Advanced Engineering Forum *Online: 2011-09-09*
ISSN: 2234-991X, Vol. 1, pp 168-172
doi:10.4028/www.scientific.net/AEF.1.168
© 2011 The Author(s). Published by Trans Tech Publications Ltd, Switzerland.

An Evaluation Survey of Score Normalization in Multibiometric Systems

Yong Li[1, a], Jianping Yin[1,b] and En Zhu[1,c]

[1] School of Computer, National University of Defense Technology, Changsha, 410073, China

{[a] liyong, [b] jpyin, [c] enzhu}@nudt.edu.cn

Keywords: Score normalization, multibiometrics, performance evaluation, fusion.

Abstract. Multibiometric fusion is an active research area for many years. Score normalization is to transform the scores from different matchers to a common domain. In this paper, we give a survey of classical score normalization techniques and recent advances of this research area. The performance of different normalization functions, such as MinMax, Tanh, Zscore, PL, LTL, RHE and FF are evaluated in XM2VTS Benchmark. We evaluated the performance with four different measures of biometric systems such as EER, AUC, GAR(FAR=0.001) and the threshold of EER. The experimental results show that there is no single normalization technique that would perform the best for all multibiometric recognition systems. PL and FF normalization outperform other methods in many applications.

1. Introduction

Biometric recognition refers to the use of distinctive physiological or behavioral characteristics for automatically confirming the identity of a person. Multibiometrics which combines more information is expected to improve the performance of biometric system efficiently. Depending on the level of information that is fused, the fusion scheme can be classified as sensor level, feature level, score level and decision level fusion [1]. Apart from the raw data and feature sets, the match scores contain the richest information about the input pattern. Also, it is relatively easy to get and combine the scores generated by biometric matchers. Consequently, score level fusion is the most commonly used approach in Multibiometric systems. Scores generated by different matchers are not homogeneous often. For example, scores of different matchers may not be on the same numerical range and may follow different probability distributions. Therefore score normalization which transforms these scores into a common domain before fusion is needed. This paper will give an overview and comparison of score normalization methods in multimodal fusion.

The remainder of this paper is organized as follows. Section 2 introduces the Fusion in multimodal biometrics include the ideal normalization function, the performance measure and the combination rules. In Section 3, several score normalization techniques are introduced include classical and the advances of normalization methods. To study the effectiveness of different normalization techniques, section 4 gives the experimental results. The last section summarizes the results of this work.

2. Fusion in multimodal biometrics

The Ideal Normalization Function. In this paper, matching score coming from samples of the same individual is noted as genuine score while that coming from samples of different individuals noted as imposter score. Since scores from different recognition systems are not comparable, the normalization step tries to find the function which can transform the scores into the common domain and make the the scores of different matchers comparable. The ideal normalization function is the posteriori probability functions which is given by

$$s^{ideal} = p(genuine \mid s) / (p(impostor \mid s) + p(genuine \mid s)) \tag{1}$$

$p(genuine \mid s)$ and $p(impostor \mid x)$ refer to conditional density of the matching score being that of a genuine user or impostor user. It is difficult to estimate the density of matching scores in that they may not obey a certain distribution model. Therefore the ideal normalization function is not easy to implement. And different normalization techniques have been proposed in literature to solve this problem. A good normalization method should be robust and insensitive [1]. Robustness refers to insensitivity to the presence of outliers and Efficiency refers to the proximity of the obtained estimate to the optimal estimate when the distribution of data is known.

