**Proceedings of the 2007 IEEE Symposium on**
**Foundations of Computational Intelligence (FOCI 2007)**

# Classifier Discriminant Analysis for Face Verification based on FAR-score normalization

Chengbo Wang, Yongping Li, Hongzhou Zhang, Lin Wang
Shanghai Institute of Applied Physics, Chinese Academy of Sciences
2019 Jialuo Road, Jiading, Shanghai, P.R. CHINA

***Abstract*- In this paper, we propose a novel matching score normalization method for multi-classifiers based on their false acceptance rate (FAR) scores to make fusion operable at the matching level. The classifier discriminant analysis (CDA) is put forward and implemented to single out the best score from the appreciate classifier as the fusion output. Experimental results of face verification on two public available face databases (ORL, XM2VTS) show our approach's efficiency and effectiveness when compared with the conventional fusion methods.**

## I. INTRODUCTION

Research interests and activities in face recognition and verification have been increased significantly during the past decades. Although many popular methods such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA) and two-dimensional LDA (2dLDA) [1-4] have been proposed, none of them can give plausible results when used solely due to the changes of the illumination, pose, etc. A possible way to get improving performance is to combine different methods together. To do this, one must find ways to combine the various data from different algorithms at feature extraction, matching score and decision levels [5]. Within those three fusion levels, researchers show special interest in the matching score level because it can utilize more information comparable with that in the decision level, but avoid the complexity of combination of various features from different algorithms.

The matching scores are the outputs of the classifiers measuring the similarities of the testing sample to the claimed class. There are two approaches for assembling the scores obtained from different matchers [6]. One is formulated as classification-based fusion, for which a new feature vector is created by concatenating scores from individual matchers, and then a classifier such as neural network [12], k-NN classifier[13] or decision tree [5] can be used to judge whether accepting or rejecting the claim. The drawback of this fusion strategy is that classifiers are sensitive to inputted data, and give total wrong decisions under noisy input condition. Another is the combination-based fusion [6], in this case, the scores are employed to compute a single scalar score which is compared with a certain threshold to make the final decision. To ensure a meaningful combination, different scores from various classifiers must be normalized into the homogeneous domain.

In this paper, we focus on the combination approach, i.e., normalizing scores from different matchers and combining them together. Min-Max, z-score and tanh methods are adopted in [6] to normalize the scores, and then combine the normalized scores with the sum, min, med and max rules. Although some methods of them are robust and efficient, there are deficiencies. At the normalizing stage, the scores normalized with a single function which does not follow all distributions of scores from different classifiers will introduce different normalization errors. When combining the normalized scores, the fusion rules focus on scores of claimed class from the different classifiers without analysis the discriminant of the classifiers.

A novel normalization method is proposed which adopts the FAR-score curve as the normalization function. It can be applied to any type of distributed scores. In the training stage, the FAR-curve of each classifier which will be used later as the normalization function can be properly obtained because there are always enough negative instances to compute FARs. Thus every classifier has its own normalization function that follows the distribution of the scores and FARs. When score is normalized by the FAR-score curves, the normalized score is the probability of the classifier to accept a negative instance. After all scores from different classifiers are normalized, the fusion rules such as sum, min, med and max can be used to compute a single scalar to judge an acceptance or rejection of the claim. In order to enhance the performance of the combination of the normalized scores, we put forward a method of classifier discriminant analysis (CDA) based on the normalized FAR-score to single out the results from the appreciate classifier as the final scores, which improves the performance significantly.

The rest of the paper is organized as follows: Section II presents the method of normalizations with the FAR-score curve and in Section III the classifier discriminant analysis (CDA) is described; the experimental results is shown in the Section IV and at last we draw the conclusion in section V.

## II. CONVERT THE MATCHING SCORE INTO THE FALSE ACCEPTANCE RATE

Score normalization for multi-classifier fusion refers to transform the various scores obtained by different classifiers into a common meaningful domain. Distinct matcher produces score diverse in numerical range and meaning, and the evaluation standards vary with different kinds of score. It is necessary to normalize the scores into the same domain prior to combination. Two factors should be considered for score

normalization. One is that the normalized score should keep the discriminating information as much as possible; the other is that the normalization method should be adaptive to various scores from different matchers.

