Study of Score Fusion and Quality Weighting in the Bio-Secure DS2 Database

Saliha Artabaz[1] and Layth Sliman[2(✉)]

[1] Laboratoire de Méthodes de Conception de Systèmes LMCS, Ecole Nationale Supérieure d'Informatique ESI, Oued-Smar, Alger, Algeria
s_artabaz@esi.dz
[2] Ecole d'Ingénieur Généraliste en Informatique et Technologies du Numérique Efrei, 30–32 Avenue de la République, Villejuif, France
layth.sliman@efrei.fr

Abstract. A uni-biometric system suffers from unbalanced accuracy because of image quality, features extraction weakness, matching algorithm and limited degrees of freedom. This can be overcome by using multiple evidences of the same identity (Multi-biometrics fusion). In a previous work, we proposed new fusion functions based on arithmetic operators and search the best ones using Genetic Programming on the XM2VTS score database. The objective function is based on the Half Total Error Rate (HTER) (a threshold dependent metrics), from the Expected Performance Curve (EPC), of fused matching scores. In this paper, we select ten functions from the generated ones and apply them on matching scores of different biometric systems, which are provided by the bio-secure database. This database provide 24 streams that we use to generate 1000 multi-biometric combinations that we, then, use to conduct our comparative study. Since the result of fusion can be biased and requires a good quality assessment to evaluate the degree of reliability of a processed scheme, we use quality weights on the proposed functions and we compare the results with existing approaches. The proposed quality weights help to reduce the Equal Error Rate (EER a threshold-independent metric) since the obtained matching scores are results of different fusions of instances, sensors and evidences. The EER range is optimized along the tested functions. To confirm that our proposed functions give better score results than the existing functions based on arithmetic rules, we perform multiple statistical significance tests to check the reliability of our experimentation.

Keywords: Multi-biometrics · Fusion · Quality weights · Genetic Programming · Optimized search

1 Introduction

Multi-biometrics address several traditional biometric systems drawbacks. Mainly, they aim at reducing system errors. In fact, experiments show that combining different evidences enhance accuracy [1–6].

Multi-biometrics fusion considers different levels and evidences such as image, feature, score, decision and rank level. The most used level in the literature is the score

M. W. Berry et al. (Eds.): SCDS 2019, CCIS 1100, pp. 252–261, 2019.
https://doi.org/10.1007/978-981-15-0399-3_20

level [16, 17]. At this level, combined scores are easier to fuse and provide rich information at the same time [4]. Furthermore, the ease of accessing and accuracy that out-perform other levels make it the best level [17]. At this level, fusion considers matching scores: the result of comparing between different evidences, instances, provided by different sensors and processed with different algorithms. Hence, to identify the best system, the evaluation must take into account all these parameters. In fact, studying different systems is necessary to conclude whether the used strategy for fusion improve baseline systems accuracy or not. In this paper, we propose a comparative study between fusion of several biometric systems. To do that, we generate different combination of biometric systems from a score database and apply our generated fusion functions from a previous work [7].

As the quality is one of the main factors affecting the overall performance of biometric systems, using quality measurement can lead the fusion process to reach better results. In this paper, we are interested in the score fusion with quality weighting. We use an optimized generation using GP in a previous work [7] to get several fusion functions. In addition, we aim to optimize the score distribution by integrating the template-query quality as weights. We perform our experiments on the Biosecure score database [13]. The proposed approach outperforms the baseline uni-biometric systems. We conduct a comparative study between the best-computed fusion functions. In addition, we give a complete view of different multi-biometric systems to test the most reliable ones according to the Equal Error Rate.

This paper is organized as follows: First, in Sect. 2, we introduce the studied field.

Section 3 illustrates the used weighted fusion functions and introduce database build on to conduct experiments. After that, we give experimental results in Sect. 4 with comparative study between proposed fusion functions and studied multi-biometric systems. Finally, we conclude and list some perspectives of our work.

