Using Score Normalization to Solve the Score Variation Problem in Face Authentication

Fei Yang[1,3], Shiguang Shan[1], Bingpeng Ma[1,3], Xilin Chen[1,2], and Wen Gao[1,2,3]

[1] Institute of Computing Technology, CAS, Beijing, 100080, China
{fyang,sgshan,bpma,xlchen,wgao}@jdl.ac.cn
[2] Computer College, Harbin Institute of Technology, Harbin, 150001, China
[3] Graduate School of Chinese Academy of Sciences, Beijing, 100039, China

Abstract. This paper investigates the score normalization technique for enhancing the performance of face authentication. We firstly discuss the thresholding approach for face authentication and put forward the "score variation" problem. Then, two possible solutions, Subject Specific Threshold (SST) and Score Normalization (SN), are discussed. But SST is obviously impractical to many face authentication applications in which only a single example face image is available for each subject. Fortunately, we have theoretically shown that, in such cases, score normalization technique may approximately approach the SST by using a uniform threshold. Experiments on both the FERET and CAS-PEAL face database have shown the effectiveness of SN for different face authentication methods including Correlation, Eigenface, and Fisherface.

1 Introduction

The development of person identity verification based on biometric information provides wide potential applications in security, law enforcement, and commerce. A number of biometric traits have been studied for identity verification in the recent years, such as face, voice, signature, fingerprint, iris, retina, palm print, gait, etc [1]. Among them, face recognition has some unique advantages. For instance, facial images are easy to capture without any invasion by using various digital cameras or scanners, while some other traits like fingerprint and iris require special equipments, some of which even may have potential invasion. In addition, the verification results using facial images can be checked easily by common people while some other traits can be distinguished only by few experts (e.g. fingerprint or iris). Furthermore, face verification systems can be deployed in the scenarios either the person cooperates or not (even not known), which provides special advantage for security surveillance.

A face verification system usually consists of four main modules: data acquisition, feature extraction, feature matching, and decision making [2]. The data acquisition module acquires the biometric data from a user; the feature extraction module processes the acquired biometric data and extracts a feature set to represent it; the matching module compares the extracted feature set with the stored templates using some matching algorithm in order to calculate matching scores; finally, the decision module compares the matching scores with a threshold. If the score is equal or larger than the threshold, the claim is conformed, or else is denied.

For face verification, many factors cause the scores to fluctuate. Variations in expression, lighting, aging will make the similarity scores of genuine faces decrease,

S.Z. Li et al. (Eds.): IWBRS 2005, LNCS 3781, pp. 31–38, 2005.

even smaller than those of imposter faces under normal conditions. What is more, different persons enrolled in the system have different characteristics. Therefore, the distribution of genuine and imposter scores may vary from person to person. This problem, which we call "score variation" problem, will make the performance degrade obviously. So, using Uniform Threshold (UT) for all subjects in the decision-making module is evidently unsuitable. Two possible solutions, Subject Specific Threshold (SST) and Score Normalization (SN) [2], can be used to solve this problem.

The SST method assigns a specific threshold for each subject in the database, in order to adjust to different genuine and imposter score distribution of different subjects. Multiple example images are needed to model the genuine and imposter distribution for each person. But for many face authentication applications, only a single face image is available for each subject, so the genuine distribution is unavailable and SST can not be applied directly. In these cases, the SN methods are more commonly exploited. SN methods are mostly used in classifier fusion [2], signature verification [3], and speaker verification [4] domains. In face authentication filed, Sanderson etc. [5] use a Gaussian Mixture Model (GMM) based classifier for classification, in which the imposter distribution is modeled for normalization. Perronnin etc. [6] use relational approaches and develop two new score normalization methods R-Norm and G-Norm. Both their work need multiple images to train Gaussian model for each subject. The FRVT 2002 test [7] uses score normalization at similarity score level as a post-processing operation.

In this paper, we deal with only the similarity score, but not caring the types of the classifier. First, we put forward the constraints of the optimal Subject Specific Threshold (SST). Then, we show theoretically that Z-Norm, one of the SN methods, can approximately approach the SST method when genuine distribution is unavailable. We perform experiments to show how Z-Norm changes the distribution of scores, and its approximate performance to SST. Our experiments on FERET [8] and CAS-PEAL [9] face-database verify the effectiveness of SN for different face authentication methods including Correlation [10], Eigenface [11], and Fisherface [12].