In the context of face recognition or verification, persons who should be recognized by classifier are referred as clients whilst those should be rejected by the classifier are identified as impostors. In practice, classifier outputs a matching score $s$ to reflect the similarity between the testing sample $Z$ and the claimed class and $s$ can be generally modeled as [7]:

$$s = f[P(genuine \mid Z)] + \eta(Z) \tag{1}$$

where $f$ is a monotonic function and $\eta()$ is the bias of the classifier. If $\eta()$ is assumed to be zero, the posteriori P(*genuine*|*Z*) can be estimated by *P*(*genuine*|*s*) [5], thus the problem reduces to compute the *P*(*genuine*|*s*) only [8]:

$$P(genuine \mid s) = \frac{p(s \mid genuine) * P(genuine)}{p(s)} \tag{2}$$

Where p(s)=p(s | genuine )P(genuine) + p(s | impostor ) P( impostor ). Before computing *P*(*genuine*|*s*), conditional density of *p*(*s*|*genuine*) and *p*(*s*|*impostor*) should be known in advance, so do the prior probabilities of genuine user *P*(*genuine*) and impostor *P*(*impostor*). However, there are difficulties in applying this model: firstly the $\eta(Z)$ does not equal to zero in most of classification system, secondly we don't know the distribution of the *p*(*s*|*genuine*) and *p*(*s*|*impostor*) beforehand, furthermore, we even have not enough training samples to estimate them.

In [5] the authors recommend to normalize scores with a certain function such as *z*-score and tanh functions as below:

$$n = \frac{s - mean(S)}{std(S)} \tag{3}$$

$$n = \frac{1}{2}[\tanh(0.01\frac{(s - mean(S)}{std(S)}) + 1] \tag{4}$$

*n* is the normalized score, *mean*() and *std*() denote the arithmetic mean and standard deviation operators, *S* is the set of all scores from the classifier. Although these functions use the statistic values such as the means and variances, they do not follow the distributions of the scores of different classifiers. When the function is used to normalize distinct distributed scores, it causes diverse normalization errors due to the deviation between the score distributions and the function.

A novel normalization method that converts scores into false acceptance rate is introduced here. When studying a typical Receiver Operating Characteristic (ROC) curve of a classifier, two kinds probabilities relate with the scores can be drawn: the FAR-score curve and the FRR-score curve as shown in Fig.1. The false acceptance rates and the false rejection rates are functions of threshold (denoted as *h*). Two functions can be written as below without assumptions of any distributions that *s* should observe:

$$f_{far}(h) = P(genuine \mid impostor, s < h) = \frac{\text{false positives}}{\text{negative instances}} \tag{5}$$

$$f_{frr}(h) = P(impostor \mid genuine, s > h) = \frac{\text{false negatives}}{\text{positive instances}} \tag{6}$$

In the real work of training, we can not get the exact forms of $f_{far}()$ and $f_{frr}()$, but a series of *h* can be used to calculate $f_{far}(h)$ and $f_{frr}(h)$, then both of the functions can be computed by interpolation. Once these two functions are available, a score can be converted into FAR or FRR. The two curves are monotonic functions with range of [0, 1] and only depend on the distribution between the samples and the scores. Both of the two curves can be used to normalize the scores of the classifier. However, only the FAR-score curve is used as normalization function parameters because there are far more samples to compute the FAR than those for the FRR. Computation of FAR introduces less error than that of FRR, which can be seen from Fig.1.