2 Multi-biometrics and Score Level Fusion

Multi-biometrics is a merged field that addresses unimodal biometric system weaknesses. Researchers who studied different fusion methods to assess their effectiveness, mostly affirm that Multi-biometrics improve the accuracy of baseline systems. The fusion is needed to enhance baseline systems accuracy or to face non-universality of all used modalities. To optimize Multi-biometrics fusion, many challenges must be handled. This includes multiple data source incompatibility, matchers' scores normalization, as well as the noise that affects system performances and can result in false positive or negative authentication. Fusion at the score level is the most used fusion [1] due to its low fusion complexity that outperforms other levels. However, data is facing the same challenges cited before (data source incompatibility, matchers' scores normalization and noise). Consequently, the quality measurement [14] is becoming inherent in biometric systems as it allows to predict biometric system performances. We can define sample quality [14] as "scalar quantity that is monotonically related to the performance of biometric matchers". The effectiveness of a sample quality evaluation and the different ways to provide this scalar can be found in [14].

Many recent research works [8–12, 15] are interested in the sample quality to get a well-adjusted fusion function, these works use the computed scalars for each sample used as weights. Other works consider the quality as a measure of the performance of a biometric system. The provided scalar from data source is used to select an adaptive solution according to sensed biometric signals, which can vary for each authentication [14]. The challenge in that case is to reach the best performances, as the signal quality is variable. In other case, the simple way is to consider the signal quality as a scalar that quantify relative signal degradation under noise. So, this scalar may be used as weight in a fusion process or contribute as an indicator in image enhancement, which is supposed to improve accuracy. *Theofanos* [17] studys the impact of image reconstruction on false acceptance and proposes a signal quality measurement to optimize it without increasing the false acceptance. In addition, the essential issue is to deal with unsupervised environment and different constraints to provide adapted systems that match with security requirements [8]. Therefore, the system quality measure can be seen as is a degree of trust that provides the reliability of the system and ensures its interoperability with regard to processed data heterogeneity [15]. All cited methods use weighting to control and take a decision depending on the data quality. Other proposals discussed using system reliability indicator [14] to estimate matcher weights.

Based on a single system, we can neither prove nor deny the effectiveness of the proposed fusion or weighting impact on system accuracy. In fact, getting excellent accuracy with a fusion strategy on a specific combination does not imply the effectiveness of this strategy. The question that we can ask is: *'given a predefined fusion strategy, in which cases the fusion strategy works better than initial fused systems'*. As well, we can look for the best combination and fusion strategy to meet security requirements. In addition, the challenge is to prove that the used fusion strategy is robust even employed for different multi-biometric systems

3 Materials and Methods

Choosing fusion method in score level is very crucial to enhance provided system performances. The rule-based function is the most relevant to be used thanks to its simplicity. In this paper, we provide experimental results of different fusion functions that we compute with GP in a previous work [7]. In our comparative study, we give an analysis of different multi-biometric systems. We test different combinations of biometric scores and compare between them. Our functions are constructed using Genetic Programming 'GP' that is based on tree structure where each node represents an operation. Using GP, we explore the search space of different trees that represent the fusion functions by applying mutation and crossover operations. The two operations apply modifications on nodes of the tree. The crossover modifies nodes of a selected tree using nodes' contents of another one. The mutation modifies randomly the nodes using other operations.

We can see in Fig. 1 the enhancement of the average HTER with GP simulation using crossover to evolve the population and roulette selection to get the best list. The HTER average is reduced to a range of [0.5%]. The graph oscillations show the progress of the GP in Fig. 1(a), which does not converge systematically.

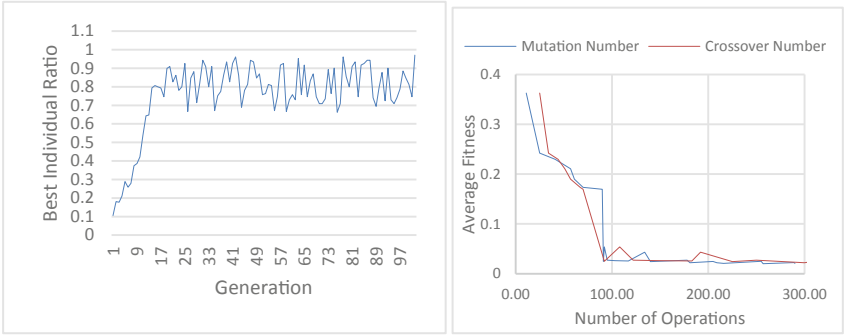

Fig. 1. Impact of mutation and crossover number on fitness average [7].