2 Solutions to "Score Variation" Problem

In face verification, the system compares the candidate image with the image of the claimed identity to get a similarity score s. The decision module accepts or rejects the identity by comparing with a threshold θ, which is referred as the Uniform Threshold (UT) method.

$$\text{if } s \geq \theta \text{ accept, otherwise reject} \tag{1}$$

There are two probability distributions of pair-wise matching scores as illustrated in Fig. 1. The *genuine distribution* characterizes the similarity scores of two images from the same person, while the *imposter distribution* characterizes the similarity scores of two images from different persons.

Ideally, the genuine distribution will show small differences and the imposter distribution will show large differences. But for a real-world face verification system, variations in pose, lighting, expression, and aging may decrease the genuine scores significantly. In addition, since different persons enrolled in the system have different

characteristics, the distributions of genuine and imposter scores may also vary for different persons. This problem, which we call as "score variation" problem, will degrade the performance of UT-based systems obviously. Two possible solutions, Subject Specific Threshold (SST) and Score Normalization (SN), are previously proposed to solve this problem.

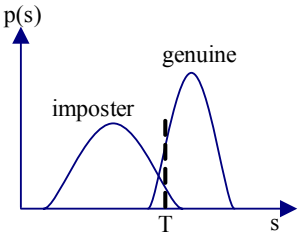

Fig. 1. Genuine and imposter distribution in a UT-based system

2.1 Subject Specific Threshold (SST)

One solution to the "score variation" problem is through the Subject Specific Threshold (SST) approach. In this approach, each subject G_i in the database has a specific threshold θ_i. For subject G_i, the decision making rule is redefined as:

$$\text{if } s \geq \theta_i \text{ accept, otherwise reject} \tag{2}$$

Then the False Alarm Rate (FAR) and Hit Rate (HR) for subject G_i are defined as:

$$FAR_i = \int_{\theta_i}^{+\infty} p(s \mid \overline{\lambda}_i)\, ds \qquad HR_i = \int_{\theta_i}^{+\infty} p(s \mid \lambda_i)\, ds \tag{3}$$

where $p(s \mid \overline{\lambda}_i)$ and $p(s \mid \lambda_i)$ are respectively the imposter distribution and the genuine distribution for the G_i subject (refer to Fig. 1). So, the total FAR and HR for all registered subjects are defined as:

$$FAR = \frac{1}{|G|}\sum_{i=1}^{|G|} FAR_i = \frac{1}{|G|}\sum_{i=1}^{|G|}\int_{\theta_i}^{+\infty} p(s \mid \overline{\lambda}_i)\, ds \tag{4}$$

$$HR = \frac{1}{|G|}\sum_{i=1}^{|G|} HR_i = \frac{1}{|G|}\sum_{i=1}^{|G|}\int_{\theta_i}^{+\infty} p(s \mid \lambda_i)\, ds \tag{5}$$

For a face authentication system, if the FAR is fixed to t (e.g. according to the practical requirement), one should set the thresholds of each subject, $\theta_1, \theta_2, \cdots, \theta_{|G|}$, in order to get a highest HR. These thresholds can be solved by using the Lagrange Multipliers. First, we define the Lagrange function:

$$\begin{aligned} F &= HR - K \cdot (FAR - t) \\ &= \frac{1}{|G|}\sum_{i=1}^{|G|}\left[\int_{\theta_i}^{+\infty} p(s \mid \lambda_i)\, ds - K \cdot \left(\int_{\theta_i}^{+\infty} p(s \mid \overline{\lambda}_i)\, ds - t\right)\right] \end{aligned} \tag{6}$$

To obtain a maximum of HR, the partial derivatives $\partial F / \partial \tilde{\theta}_i$ should be zero, that is,

$$\frac{\partial F}{\partial \theta_i} = -\frac{1}{|G|}(p(\theta_i \mid \lambda_i) - K \cdot p(\theta_i \mid \overline{\lambda}_i)) = 0 \tag{7}$$

So, finally we have:

$$\frac{p(\tilde{\theta}_i \mid \lambda_i)}{p(\tilde{\theta}_i \mid \bar{\lambda}_i)} = K, \quad (i = 1,2,\cdots \mid G \mid) \tag{8}$$

.