Given a training set of *K* classes, (*K*>>*1*), *m* samples for each class, we have *mK*(*K-1*) negative instances for computing the FAR, while the instances for FRR are *mK* which is usually much less than *mK*(*K-1*). After the FAR-score curve is obtained for each classifier, scores can be re-normalized by the FAR-score curve to reflect the global probability of a negative instance being accepted by the classifier, and each classifier will have its own normalization function learned especially through experiments. This function represents the classification capability of the classifier. A. Ross, et al in [5] normalized scores obtained from different classifiers by a single function with the assumption of that scores follow Gaussian distribution whilst our FAR-score normalization method compute the posteriori of *P*(*genuine*|*imposter*,*s*<*h*) through experiments without any assumption of original score distribution but only suppose that samples follow a certain distribution as proposed in [8].

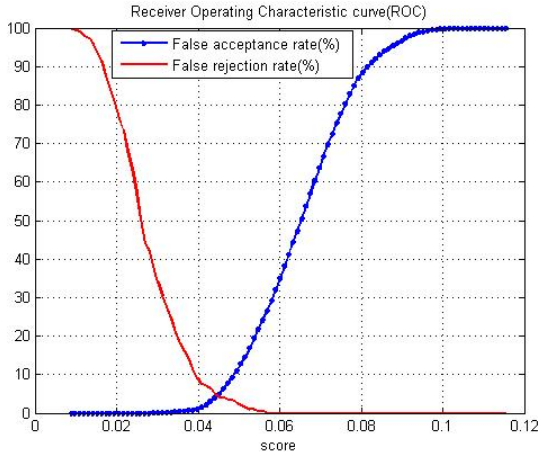


Figure 1: A typical ROC curve

Before using the FAR-score curve to normalize the scores from the classifier, the FARs of a series $s$ should be computed beforehand in order to learn the FAR-score curve. For the above mentioned training set, the $j$th classifier at the threshold of $h^j_i$, if $e$ impostor samples are accepted, then $far^j_i=e/(K(K-1)m)$, using a set of thresholds $(h^j_{i-1}<h^j_i<h^j_{i+1})$, the FAR-score curve can be calculated. When a testing sample $Z$ comes with a claim, the score $s^j$ from the $j$th matcher can be normalized by the curve. If FAR monotonically increase with $h^j_i$, $s^j$ is normalized by (7):

$$n^j = far_i^j(h_i^j) + \frac{dfar_i^j}{dh^j}\Big|_{h^j=h_i^j}(s_i^j - h_i^j) \qquad (7)$$

for $h_i^j \leqslant s_i^j \leqslant h_{i+1}^j$, otherwise, (8) is used to normalize $s$:

$$n^j = far_i^j(h_{i+1}^j) + \frac{dfar_i^j}{dh^j}\Big|_{h^j=h_i^j}(s_i^j - h_{i+1}^j) \qquad (8)$$

for $h_i^j \geqslant s_i^j \geqslant h_{i+1}^j$. When scores from all classifiers are normalized into FARs, the common fusion methods such as sum, min, med and max can be adopted to compute a single scalar to judge if accept or reject the claim.

## III. Classifier Discriminant Analysis

The traditional fusion rules, such as the sum, min, med, max, etc., only focused on the scores of the claimed class for the testing sample, while neglecting to analysis the discriminant of the classifiers, so they can not judge which classifier gives the minimum error. To achieve the minimum combination error, the discriminant analysis must be performed for every classifier before the fusion. When combining the scores from different classifiers, the sum rule in [9] adds the error together when summing the scores of the claimed class from all classifiers without analyzing the classifier discriminants, thus it can not take the minimum error; as for the min, max and med rules, they simply compare the scores from different classifiers of the claimed class while not comparing the discriminant of the classifiers.