We select a number of trees obtained from the simulation. We use them to fuse different biometric systems of the Bio-secure database of scores. We use a user-specific weighting in order to improve baseline systems accuracy. The weight gives the sample fidelity to the claimed ID. We use the following formula to compute weight:

$$\left(\frac{1}{2} * \left((x_t - \bar{x})^2 + (x_q - \bar{x})^2\right)^{1/2}\right)^{-1} \tag{1}$$

where x_t is the template quality (claimed id) and x_q is the query quality (true id).

An insignificant value means that the two measurements are distinct and inversely. To normalize the quality, we use the standard deviation between max and min value computed from the development set.

4 Experiments and Discussion

We choose the Biosecure score database [13] as it is the only score database that offers a quality evaluation of its sets. We select a subset of functions that provide low HTER error in XM2VTS database. We take different functions along processed generations. Experiments are done according to these steps:

1. Generate 1000 different configurations to select 8 scores from 24 scores of the database. This allows testing different combinations of biometric systems in order to select the best ones. The scores must be filtered to get subset that contains sufficient number of data;
2. Test the baseline systems of these configurations;
3. Test each function for:
   (a) fusion of the baseline scores;
   (b) fusion using the weights scalars on normalized scores.
4. Compare between functions and analyze statistically the improvement of EER comparing to usual operators used for fusion.

Here is the list of functions used for our experimental study. These functions were selected according to their accuracy and number of fused scores.

Table 1. List of used functions for fusion.

Identifier	Function
Fct1	`avg (S1, avg (min (min (S2, S3) + S4 * S5, S6), min (S7, S8)))`
Fct2	`S1 + avg (min (avg (S2 + S3, min (S4, S5)), S6), min (S7, S8))`
Fct3	`S1 + avg (max (avg (S2 + S3, avg (S4, S5)), S6), min (S7, S8))`
Fct4	`S1 + avg (min ((min (S2, S3)) + S4 + S5, S6]), min(S7, S8))`
Fct5	`avg (max (max (S1, S2), (S3 + ((S4 * S5)- S6))), avg (S7, S8))`
Fct6	`max (avg (S1, S2), S3) + avg (S4, S5) + min (S6, S7) + S8`
Fct7	`avg (S1, S2) * S3 + avg (S4, S5) + min (S6, S7) + S8`
Fct8	`S1 + avg (max (avg (S2 + S3, S4 + S5), S6), min(S7, S8))`
Fct9	`avg (avg (min (S1, min (max (S2, S3), S4)), S5), min (avg (S6, S7), S8))`
Fct10	`avg (avg (S1, avg (S2, S3)), min (S4, S5) + S6 + S7 + S8);`

We achieve these experiments using the Biosecure protocol with the selected functions cited in Table 1 above. This protocol uses datasets of two sessions with different impostors.

Figure 2 compares between some statistics of the tested functions upon the evaluation set. We rank functions according to the variance of the used statistics between the two sessions to evaluate quality of each function. Then, we use the sum of these ranks to get the best ones. We can observe that the best functions are respectively Function 3, 8, 5 and 2 whose average of ranks does not exceed rank 5. The function 3 applied gives small Standard Deviation on a limited range of EER values (max EER = 33%). To verify our results, we take, as an example, selected input scores. In Fig. 3, we see that function 2 is more relevant in this case since it gives the lowest variance with EER = 0.85% on session 2 of the dataset.

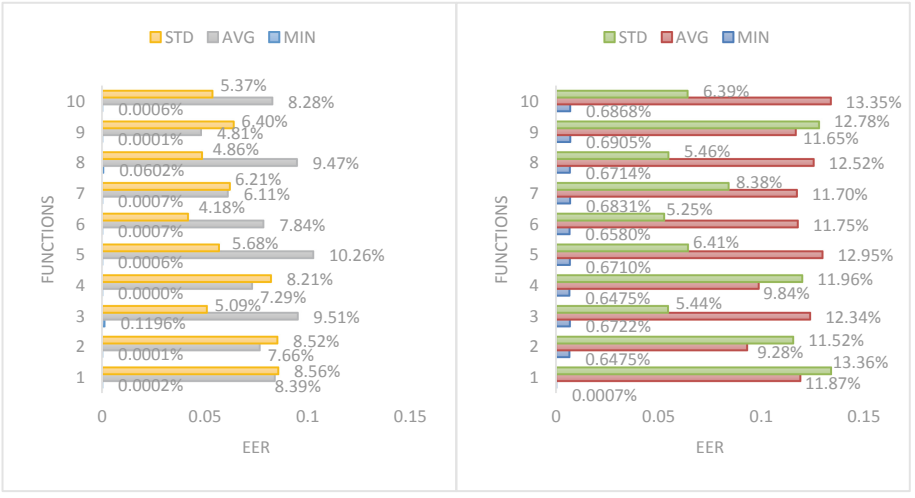

Fig. 2. MIN, AVG and STD equal error rate of the studied functions on the two sessions of the evaluation set.

