}_i - \boldsymbol{\mu})$$

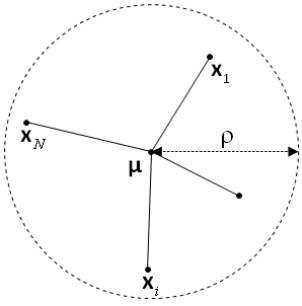

Figure 3. Distances of points of training set with Gaussian distribution $\mathbf{\Sigma} = \rho\mathbf{I}$

The Lemma 1 shows that the information of distances from data points to the mean point is sufficient for computing ρ (see Fig. 3).

Recall that two subspaces $\mathbf{V}$ and $\overline{\mathbf{V}}$ are orthogonal. Therefore their distributions are independent and can be estimated separately. Distances of data points in subspace $\overline{\mathbf{V}}$ can be computed from those in subspace $\mathbf{V}$ using Pythagorean theorem (see Fig. 4).

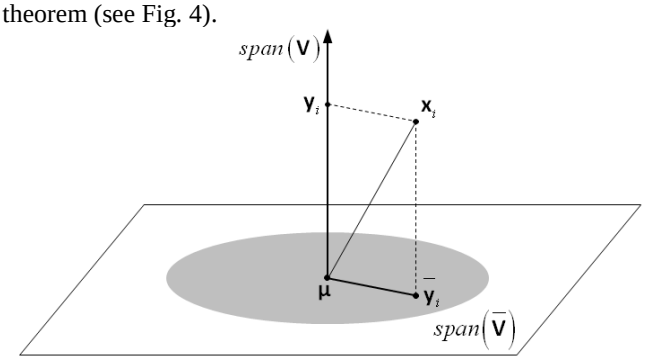

Figure 4. Distances of $\mathbf{x}_i$ in subspace $\mathbf{V}$ and $\overline{\mathbf{V}}$

Algorithm to compute ρ

Input: training data set $\mathbf{X} = \{\mathbf{x}_i; i = 1,..., N\}$

Output: ρ

Step 1: Determine subspace $\mathbf{V}$

Step 2: Compute distance to mean in subspace $\mathbf{V}$

Step 3: Compute distance to mean in subspace $\overline{\mathbf{V}}$

Step 4: Compute ρ using Equation 7

3.2 Similarity score

Assumption 2 in Section 2.2 on the distribution of intra-variation could provide better recognition results. However, the assumption on the distribution of extra-variation is not persuading us. In fact, in our first experiment, we repeated the experiments presented in [5-7] but we could not get the same results. We assume that the contributions of intra-variation and extra-variation are not equal as seen in Equation 6. We propose a scale factor w for the second term to adjust its contribution. The proposed score is of the form

Proposed soft similarity score

$$s_{soft}(\Delta) = \left[\sum_{i=1}^{k}\frac{y_{I,i}^2}{\lambda_{I,i}} + \frac{\eth^2(\Delta)}{\rho_I}\right] - w\left[\sum_{i=1}^{k}\frac{y_{E,i}^2}{\lambda_{E,i}} + \frac{\eth^2(\Delta)}{\rho_E}\right], w \geq 0 \quad (8)$$

We can use cross-validation task to find the optimal value for this scale factor w.

It can be seen that our proposed soft similarity score has a close relationship to those scores in Section 2.2. For example, if $w = 0$ then $s_{soft} = s_{ML}$ and if $w = 1$ then $s_{soft} = s_{MAP}$.

4 Data sets, experiments and results

We evaluate our approach based on two well-known datasets which are FERET and AT&T. We use FERET protocol to evaluate our experiments.

4.1 FERET protocol for face recognition

The goal of FERET evaluation protocol is to provide a standard method to access an algorithm. The evaluation design cannot be too hard or too easy. In the protocol, an algorithm is given two sets of images which are target set T (training set) and query set Q (testing set). The algorithm reports the similarity score $s(\mathbf{q}_i, \mathbf{t}_j)$ between all query images $\mathbf{q}_i$ in the query set Q and all target images $\mathbf{t}_j$ in the target set T. In face recognition, the query is not always "Is the top match correct?" but "Is the correct answer in the top *n* matches?" like Google search engine responds our query by listing more possible answers for each query. For our method, we assume that a larger similarity score implies a closer match. Finally, the performance statistics are reported as cumulative match scores.

4.2 Data sets

AT&T dataset was taken at University of Cambridge. It includes 400 images of 40 individuals and each person has 10 images in different illumination, pose and expression. Grayscale FERET is standard data set and is widely used for evaluation. There are about 14000 images of more than 1000 individuals.