To improve the fusion result, a classifier discriminant analysis (CDA) algorithm which can adaptively select the appreciate classifier for every testing sample is developed to combine the resulting scores normalized by the FAR-score curves. The CDA algorithm first evaluates the performance of every classifier for each testing sample based on the FAR-score curve normalized scores, then takes the scores from the appreciate classifier as the fusion results. When a testing sample comes with a claim of belonging to the $t$th class, besides the claimed class $t$, each class $i$ ($i\neq t$) in the trained pattern space can also give a score $s_i$ to manifest the similarity to the potential claims that the testing sample belongs to $i$th class. When evaluate the performance of the classifiers, the $k$ classes that show the most similarities to the testing sample are especially important. In this paper, the $s$ is normalized into FAR $n$ by the FAR-score curves, the less of $n$, the less risk to accept an imposter sample, the more similarity of the testing sample to the claimed class. The CDA first consider the $k$ least $n_i$ to evaluate the perform error of the classifiers. If all the mean $n_i$ from the $k$ nearest classes of different classifiers are less than a certain threshold, it is hard to judge which classifiers is the appreciate classifier, a kind of relative distance is used to evaluate the classifier. Otherwise, the classifier which take the minimum risk to accept an impost sample is looked as the appreciate classifier, thus the minimum rule is applied to take the minimum risk in this case. For a system of $N$ classifiers and $K$ clients, if a testing sample comes with a claim of belonging to the $t$th class, besides the potential claim that the testing sample belong the class $i$ ($i\neq t$) in the trained pattern space, the $j$th classifier gives $K$ scores $s_i^j$ ($i$=1,…$K$) to show the similarities of the testing sample and the $i$th class, normalize the scores with the Far-threshold curve into $n_i^j$, then the fusion result $fs_t$ is computed by the CDA algorithm is as following:

(i) sort the normalized score with increase sequence, the sorted normalized score $sn_i^j$, $sn_i^j \leqslant sn_{i+1}^j$, for $j$=1,2…$N$;

(ii) calculate the mean of the scores for the $k$ nearest classes, $\mu^j=\sum_{i=1}^k sn_i^j/k$, for $j$=1,2…$N$;

(iii) if $\mu^j \leq th$, for $j$=1,2…$N$, ($th$ is a given threshold) goto iv, otherwise goto v ;

(iv) calculate the relative distance $d^j=1/sn_1^j\sum_{i=1}^k sn_i^j$ for $j=1,2,...N$, if $dt=\min_j d^j$ the $j$th classifier is considered the appreciate classifier and the fusion result is $fs_t=n_t^k$;

(v) in this case, the classifier that take the minimum risk to accept an imposter sample is the appreciate classifier, thus $fs_t=\min_j n_t^j$.

## IV. Experiments

In the experiments, LDA, ICA and 2d-LDA are three individual classifiers and their scores have different distributions. These scores are first normalized and then combined into an aggregated score in order to decide if to accept or reject the claim. Our experiments are conducted on two public available face databases: ORL and XM2VTS. The ORL database [10] consists of 40 individuals and 10 images for each individual, the first 5 images are used as training samples and the remains as the testing samples. In the training process, each sample is used as the negative instance for the else classes, so there are 7800 negative instances to compute the FARs. In the testing stage, a testing sample is assigned to the positive set and negative set, and then all the testing samples yield 200 positive instances and 7800 negative instances.

For XM2VTS database under the configuration I [11], training system has 200 clients with 3 samples for each client, again we use each training sample as the negative instances for the else classes, thus there are 119600 instances to calculate the FARs. In the testing stage, every clients has 2 samples to form the positive set, thus the positive set has 400 instances, 70 impostors, each impostor with 8 samples form the negative set is used to attack each client class, which sum up 112000 negative instance.

Fig.2 show the ROCs on ORL database and table 1 gives the error rates for several key points on ROCs. Fig. 2a is the

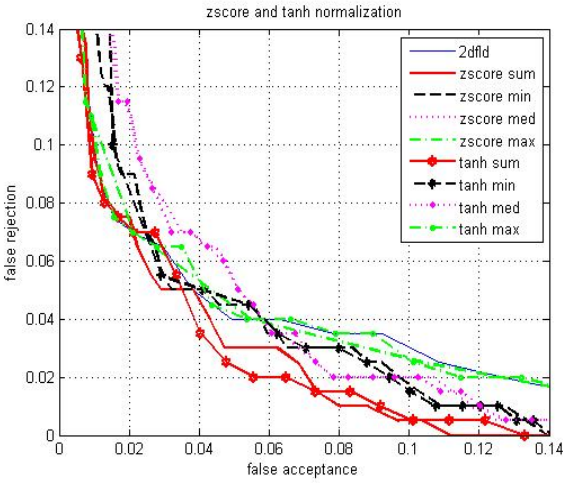


Figure 2a: Result of normalization with *z*-score and tanh functions and fusion with min, med, max and sum rules on ORL database.