For instance, in Fig. 3, which shows the multi-biometric system minimizing the EER, we can see that function 1 gives the lowest EER and the best functions EER is near the computed average since the standard deviation do not exceed 10%. The best combination in this case is function 1, 2 and 4 if we consider the EER and disparity between the two sessions of the evaluation set at the same time.

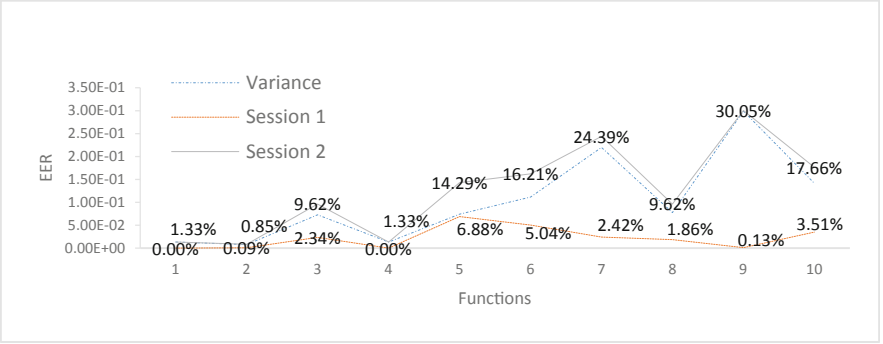

Fig. 3. Results of the multi-biometric system with the lowest EER on the two sessions of the evaluation set.

The example illustrated in Fig. 4 shows that score fusion in these cases allows to reduce errors caused by divergent scores. Indeed, functions 2, 5 and 8 give already good results and optimize Area Under Curve. For instance, the fused scores represent respectively: Face (CANON), Face (CANON with Flash), Iris, two fingerprints taken with different devices as described in the database. The other scores are filtered for dummies values included in the tested database.

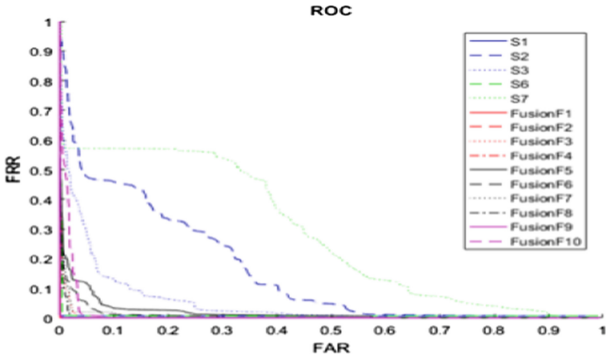

Fig. 4. ROC curves (EER = 0.8% for Function 2) of all used functions compared to a subset of fused scores and baseline systems.

To analyze the effectiveness of our results, we perform a significance test to compare between applying fusion of normalized data and one of the proposed weighted

fusion function that give the best statistics. We perform a t-test on two samples of equal averages and unequal variances for the same sample size. The obtained probability (=0.027) is under the tail probability fixed to 0.05. The obtained t value 1.9607 is found to be less than the standard t value 2.2126. Therefore, statistically, we conclude that our approach is significantly better than the fusion applied with normalization.

In order to show our results, we present the whole combined systems in Fig. 5. The figure shows the ratio of multi-biometric systems depending on the number of functions that give an EER less than the referenced EER. The ratio of multi-biometric systems decreases gradually according to the number of functions with an EER under a predefined value. For example, five functions reduce the EER of 21% of multi-biometric systems below 5%. For 26% of the systems, we get a function that reduce the EER below 0.1% (only 3% of these systems get an insignificant error (less than 0.001%) for only one function).