If we further assume the genuine scores of subject G_i satisfy a Gaussian distribution $N(\mu_i, \sigma_i)$ and imposter scores be a Gaussian distribution $N(\bar{\mu}_i, \bar{\sigma}_i)$, given a certain value K, the subject specific threshold can be solved by

$$\tilde{\theta}_i = f(K, \mu_i, \sigma_i, \bar{\mu}_i, \bar{\sigma}_i). \tag{9}$$

2.2 Score Normalization (SN)

Previous studies have shown that the performance of a number of biometric verification systems, especially those based on behavioral traits such as written signature or voice, can be improved by using score normalization method. In score normalization, the normalized score is a function of original similarity score, the input sample, the client specific information, and the information of imposters [6]. Researchers have proposed several SN methods, such as Z-Norm [3], T-Norm [6], G-Norm [8], etc. Among them, the Z-Norm method normalizes the similarity score by using the mean and standard deviation of the imposters:

$$s_{znorm} = \frac{s - u_{\bar{\lambda}_i}}{\sigma_{\bar{\lambda}_i}} \tag{10}$$

where $u_{\bar{\lambda}_i}$ is the mean of the scores from imposter images of subject G_i, $\sigma_{\bar{\lambda}_i}$ is the standard deviation.

Z-Norm actually aims to make the imposter distribution of different subjects the same standard normal distribution $N(0,1)$. Thus, one can easily configure the final verification system by setting a uniform threshold for all subjects.

3 Using Z-Norm to Approximately Approach SST

In the SST method, optimal target specific thresholds $\theta_1, \theta_2, \cdots, \theta_{|G|}$ are estimated by the candidate facial image, the genuine and imposter score distribution of identity G_i. However, in many face verification systems, only a single image per subject is enrolled in the system. Thus, the genuine distribution can not be obtained. In this case, we can neglect the genuine distribution in Eq.8, using only the imposter term:

$$p(\tilde{\theta}_i \mid \bar{\lambda}_i) = 1/K, \quad (i = 1,2,\cdots \mid G \mid) \tag{11}$$

Suppose the imposter scores satisfy a Gaussian distribution $N(\bar{\mu}_i, \bar{\sigma}_i)$, then

$$\tilde{\theta}_i = \mu_{\bar{\lambda}_i} + \sigma_{\bar{\lambda}_i} \cdot \sqrt{2\ln(\sigma_{\bar{\lambda}_i} \cdot K) + \ln(2\pi)}, \quad (i = 1,2,\cdots \mid G \mid) \tag{12}$$

We will see in the experimental part that $\sigma_{\bar{\lambda}_i}$ varies little for all subjects, so

$$\tilde{\theta}_i = \mu_{\bar{\lambda}_i} + \sigma_{\bar{\lambda}_i} \cdot K', \quad (i = 1,2,\cdots \mid G \mid) \tag{13}$$

This form is equivalent to Z-Norm since

$$s \geq \tilde{\theta}_i \Leftrightarrow s \geq \mu_{\bar{\lambda}_i} + \sigma_{\bar{\lambda}_i} \cdot K' \Leftrightarrow \frac{s - \mu_{\bar{\lambda}_i}}{\sigma_{\bar{\lambda}_i}} \geq K' \Leftrightarrow s_{znorm} \geq K' \tag{14}$$

This means that the Z-Norm method approximately approaches the SST method when only imposter distribution is available in the system.

4 Experiments and Analysis

Experiments are performed on two public face databases, FERET and CAS-PEAL. For FERET database, the training set contains 1002 frontal images of 429 persons. Four testing sets Fafb, Fafc, DupI, DupII are used to test the performance under variations of expression, lighting and aging. For CAS-PEAL database, the training set contains 300 persons with 4 images per person. Six testing sets Accessory, Aging, Background, Distance, Expression and Lighting are used to test performance under corresponding variations. For each facial image used for training and testing, the preprocessing procedure consists of locating the centers of two eyes, geometrical transformation to place the center of two eyes on specific position. Each image is cropped to the size 64*64, processed by histogram equalization, and concatenated by rows to form a vector of 4096 dimension.