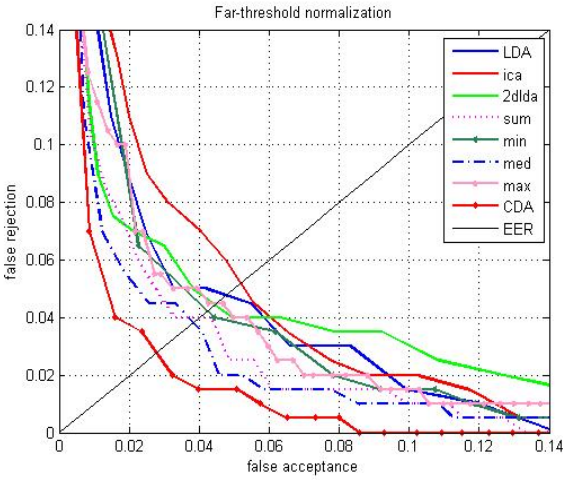


Figure 2b: Result of normalization with FAR-score curve, fusion with simple rules and CDA algorithm on ORL database.

results of the *z*-score and tanh normalization with simple fusion methods. In Fig.2b, scores are normalized by FAR-score curves. We can see from Fig.2a the fusion results are not always better than the original methods, but in Fig.2b, when the scores are converted into false accepting rates, the results are improved significantly, especially when the CDA is applied, even better results are achieved. From Table 1, wecan see the equal error rate (EER) improves to 3.8% from 2d-LDA's 4.5% in the FAR-score normalized fusion methods, 0.3% better than the best result in the tanh normalized algorithm, the CDA reduces the EER to 2.7% further. For other key points at ROC on the FAR-score normalized methods, improvements are remarkable as well, e.g. when FA=1%, FRR is 5.5% for the CDA, when FR=1%, the CDA improves the FA from 12% of the best original methods to 4.7%. Fig.3a and Fig.3b are ROCs for the experiments on the XM2VTS database at configuration I. They show the same trends of performance improvement as the experiments on the ORL face database. The CDA improves EER to 4.5% after the scores are normalized by FARs-score curve, which is 0.3% better than the best of tanh-sum algorithm from Table 2.

Table 1: Key error rate data (%) in Fig.2a and Fig.2b

| | | EER | FR (FA =1) | FR (FA=0.1) | FA (FR=1) | FA (FR=0.1) |
|---|---|---|---|---|---|---|
| FLD | | 4.7 | 14.5 | 20.5 | 12 | 13 |
| ICA | | 5 | 15 | 26 | 12 | 13.25 |
| 2DFLD | | 4.5 | 9 | 30.5 | 16.91 | 21.64 |
| *z*-score | min | 4.7 | 14.5 | 31 | 10.9 | 13.35 |
| | med | 5 | 15.5 | 31 | 11.79 | 12.51 |
| | max | 5 | 9.7 | 31 | 16.94 | 20 |
| | sum | 4.1 | 9.1 | 31 | 16.94 | 20 |
| Tanh | min | 4.7 | 14.5 | 20.5 | 10.9 | 13.35 |
| | med | 5 | 15.5 | 22.5 | 11.79 | 12.53 |
| | max | 5 | 9.7 | 31 | 16.79 | 20 |
| | sum | 4.2 | 9.5 | 31 | 8.06 | 9.77 |
| FAR-score curve normalization | min | 4.3 | 14.5 | 20.5 | 11.86 | 13.1 |
| | med | 3.8 | 8 | 20.5 | 10.79 | 11.29 |
| | max | 4.5 | 11.5 | 23.5 | 10.56 | 14.44 |
| | sum | 4 | 8.8 | 20.5 | 9.97 | 11.58 |
| | CDA | 2.7 | 5.5 | 22.5 | 4.7 | 5.14 |