Performance Measures. Let us denote with ***t*** an acceptance threshold so that users whose score is larger than ***t*** are assigned to the genuine class, while users whose score is smaller than ***t*** are assigned to the impostor class. The two errors, respectively the False Rejection Rate (FRR), and the False Acceptance Rate (FAR) are defined as follows.

$$FAR(t) = p(\mathrm{s} \ge t \mid impostor) = \int_t^{\infty} f_{imp}(s)ds \text{ and } FRR(t) = p(\mathrm{s} < t \mid genuine) = \int_{-\infty}^{t} f_{gen}(s)ds \quad (2)$$

The Genuine Accept Rate (GAR) is the fraction of genuine scores exceeding the threshold t. Therefore GAR=1-FRR. The most widely accepted method used to evaluate the performance of a biometric system is the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the GAR (or FRR) against the FAR. The Equal Error Rate (EER) is the point of the ROC curve where the two errors, i.e. the FAR and the FRR, are equal. EER is widely used in the biometric field to assess the performance of biometric systems. GAR(FAR=0.001 or else) is another performance measure which is also widely used in biometric performance evaluation[1]. In ROC analysis the Area Under the Curve (AUC) [2] is the also used evaluate the performance of a two-class system because it is a more discriminating measure than the accuracy. In biometric recognition systems, we always try to make EER smaller and GAR(FAR=0.001) as well as AUC larger.

Combination Rules. After normalizing the matching scores and then we need to acquire a new score through a certain combination fusion rule to make final decision. Kittler [3] et al. proposed a general fusion theory framework and deduced five basic fusion rules: *Sum, Product, Max, Min* and *Median*. Since *Sum* rule works better in most applications [4], we use *Sum* rule to get the final mark in our experiments to evaluate the performance of the normalization techniques.

3. Score normalization schemes

Several classical score normalization techniques such as MinMax, Tanh, Z-score, Median, Median/MAD and Decimal Scaling have been described in Ref. [1]. Among the classical normalization techniques, Median/MAD and decimal scaling are not robust and Efficiency, therefore, we choose MinMax, Tanh and Z-score in the experiments in Section 4. Then we describe the progress of normalization techniques in recent years. In this section, let *X*, X_G and X_I denote the set of raw matching scores, genuine scores and imposter scores of training data. And let ***s*** denotes the new score which associated with the same matcher. The normalized score of ***s*** is then denoted by s' . *Max*, *Min*, *Median*, *μ and σ* are the maximum, minimum, median, mean and standard deviation values**.**

Piecewise linear (PL) [5] normalization technique transforms the scores in the range of [0, 1]. The normalization function of PL maps the raw scores using piecewise linear function as,

$$s' = \begin{cases} 0 & s < \min(X_G) \\ 1 & s > \max(X_I) \\ (s - \min(X_G)) / (\max(X_I) - \min(X_G)) & else \end{cases} \quad (3)$$

Four Segment Piecewise-Linear(FSPL)[6] technique divides the regions of impostor and genuine scores into four segments and map each segment using piecewise linear functions. The scores between two extremities of the overlap region are mapped using two linear functions separately in range of [0, 1] and of [1, 2] towards left and right of ***t***, respectively as equation (4).

$$s' = \begin{cases} 0 & s < \min(X_G) \\ (s - \min(X_G)) / (t - \min(X_G)) & s < t \\ 1 + (s - t) / (\max(X_I) - t) & s > t \\ 2 & s > \max(X_I) \end{cases} \quad \text{where } (\max(X_I)<t<\min(X_G)) \quad (4)$$

Linear Tanh Linear(LTL) [6] normalization technique takes the advantage of the ***tanh*** estimator and the PL normalization. Normalization function of LTL maps the non overlap region of impostor scores to a constant value 0 and non overlap region of genuine scores to a constant value 1. The overlapped region between **max(*X_I*)** and **min(*X_G*)** is mapped to a nonlinear function using ***tanh*** estimator as,

$$s' = \begin{cases} 0 & s < \min(X_G) \\ 1 & s < \max(X_I) \\ 0.5*[\tanh\{0.01*((s-u(X_G))/\sigma(X_G))\}+1.5] & else \end{cases} \quad (5)$$