First, we conduct an experiment to show the effectiveness of Z-Norm on CAS-PEAL training set. The CAS-PEAL training set contains 300 subjects, 4 images per subject. Thus, for every subject, 6 genuine scores and 4784 imposter scores can be computed by using correlation method. The mean and variance of genuine and imposter scores before and after Z-Norm for each subject are shown in Fig. 2, whose horizontal axis denotes the subject number. The two figures in the top row show the mean and variance before Z-Norm and the bottom row shows them after Z-Norm. One can find that before Z-Norm, the distribution of scores fluctuates a lot for each subject, while the variances of imposter scores for each subject are almost the same. After Z-Norm, the mean and variance of imposter scores are 0 and 1 for each subject, which will greatly facilitate the setting of a uniform threshold.

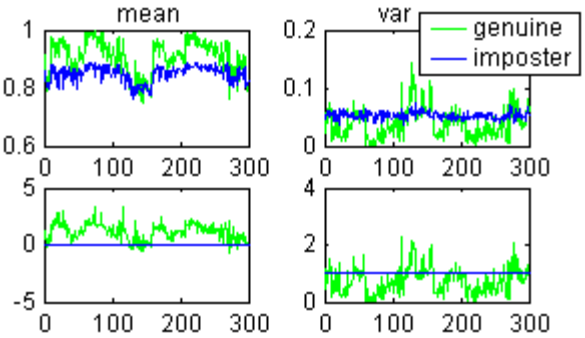

Fig. 2. Genuine and imposter scores distribution of each subject before and after Z-Norm

Secondly, we conduct experiments on FERET face database to verify the effect of Z-Norm, and SST (using Eq.12). Figure 3 shows the ROC curve of normalized corre-

lation method testing on the FERET-FB probe set. As we can see clearly, the ROC of Z-Norm and SST are approximate, and both outperform UT impressively.

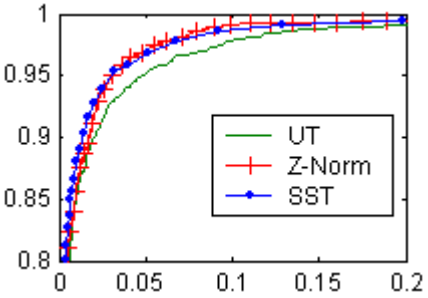

Fig. 3. ROC curves of correlation method with UT, Z-Norm and SST on fafb probe set

Finally, verification testing experiments are conducted on the FERET and CAS-PEAL face database to compare the EER of Z-Norm and UT. Since SST performs similarly to Z-Norm, it is not tested in these experiments. We test three different verification methods: Correlation, Eigenface, and Fisherface. Figure 4 and 5 show comparison results on the four probe sets of FERET and the six probe sets of CAS-PEAL respectively. For each classifier, the left column shows the EER by using UT, while the right column shows that of the Z-Norm method. From these figures, one can see that Z-Norm method can get better performance on FERET fafb, DupI, and DupII testing sets. On most of the testing sets of CAS-PEAL database, Z-Norm enhances the performance with only one exception (Fisherface on Aging test set).

The experimental comparisons on the FERET and CAS-PEAL database show that Z-Norm can generally enhance the performance for Correlation and Eigenface method. For Fisherface method, Z-Norm does not improve the verification performance obviously.

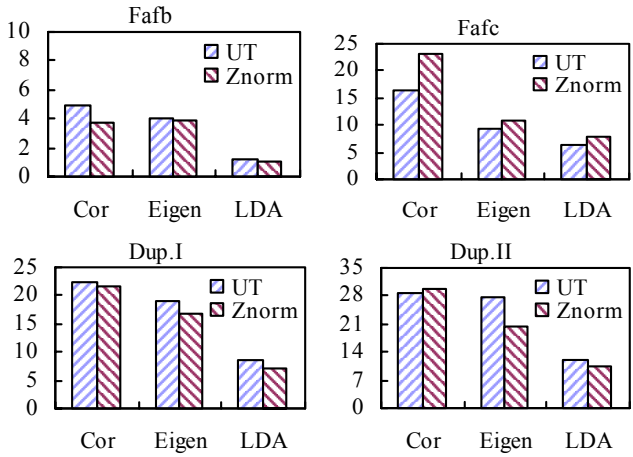

Fig. 4. EER on FERET database (Cor: Correlation, Eigen: Eigenface, Fisher: Fisherface)