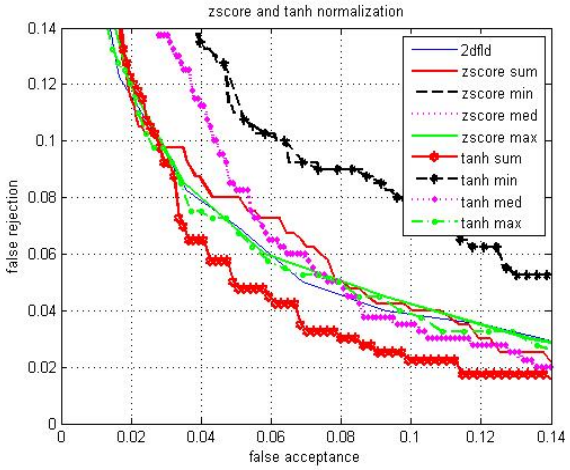


Figure 3a: Result of normalization with *z*-score and tanh and fusion with min, mediean, max and sum rules on XM2VTS database.

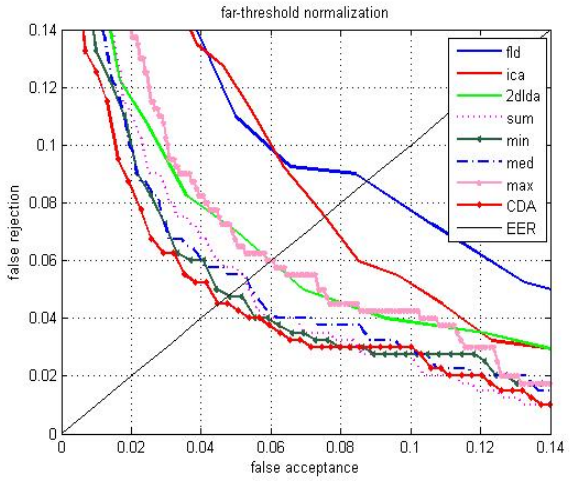


Figure 3b: Result of normalization with FAR-score curve, fusion with simple rules and CDA algorithm on XM2VTS database.

Table 2: Key error rate data (%) in Fig.3a and Fig.3b

| | | EER | FR (FA =1) | FR (FA=0.1) | FA (FR=1) | FA (FR=0.1) |
|---|---|---|---|---|---|---|
| FLD | | 9 | 21.75 | 36.5 | 34.06 | 38.07 |
| ICA | | 7.3 | 26.25 | 39.75 | 26.88 | 32.72 |
| 2DFLD | | 5.8 | 20.5 | 38.25 | 30 | 37.36 |
| *z*-score | min | 9 | 21.75 | 36.5 | 32.86 | 37.44 |
| | med | 6.28 | 22 | 41.25 | 22.79 | 36.64 |
| | max | 6 | 17.75 | 30.25 | 27.85 | 38 |
| | sum | 6.6 | 22 | 41.5 | 18.2 | 29.35 |
| tanh | min | 9 | 21.75 | 36.5 | 34.06 | 38.07 |
| | med | 6.28 | 22 | 41.25 | 22.79 | 36.65 |
| | max | 6 | 17.75 | 30.25 | 27.55 | 36.64 |
| | sum | 4.8 | 18.5 | 32 | 18.32 | 29.44 |
| FAR-score curve normalization | min | 4.7 | 13 | 26.5 | 15.81 | 19.79 |
| | med | 5.27 | 15.25 | 26.5 | 22.57 | 30.21 |
| | max | 5.8 | 20.5 | 38.25 | 22.57 | 30.21 |
| | sum | 5.23 | 16.25 | 35.25 | 12.78 | 30.21 |
| | CDA | 4.5 | 11 | 21.5 | 12.66 | 29.41 |

From above experimental results we can see, when the scores from different classifiers are converted into false accepting rates (FAR), the fusion results improve prominently. This manifests that the FAR-score normalization method is more reliable than both the *z*-score and the *tanh* normalization. The classifier discriminant analysis (CDA) improves the fusion results further, which shows the CDA is more adapt to fusion the FAR-score normalized scores.