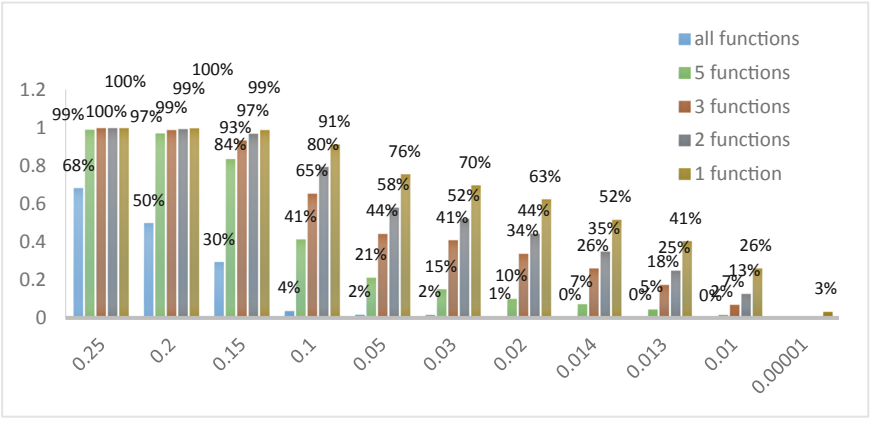

Fig. 5. Cumulative ratio of multi-biometric systems depending on the number of functions verifying the corresponding EER.

From the above analysis, we can observe that the improvement is guaranteed comparing to initial operators (Sum, Product, Min, Max) since the EER provided by our functions is less than the one achieved by the best operator for 88.6% of multi-biometric systems. To prove that at least one of the proposed functions outperforms the usual fusion operators, we must use t-test to verify whether the difference between the outputs is significant. To do so, as a first assessment, we use a paired test to analyze average difference in one direction without taking into consideration the variance. As a result, a t-value equal to 1.63 is obtained which is less than the critical t-value (equal to 1.64). Hence, we can conclude that the hypothesis of significant improvement can be assumed (i.e. P-value <0.1). Consequently, the proposed fusion is significantly better than usual fusion operators with 99% confidence interval. As a second assessment, we use a paired test to analyze average difference assuming that the two variances are different. The test is almost successful (P-value = 0.05 < 0.1 and t-value = 1.58 < 1.64).

In our study, we reach an EER under 0.013% and we get two multibiometric systems that optimize the EER to the range of [0.12%, 1.30%] from baseline scores with EER in the range of [1%, 99.55%] (see Fig. 6). This means that we reach a range improvement of 98%. As a result, we can conclude that our functions outperform the results obtained in [16] using RS-ADA, on the same database, for fusion that reaches an EER equal to 1.98%.

Table 2. Multi-biometric systems details.

	Face	Fingerprints with the same sensor for template and query	Fingerprints with different sensors for template and query
1	Webcam (low resolution) LDA-based face verifier	Thermal: right/left thumb, right index, Optical: left thumb, right index NIST fingerprint system	Left index, left middle finger NIST fingerprint system
2	Webcam (low resolution) LDA-based face verifier	Thermal: right thumb, right index, Optical: right/left thumb, right index	Left index, left middle finger

Table 2 gives details of the fused evidences in these resulting systems (see Fig. 6). The selected multi-biometric systems use face, multiple instances of fingerprint and multiple captures using different sensors. Furthermore, the fusion is done on scores comparing between fingerprint queries and templates taken with different sensors.

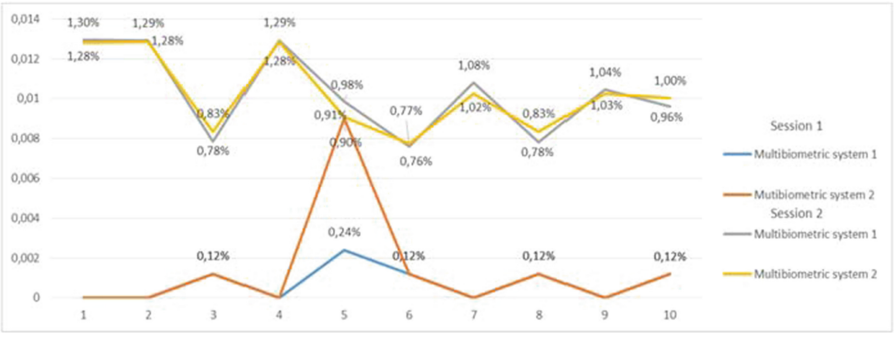

Fig. 6. Equal Error Rate for the best two multibiometric systems (EER less than 1.3% for all functions).