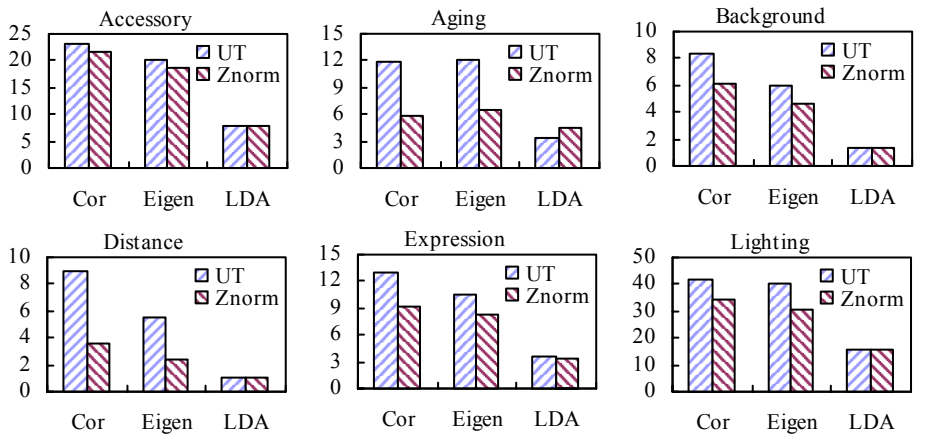

Fig. 5. EER on CAS-PEAL database (Cor: Correlation, Eigen: Eigenface, Fisher: Fisherface)

5 Conclusions

This paper discusses two methods, Subject Specific Threshold (SST) and Score Normalization (SN), aiming at the "score variation" problem caused by the significant variations in testing images due to varying lighting, pose, and expressions. We have theoretically and experimentally shown that a uniform threshold after score normalization using Z-Norm can approximate the SST method. In addition, our experiments on FERET and CAS-PEAL face database with three different face verification methods also illustrate the effectiveness of the Z-Norm method compared with the uniform threshold (UT) without score normalization.

Acknowledgements

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, and No.60473043, "100 Talents Program" of CAS, ShangHai Municipal Sciences and Technology Committee (No.03DZ15013), and ISVISION Technologies Co., Ltd.

References

1. A.K. Jain, A. Ross and S. Prabhakar: An Introduction to Biometric Recognition, IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Image and Video Based Biometrics, Vol. 14, No. 1, pp. 4-20, Jan. 2004.
2. A.K. Jain, K. Nandakumar, A. A. Ross: Score Normalization in Multimodal Biometric Systems, to appear in Pattern Recognition, 2005.
3. J.F. Aguilar, J. O. Garcia, J. G. Rodriguez: Target Dependent Score Normalization Techniques and Their Application to Signature Verification, Proc. International Conference on Bioinformatics and its Applications (ICBA), LNCS 3072, pp. 498-504, Dec. 2004.
4. C. Barras, J. L. Gauvain: Feature and Score Normalization for Speaker Verification of Cellular Data, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. II, pp. 49-52, Hong Kong, Apr. 2003.

5. C. Sanderson, K. K. Paliwal: Likelihood Normalization for Face Authentication in Variable Recording Conditions, Proc. IEEE International Conference on Image Processing (ICIP), Vol. I, pp. 301-304, Rochester, Sep. 2002.
6. F. Perronnin, J. L. Dugelay: Robust Score Normalization for Relational Approaches to Face Authentication, 12th European Signal Processing Conference, Sep. 2004.
7. P.J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, J. M. Bone: Face Recognition Vendor Test 2002 Evaluation Report, NISTIR 6965, Mar. 2003.
8. P.J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi: The FERET Evaluation Methodology for Face Recognition Algorithms, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, Oct. 2000.
9. Bo Cao, Shiguang Shan, Xiaohua Zhang, Wen Gao: Baseline Evaluations on the CAS-PEAL-R1 Face Database, SinoBiometrics 2004, pp. 370-378.
10. R. Brunelli, T. Poggio: Face Recognition: Features vs. Templates, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15, No. 10, pp. 1042-1053, Oct. 1993.
11. M. Turk and A. Pentland: Eigenface for Recognition, Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.
12. P.N. Belhumeur, J. P. Hespanha, D. J. Kriegman: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, Issue 7, Jul. 1997.