4.3 Experiments

4.3.1 Preprocessing

There was no image processing task for AT&T data set. For FERET data set, we used the data set conducted in 1996 for independent tests including gallery set **fa** and four probe sets **fb**, **fc**, **dup1** and **dup2**. Based on the information containing in ground truth files, we cropped, aligned and masked images to normalize images. We used procruste analysis for aligning images based on the positions of mouths and eyes. Fig. 5 shows some sample images of AT&T and FERET after cropping and aligning.

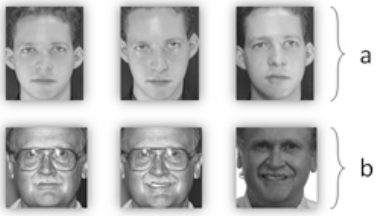

Figure 5. a) images from AT&T b) images from FERET

4.3.2 Building two subspaces from training set

For AT&T data set, we divided the training data set into four partitions and each partition included 200 images for training and gallery sets and the other 200 images for probe sets. There were 1000 intra-variation pairs and 39000 extra-variation pairs in those training sets. From these pairs of data sets, we could find the parameters in Equation 2 or build two subspaces intra and extra-subspaces (see more details in [10-14]). For FERET dataset, we randomly selected a subset including 250 **fa** images and 250 **fb** images from the training set in FERET CD-ROM. There are 1230 intra-variation pairs and 248770 extra-variation pairs.

We used Equation 2 for subspace projections and distribution computations. For the AT&T dataset, we kept 160 first eigenvectors for intra-variation subspace and 199 for extra-variation subspace. For the FERET dataset, we kept 200 first eigenvectors for intra-variation and the same for extra-variation.

4.3.3 Experiments

To evaluate our approach with state-of-the-art approaches, we conducted six algorithms on each data set. They were raw matching method (Baseline) using metric L_2, Eigenface method (Eigenface) using PCA to extract feature vectors and using L_2 + Mah [16] for computing similarity scores, Bayesian method using scores s_{ML} in Equations 5 and 6 (ML1 and ML2) and s_{MAP} (MAP), and our proposed scores s_{soft} (Soft). Note that the best results could be obtained by selecting an appropriate number of Eigenfaces for Eigenface method. In our experiments, they were 200 for FERET and 199 for AT&T. To validate our proposed score, we used three partitions as validation sets and the remaining as test set. The best accuracy rate is corresponding to w = 0 for AT&T data set (see Fig. 6) and w = 7.25 for FERET (see Fig. 11). It means that s_{soft} becomes s_{ML} in AT&T case. The detailed experimental results on AT&T are presented in Table 1 and Figure 6. The detailed experimental results on FERET are presented in Figures 7-11.

Table 1. Experimental results on AT&T data set

Method	Metric	P1	P2	P3	P4	avg
Baseline	L_2	95	95	93.5	93.5	94.25
Eigenface	L_2+ Mah	89.5	90.5	87	86.5	88.38
Bayesian	ML1	94	95.5	94.5	93	94.25
Bayesian	ML2	94	95.5	94.5	93	94.25
Bayesian	MAP	52.5	42.5	56	57.5	52.13
Bayesian	Soft	94	95.5	94.5	93	94.25

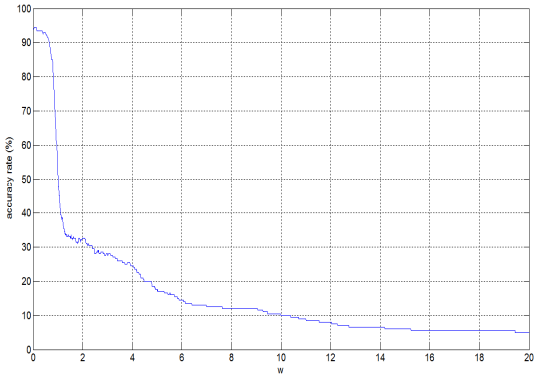

Figure 6. Average accuracy rates corresponding with soft score w ranging from 0 to 20

4.3.4 Discussion

Our experimental results on FERET using traditional Bayesian scores are quite similar to some public face recognition evaluation systems such as "The CSU face identification evaluation system" although our implementation is based on MATLAB platform and independently written. As mentioned above, the results are slightly different from the results published in 1996 FERET independent tests. For example, the rank-1 result on **fb** is 95% in 1996 FERET test; but it is 85% in our experiments and in other reports [15, 17]. Because Moghaddam and his colleagues have not published more details on their implementation, we suggest two explan-