5 Conclusion

In this paper, we study some generated functions based on primitive fusion rule. These are the result of applying Genetic Programming proposed in a previous work to get the best rules combination using the XM2VTS. Afterward, we apply weighting on the generated functions and then we perform experiments on the Biosecure score database to compare between different combinations of the provided scores and find the best solution for fixed range of EER. The significance test confirm the improvement comparing to usual fusion operators. The provided functions can be tested with multi-algorithm biometric systems or using other databases in order expand the study and validate the results. We aim, later, to study fusion of features and classifiers using the proposed functions.

References

1. AlMahafzah, H., AlRawashdeh, M.Z.: Performance of multimodal biometric systems at score level fusion. In: Zeng, Q.-A. (ed.) Wireless Communications, Networking and Applications. LNEE, vol. 348, pp. 903–913. Springer, New Delhi (2016). https://doi.org/10.1007/978-81-322-2580-5_82
2. Kadam, A., Ghadi, M., Chavan, A., Jawale, P.: Multimodal biometric fusion. Int. J. Eng. Sci. Comput. **7**(5), 12554–12558 (2017)
3. Anzar, S.M., Sathidevi, P.S.: Optimization of integration weights for a multibiometric system with score level fusion. In: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13–15, 2012, Chennai, India – vol. 2. Springer, Heidelberg, pp. 833–842 (2013). https://doi.org/10.1007/978-3-642-31552-7_85
4. Conti, V., Militello, C., Sorbello, F., Vitabile, S.: A frequency-based approach for features fusion in fingerprint and iris multimodal biometric identification systems. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**, 384–395 (2010)
5. Damer, N., Opel, A.: Multi-biometric score-level fusion and the integration of the neighbors distance ratio. In: Campilho, A., Kamel, M. (eds.) ICIAR 2014. LNCS, vol. 8815, pp. 85–93. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11755-3_10
6. Eskandari, M., Toygar, Ö.: Score level fusion for face-iris multimodal biometric system. In: 2013 Proceedings of the 28th International Symposium on Computer and Information Sciences, pp. 199–208. Springer, Cham. https://doi.org/10.1007/978-3-319-01604-7_20
7. Artabaz, S., Sliman, L., Benatchba, K., Dellys, H.N., Koudil, M.: Score level fusion scheme in hybrid multibiometric system. In: Badioze Zaman, H., et al. (eds.) IVIC 2015. LNCS, vol. 9429, pp. 166–177. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25939-0_15
8. Alonso-Fernandez, F., Fierrez, J., Bigun, J.: Quality Measures in Biometric Systems. Encyclopedia of Biometrics, pp. 1287–1297. Springer, US (2015). https://doi.org/10.1007/978-3-642-27733-7
9. Alonso-Fernandez, F., Fierrez, J., Ramos, D., Gonzalez-Rodriguez, J.: Quality-based conditional processing in multi-biometrics: application to sensor interoperability. IEEE Trans. Syst. Man Cybern. - Part A Syst. Hum. **40**, 1168–1179 (2010)
10. Mohammed Anzar, S.T., Sathidevi, P.S.: On combining multi-normalization and ancillary measures for the optimal score level fusion of fingerprint and voice biometrics. EURASIP J. Adv. Signal Process. **1**, 1–17 (2014)

11. Zhang, D., Lu, G., Zhang, L.: Finger-knuckle-print verification with score level adaptive binary fusion. Advanced Biometrics, pp. 151–174. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-61545-5_8
12. Poh, N., Bourlai, T., Kittler, J.: A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. Pattern Recogn. J. **43**(3), 1094–1105 (2009)
13. Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 531–543 (2007)
14. Kabir, W., Ahmad, M.O., Swamy, M.N.S.: Score reliability based weighting technique for score-level fusion in multi-biometric systems. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–7 (2016)
15. Lip, C.C., Ramli, D.A.: Comparative study on feature, score and decision level fusion schemes for robust multibiometric systems. Frontiers in Computer Education. Springer, Heidelberg, pp. 941–948 (2012). https://doi.org/10.1007/978-3-642-27552-4_123
16. Lumini, A., Loris, N.: Overview of the combination of biometric matchers. Inf. Fusion **33**, 71–85 (2017). Elsevier
17. Theofanos, M.: Biometrics systematic uncertainty and the user. In: Proceeding of Biometrics: Theory Applications and Systems (BTAS 2007), IEEE, pp. 1–6 (2007)