P-score: Performance Aligned Normalization and an Evaluation in Score-level Multi-biometric Fusion

Naser Damer*, Fadi Boutros*, Philipp Terhörst*, Andreas Braun*, Arjan Kuijper*[†]
*Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[†]Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
Email:{naser.damer, fadi.boutros, philipp.terhoerst, andreas.braun, arjan.kuijper}@igd.fraunhofer.de

Abstract—**Normalization is an important step for different fusion, classification, and decision making applications. Previous normalization approaches considered bringing values from different sources into a common range or distribution characteristics. In this work we propose a new normalization approach that transfers values into a normalized space where their relative performance in binary decision making is aligned across their whole range. Multi-biometric verification is a typical problem where information from different sources are normalized and fused to make a binary decision and therefore a good platform to evaluate the proposed normalization. We conducted an evaluation on two publicly available databases and showed that the normalization solution we are proposing consistently outperformed state-of-the-art and best practice approaches, e.g. by reducing the false rejection rate at 0.01% false acceptance rate by 60-75% compared to the widely used z-score normalization under the sum-rule fusion.**

I. Introduction

Normalization is essential in a wide range of statistical based solutions where measurements have different origins, or are calculated on different scales. Information fusion, more specifically multi-biometric fusion, is one of these applications where combined information should have a common measure of inference in the fusion process [1]. Biometric verification is a typical binary classification problem where comparisons between reference and probe captures have to be classified into genuine or imposter comparisons. This is usually achieved by thresholding the comparison score between these captures.

Previous normalization solutions focused on aligning the measurement values (scores) of different origins to fit in a predefined range. Other solutions extended this concept to bring these values to common distribution characteristics [2]. A modification to some of the traditional normalization approaches have been proposed to anchor one point in the values distributions based on the performance it induces [3]. However, due to the diverse performance behavior around these single performance points, the performance alignment between different sources is not achieved on other points.

This work presents a normalization approach aiming at transferring the score values into a space where a similar value from a difference source will induce similar relative performance, and thus its interpretation in further processing steps (e.g. fusion). This is done by considering the relation between the performance and score values, represented by the half total error rate (HTER) at different score thresholds, and trying to normalize these values so they will achieve such a relationship with common properties.

The proposed normalization approach can be utilized in any problem where the normalized values aim at influencing a binary decision, such as biometric verification. The proposed approach is evaluated along with a number of state-of-the-art and baseline approaches on two publicly available multi-biometric score databases, the XM2VTS LP1 and LP2 [4]. The evaluation results proved consistent superiority of the proposed approach under different experimental settings with the false rejection rate (FRR) at 0.01% false acceptance rate (FAR) reduced by 60% and 75% on the LP1 and LP2 databases when compared to the z-score normalization and by 71% and 81% when compared to min-max normalization under the sum-rule fusion.

II. Related Work

Different normalization approaches have been suggested in statistical applications, some focus on a unified scale and others extended this into achieving a common probability distribution. Min-max normalization is a simple technique to rescale the range of data to fit into a [0,1] range and it only depends on the minimum and maximum values of the training data, which makes it vulnerable to outliers. It also does not consider the nature of the distribution of the values. Due to its simplicity, min-max normalization was used regularly in score-level multi-biometric fusion [5][6][7]. The median absolute deviation normalization (MAD) tries to capture information about the values distributions by considering their median and median absolute deviation assuming a Gaussian distribution of the data. MAD was popularized by the work of Hampel in 1974 [8][9]. Just like MAD normalization, z-score normalization tries to unify the normalized values distribution based on their standard deviation and mean value [2]. z-score transforms the values to have a standard deviation of 1 and a mean value of zero. TanH was introduced by Hample et al. [10] and is based on the standard deviation and mean value of the genuine comparisons values in a way that reduces the influence of the points at the tails of the distribution. Just like the previously presented normalization approaches, TanH construct no clear link between the normalized values and their influence on the performance and thus their real effect in a fused system. A work by Jain et al. [6] provided a thorough study of these classical normalization approaches and their

ISBN 978-90-827970-1-5 © EURASIP 2018

effect on the multi-biometric fusion process emphasizing the normalization critical role.