Reduction of High-scores Effect normalization(RHE) [7] is derived from min–max normalization scheme. The idea behind RHE is based on following observations: Any kind of normalization always causes loss of information content. Multimodal biometric systems suffer mainly from the 'low' genuine scores instead of 'high' impostor scores. So the RHE normalization method is given by

$$s' = (s - \min(X)) / (u(X_G) + \sigma(X_G) - \min(X)) \quad (6)$$

FRR and FAR based normalization (FF)[8] can be regard as a new normalization method: FRR and FAR-based normalization (FF). The FF normalization use FRR/(FAR+FRR) as a means to normalize matching scores, Therefore, we firstly calculate FRR and FAR of each matching score based on training samples, then compute FRR and FAR of each matching score through interpolation based on testing samples.

$$s' = FRR(s) / (FRR(s) + FAR(s)) \quad (7)$$

Because the FSPL and LTL are both improvement of the PL normalization and LTL show better performance than the FSPL methods. The PL, LTL, RHE and FF normalization methods were chosen during the experiments.

4. Experimental Results

Database. The XM2VTS-Benchmark[9] database consists of five face matchers and three speech matchers and was partitioned into training and evaluation sets according to the Lausanne Protocol-1(LPI). The benchmark of LPI includes two files, one is dev.label and the other is eva.label. We use dev.label as training data and eva.label as test data. Our experiments are conducted based on this match score benchmark. We sign the face matcher as face-1, face-2, face-3, face-4 and face-5 and the speech matcher as speech-1, speech-2 and speech-3 respectively.

Experimental Results. We conducted experiments to measure the benefits between the 7 normalization methods: MinMax(MM), Tanh, Zscore, PL, LTL, RHE and FF. The EER of all the matchers can be found in Table 1. As shown in Table 1, among face matchers, matcher face-3 and face-5 gain the best and worst performance respectively. And among speech matchers, the performance order is speech-1, speech-3 and speech-2. The experiments are conducted with 15 kind multimodal combinations. In each combination, the scores of different matchers are normalized first and *Sum* rule is used to get the final score. Then different thresholds are set to compute the FRRs and FARs. Table 2 shows the EER of multi-modal fusion among the 7 normalization methods. In order to evaluate the performance precisely, for each fusion, we give each matcher the performance mark. The performance mark for the best matcher is 7 and followed by 6, 5, 4, 3, 2 and 1. If the performance of two matchers are the same, for example, both are the second best, then the two matchers get the same mark (6+5)/2=5.5. Table 3 is the performance mark of different fusion techniques which is measured by EER. From Table 3, we can easily find that the proposed fusion method FF shows the best performance because the total mark is the largest one. And we observe that the PL and Zscore methods also perform well. To show the comparison of all the algorithms in multimodal biometric systems, Fig 1 shows the EERs of the 7 normalization algorithms. From the last column of Table 3, the sum of performance mark summaries the performance from EER aspects. It is easy to find that FF, PL and Zscore methods give better performance than other normalization methods. Fig 2 and Fig 3 shows the GARs(FAR=0.001) and AUCs of the 7 normalization methods.

Table 1: EERs of each simple recognition system on XM2VTS-Benchmark

matcher	S1	S2	S3	F1	F2	F3	F4	F5
EER(%)	1.109	6.500	4.500	1.814	4.115	1.767	3.500	6.500

Table 2: EERs of different normalization methods based on XM2VTS-Benchmark

EER(%)	F1S1	F1S2	F1S3	F2S1	F2S2	F2S3	F3S1	F3S2	F3S3	F4S1	F4S2	F4S3	F5S1	F5S2	F5S3
MM	0.9	1.2	1.0	0.5	1.6	1.2	0.4	1.2	0.7	1.1	1.3	1.5	2.6	3.5	3.7
Tanh	0.5	1.7	0.6	0.5	1.7	1.2	0.5	1.6	1.2	0.7	1.4	1.2	1.0	3	2.5
Zs	0.7	1.2	0.9	0.5	1.5	1.2	0.4	1.2	0.7	0.7	0.9	1.2	1.1	2.8	2.2
PL	0.5	0.9	0.7	0.6	1.7	1.2	0.5	1.2	0.7	0.6	0.9	1.2	0.6	2.7	2.1
LTL	0.7	2.7	1.2	0.6	2.7	1.2	0.6	2.7	1	0.5	2.2	0.8	0.9	3.3	2
RHE	0.7	1.1	0.8	0.5	1.7	1.2	0.5	1.2	0.7	0.7	0.9	1.4	1.5	3	3
FF	0.2	1.2	0.7	0.5	1.7	1.0	0.3	1.5	1.1	0.3	1.0	0.7	1	2.4	1.7