## V. CONCLUSION

In this paper, a novel matching score normalization method for multi-classifiers based on their false acceptance rate (FAR) scores is proposed. When the scores are normalized with FAR-score curve, the normalized scores show the probabilities of accepting an impostor. The FAR-score curve of each classifier is computed without assumptions of observing any distributions, so scores from all classifiers can be normalized by its own FAR-score curve. Therefore, the method can be adapted to scores from any classifiers. To achieve better the fusion result, we give a method of classifier discriminant analysis to evaluate the performance of each classifier in order to choose the scores from the appreciate classifier for every testing samples. Experimental results of face verification on both ORL and XM2VTS databases show our approach efficiency and effectiveness compared to conventional methods.

The false rejection information of the matching scores has been discarded in our current work because there are usually not enough samples to compute FRR-score curve precisely. When the scores are normalized by the FRR-score curve, more normalization errors will be introduced that may lead to even worse fusion results. However, matching scores normalized by FRR-score are another part of important information of the matching scores that might be making use of by second-order combination. These will be part of our further work to do. We will continue our experimental testing for more face databases, i.e. the XM2VTS database at configuration II, FERET database, etc. and explore the sample influences of system-discriminated impostor (i.e. un-registered with the system) and client-discriminated imposter (i.e. other clients' sample).

## ACKNOWLEDGMENT

This research work is supported by the "Hundred Talent Program" of the Chinese Academy of Sciences (Grant No. sinap-26050432). Thanks also give to the AT&T Lab. at Cambridge University and CVSSP at Surrey University for their permission to use their face databases.

## REFERENCES

[1] M. A.Turk, A.Pentland,:"Eigenfaces for recognition'" *Journal of Cognitive Neuro-Science* 3(1) (1991) 71-86.

[2] P. N.Belhumeur, J. P.Hespanha, and D. J. Kriegman: "Eigenface vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans on PAMI* 19(7)(1997)711-720.

[3] M.S.Bartlett, J.R.Movellan and T.J.Sejnowski: "Face recognition by independent component analysis", *IEEE Trans. On Neural Networks*, vol. 13, no.6, pp. 1450-1464, 2002.

[4] M. Li, B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix", *Pattern Recognition Letters*, vol. 26, no.5, pp. 527-532, 2005.

[5] A. Ross and A. Jain: "Information fusion in biometrics", *Pattern Recognition Letters*, 24(2003) 2115-2125.

[6] A. Jain, K. Nandakumar and A. Ross: "Score normalization in multimodal biometric system", *Pattern Recognition* 38(2005) 2270-2285.

[7] J. Kittler, M. Hatef, R. P.W. Duin and J. Matas: "On combining classifiers", *IEEE Trans. on PAMI* vol. 20, No.3 pp.226-239, 1997.

[8] R.O. Duda, P.E. Hart and D.G. Stork: *Pattern Classification*, Wiley, New York, 2001.

[9] K. Tumer and J. Ghosh: "Analysis of decision boundaries in lLinearly combined neural classifiers," *Pattern Recognition*, vol.29, No.2, pp.341-348, 1996.

[10] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[11] K. Messer, J. Matas, J. Kittler, J.Luettin, G. Maitre, "XM2VTSDB: the extended M2VTS database", *Proceeding of Audio and Video-based Biometric Person Authentication*, Washington DC, USA, pp.72-77, 1999.

[12] Y. Wang, T. Tan and A.K. Jain: "Combing face and iris biometrics for identity Verification", Proceedings of Fourth International Conference on AVBPA, Guildford, UK, pp. 805-813, 2003.

[13] P. Verlinde, G. Cholet, "Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application", Proceedings of Second International Conference on AVBPA, Washingtom, DC, USA, pp. 188-193, 1999.