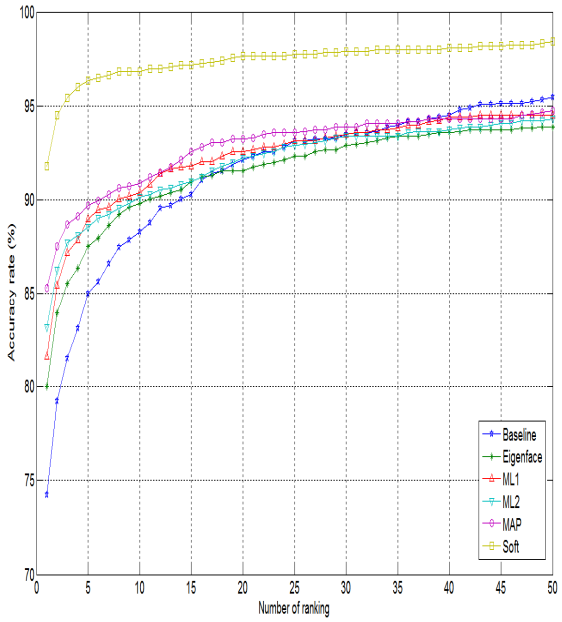

Figure 7. Experimental results on probe set **fb**

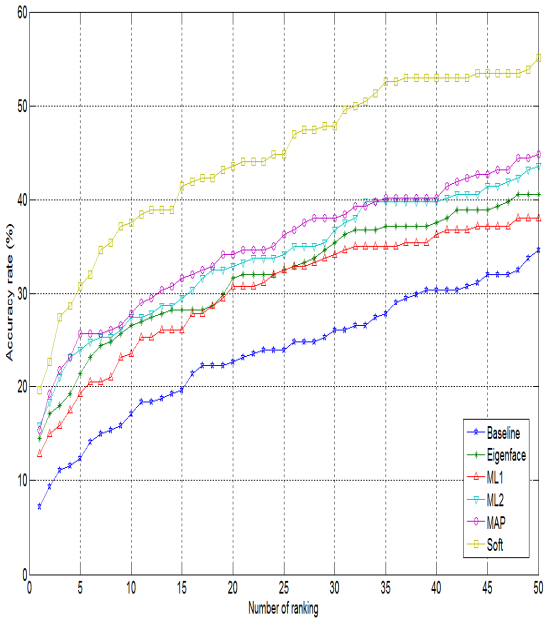

Figure 9. Experimental results on probe set **dup2**

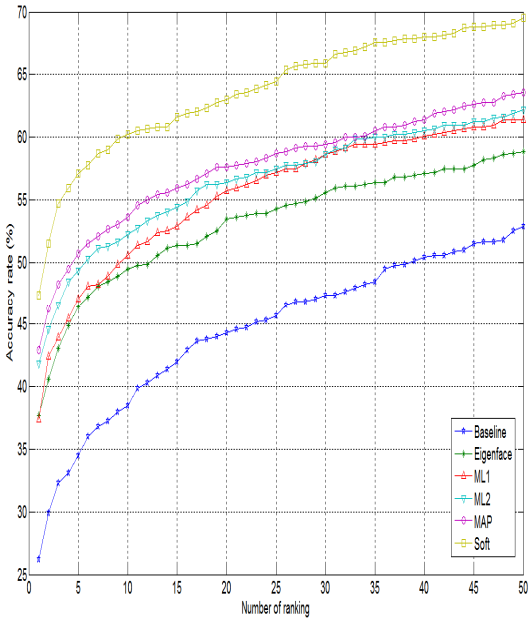

Figure 8. Experimental results on probe set **dup1**

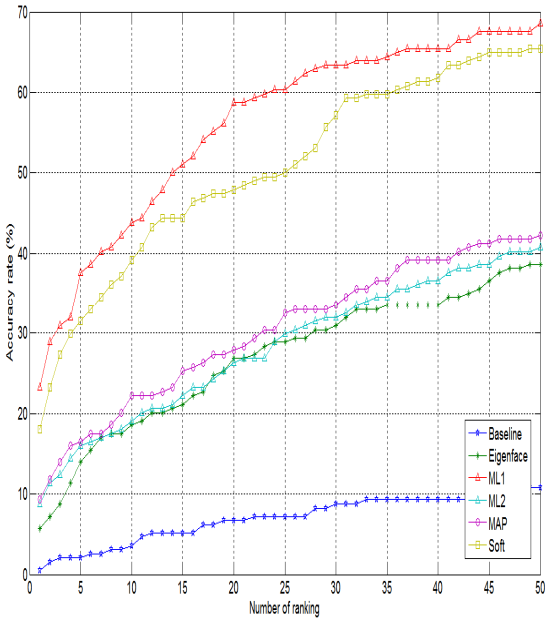

Figure 10. Experimental results on probe set **fc**

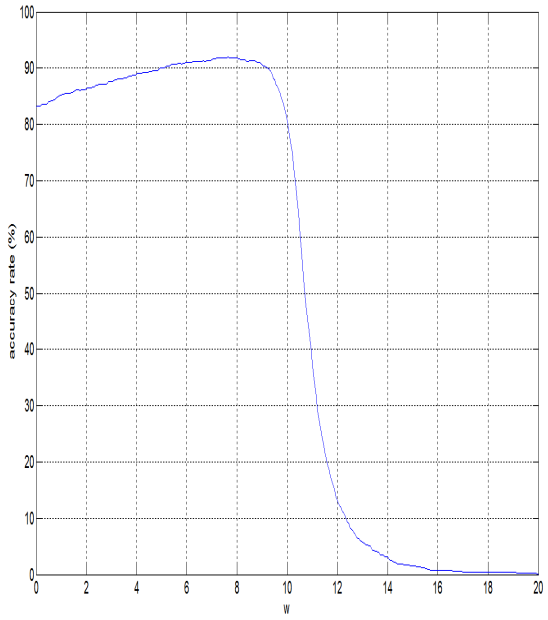

Figure 11. Accuracy rates corresponding with soft score w ranging from 0 to 20 on validation set **fb**

-ations for this issue. The first one is numerical issues. If the original formulas of Bayesian scores [4, 5] are used to compute probabilities, it can lead to unstable results (See more details in [17]). The second one is that the quality of face recognition system depends not only on classification method (Bayesian method) but also on other tasks in the system such as the quality of preprocessing tasks.

According to our experiments, it proves that the Assumption 3 is not good. Table 1 and Fig. 6 present experiments on AT&T and show that the MAP score gives the worst result (52.5%) compared with the best result (94.25%). Obviously, the distributions of extra-variations are not Gaussian because the training sets contain individual varition and strong different variations such as illumination variations and pose variations. Therefore, these distributions can not improve accuracy rates. Figures 7 – 11 present experiments on FERET and show that the MAP score gives good results (85% rank-1 for proble set **fb** in Fig. 7) but our approach can achieve the best results (92% rank-1 for probe set **fb** Fig. 7). It proves that the contribution of extra-variation to the similarity score is not equal to that of intra-variation. To get good soft scores, we used validation set to find optimal value for w. Fig. 11 shows the accuracy rate curve for rank-1 with respect to w. The maximum value 92% is corresponding to w = 7.25. In some experiments [4, 5, 8], researchers scaled down normal images to small images. It can make the extra-variation distributions to approach Gaussian distributions. In these cases, using extra-variation distributions would improve the accuracy rates.