Table 3: Sum of performance mark of different normalization methods based on XM2VTS-Benchmark

Mark	F1S1	F1S2	F1S3	F2S1	F2S2	F2S3	F3S1	F3S2	F3S3	F4S1	F4S2	F4S3	F5S1	F5S2	F5S3	***SUM***
MM	1	4	2	5	6	6	5	6	5.5	1	3	1	1	1	1	***48.5***
Tanh	5	2	7	5	3.5	2.5	3	2	1	4	2	4	4	3	3	***51***
Zs	2.5	4	3	5	7	4	6	5	5.5	2.5	6	4	3	5	4	***66.5***
PL	6	7	5.5	2	3.5	2.5	3	4	5.5	5	5	4	7	6	5	***71***
LTL	4	1	1	1	1	1	1	1	3	6	1	6	6	2	6	***41***
RHE	2.5	6	4	5	3.5	5	3	7	5.5	2.5	7	2	2	4	2	***61***
FF	7	4	5.5	5	3.5	7	7	3	2	7	4	7	5	7	7	***81***

Table 4: Sum of performance mark based on EER, AUC and GAR(FAR=0.001)(x001gar)

MarkSum	MM	Tanh	Zscore	PL	LTL	RHE	FF
EER	48.5	51	66.5	71	41	61	***81***
AUC	48	73	66	***82***	38	68	45
001gar	39.5	49.5	59.5	***78.5***	68.5	59.5	65

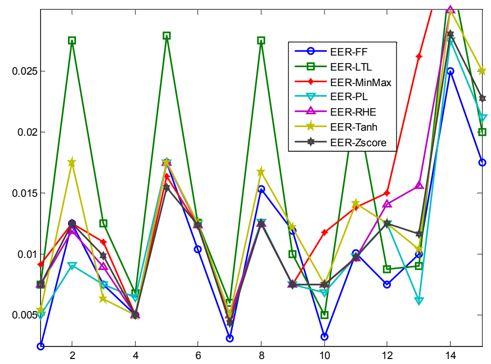

Figure 1. EER of different normalization techniques.

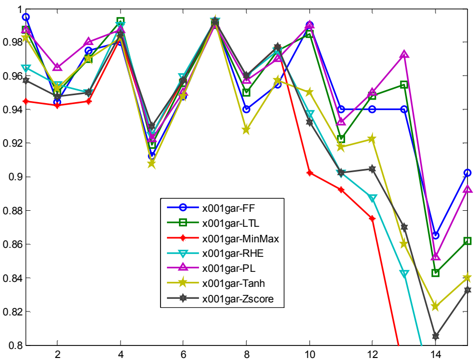

Figure 2. GAR(FAR=0.001) of different techniques.

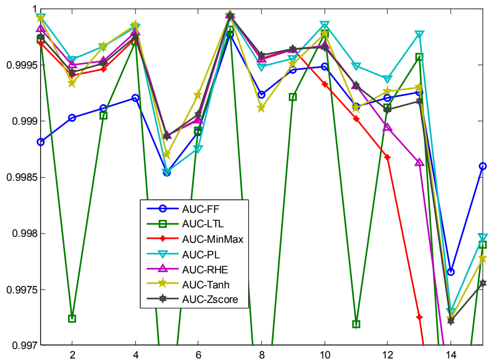

Figure 3. AUC of different normalization techniques

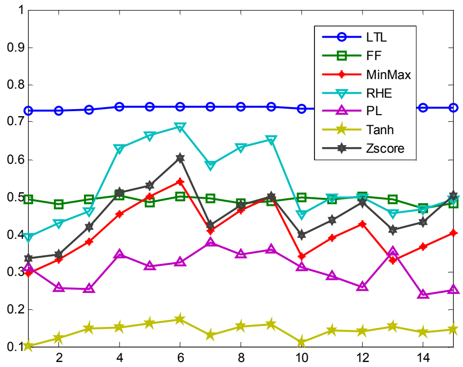

Figure 4. TE of different normalization techniques

Table 4 shows the sum of performance mark of different normalization techniques based on EER, AUC and GAR (FAR=0.001). From AUC aspects, PL, RHE and Zscore techniques outperforms other normalization methods. From the GAR (FRR=0.001), PL, LTL and FF normalization algorithms give better performance than other algorithms.In order to verify the stabilization of different normalization techniques, Fig 4 shows the Thresholds of EERs(TE). From Figure 4, we observe that the TE of FF normalization varies slowly and is about 0.5. Also, the TE of Tanh and LTL normalization techniques vary slowly also. FF, LTL and Tanh show better performance than other normalization methods from the change of TE.

In section 3, we have introduced that LTL and FSPL are the improvement of PL. In Ref. [6], LTL showed better performance than LTL normlization method, and LTL and FSPL outperformed PL nomalization method. However, our experimental results show that PL works better than LTL nomalizaiton method with EER, AUC and GAR(FAR=0.001).

5. Conclusions

The experimental results suggest that there is no single normalization technique that would perform the best for all multibiometric recognition systems. Four measures: EER, AUC, GAR(FAR=0.001) and the threshold of EER, are selected to evaluation of different normalization techniques. Different normalization functions should be choosing according to different applications. FF, PL and Zscore should be chosen if EER is the performance measure; PL, LTL and FF should be chosen if GAR(fixed FRR) is the performance measure; PL, Tanh and RHE should be chosen if AUC is the performance measure; FF, LTL and Tanh should be chosen if we want threshold of EER to be fixed. We can conclude that PL and FF normalization work better than other methods in many applications.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 60603015 and 60970034), the Foundation for the Author of National Excellent Doctoral Dissertation (Grant No. 2007B4), and the Scientific Research Fund of Hunan Provincial Education (the Foundation for the Author of Hunan Provincial Excellent Doctoral Dissertation).

References

[1] A. Ross, K. Nandakumar, A. Jain, Score normalization in multimodal biometric systems, Pattern Recognition 38 (2005):2270–2285.

[2] R. Tronci, G. Giacinto, F. Roli, Dynamic score combination of binary experts, 19th International Conference on Pattern Recognition, New York, 2008, 1-6: 2420-2423

[3] J.Kittler, M. Hatef, R.P.W. Duin, and J. Matas, On combining classifiers, IEEE Trans. on Pattern Anal. Machine Intell., 20(3):226-239 (1998)

[4] D.Tax, M.Breukelen, R. Duin,: Combining multiple classifiers by averaging or by multiplying, Pattern Recognition, 33:1475-1485 (2000)

[5] S. Ribaric, I.Fratric, A Matching-Score Normalization Technique for Multimodal Biometric Systems, Proc. 3rd COST 275 Workshop:Biometrics on the Internet, Hatfield, UK, 27-28 October 2005, pp. 55-58.

[6] Y.N. Singh and P. Gupta, Quantitative Evaluation of Normalization Techniques of Matching Scores in Multimodal Biometric Systems, ICB 2007.

[7] M.X. He, S.J. Horng, Performance evaluation of score level fusion in multimodal biometric systems, Pattern Recognition. 2010, 43(2): 1789-1800

[8] Li Yong, YIN Jian-ping, ZHU En, LI Kuan. Multibiometric Fusion Based on FAR and FRR. Acta Automatic Sinica 2011, 37(4):408-417(in Chinese)

[9] N. Poh, S. Bengio, Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication, Pattern Recognition, 39(2):223-233(2006)