We would also discuss on Probabilistic Eigenfaces method proposed by Shakhnarovich and Moghaddam [5] in this paper, however they did not present their experiments related to this approach. In fact, there is no evidence to persuade that it will increase the accuracy rates. In our experiments, it achieved only 48% for probe set **fb** and the lowest for the other proble sets. In our opinions, this method is appropriate to distinguish face or not face – face detection or objection detection.

5 Conclusions

We have reviewed conventional Bayesian, given a clear picture about some confused experimental results published previously and presented some weak points. We have also proposed a soft similarity score to deal with this problem and suggested a closed-form formula to estimate very high dimensional Gaussian distributions.

6 References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of cognitive neuroscience, vol. 3, pp. 71-86, 1991.

[2] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 12, pp. 103-108, 1990.

[3] P. N. Belhumeur, et al., "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, pp. 711-720, 1997.

[4] B. Moghaddam, et al., "Bayesian face recognition," Pattern Recognition, vol. 33, pp. 1771-1782, 2000.

[5] S. Z. Li and A. K. Jain, Handbook of face recognition: Springer, 2005.

[6] P. J. Phillips, et al., "The FERET evaluation methodology for face-recognition algorithms," in Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, 1997, pp. 137-143.

[7] P. Phillips, et al., "The FERET evaluation methodology for face-recognition algorithms," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, pp. 1090-1104, 2002.

[8] X. Wang and X. Tang, "A unified framework for subspace face recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, pp. 1222-1228, 2004.

[9] F. Perronnin and J. Dugelay, "Discriminative face recognition," 2003, pp. 446-454.

[10] C. M. Bishop, Pattern recognition and machine learning: Springer New York., 2006.

[11] R. O. Duda, et al., Pattern classification: Wiley New York, 2001.

[12] K. Fukunaga, Introduction to statistical pattern recognition: Academic Pr, 1990.

[13] A. Webb, Statistical pattern recognition: A Hodder Arnold Publication, 1999.

[14] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, pp. 696-710, 1997.
[15] J. Beveridge, et al., "The CSU face identification evaluation system," Machine vision and applications, vol. 16, pp. 128-138, 2005.
[16] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," PERCEPTION-LONDON-, vol. 30, pp. 303-322, 2001.
[17] M. L. Teixeira, "The bayesian intrapersonal/extrapersonal classifier," Colorado State University, 2003.