A simple link between comparison score normalization and multi-biometric verification performance was previously proposed by Damer et al. [3]. This work described modifications to min-max, TanH, and MAD normalizations that align one performance related point (anchor), namely the threshold at equal error rate (EER), in the distributions of different sources and didn't consider the relation between the scores and the performance beyond this point. These approaches were noted as performance anchored normalization (PAN) and the three variations are abbreviated as PAN-min-max, PAN-TanH, and PAN-MAD. Two works by Kabir et al. proposed modifications to the PAN-min-max normalization by proposing anchor values calculated from comparison scores after neglecting possible outliers [11][12]. The first proposed anchored value is based on the mean comparison score value (noted as anchored min-max, AMM [12]) and the second included the standard deviation of these values (improved anchored min-max, IAMM [11]). However, these works didn't consider any link between score values and performance.

III. Methodology

The goal of this work is to normalize biometric comparison scores so that common values from different sources would have common relative inference on performance, and thus a common interpretation. This also applies to normalizing any measurement that is intended to be used in a binary stump decision. To achieve this goal, score values have to be transformed so that a certain relation between these values and a measure of performance is similar after normalization for different sources. A good candidate for such a relation is the one between the score values (seen as a decision threshold) and the HTER as a measure of performance.

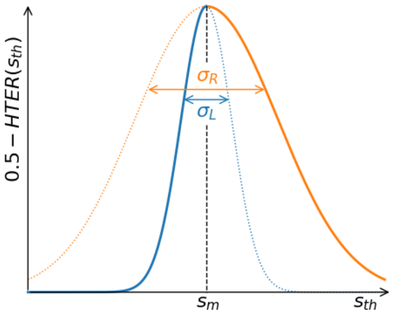

Fig. 1: Modeling the relation between performance (HTER) and score values (s_{th}) shown by the solid curve. The dotted curves extends to the left and right to simulate two Gaussian distributions and their standard deviations, σ_L and σ_R.

HTER is the average of the two trade-off error rates, the FAR and FRR of a biometric verification system (or any binary decision maker) at a certain operation point (decision threshold). This corresponds to the false positive rate (FPR) and false false negative rate (FNR) rates usually used to evaluate binary classification systems. The $HTER(s_{th})$ is a function of decision threshold s_{th}, which corresponds to a comparison score threshold in biometric verification. A distinctive value in the $HTER(s_{th})$ is the minimum half total error rate (minHTER) value that occurs at the certain decision threshold $s_{th} = s_m$.

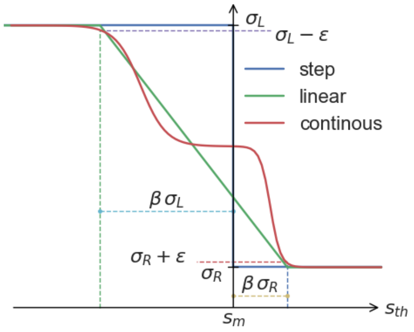

Fig. 2: Different strategies to assign the normalization standard deviation σ_N around s_m given σ_L and σ_R.

An example of a typical $HTER(s_{th})$ is shown in Figure 1, to facilitate an easier explanation this figure plots $0.5 - HTER(s_{th})$. The maximum HTER value occurs on the edges of the score values range and is equal to 0.5, this is because a threshold set at the edges of the range would produce a very biased decision leading to one of the trade-off errors (FAR and FPR) to be 100% and while the other is 0%. Splitting the $0.5 - HTER(s_{th})$ curve into two parts along the vertical line $s_{th} = s_m$ results in two curves (right and left) seen in solid line (orange and blue) in Figure 1. Each of these curves can be modeled as a part of a Gaussian distribution and thus can be associated with a standard deviation (σ_R and σ_L) as follows given that $\mathcal{S}$ set of all scores s_{th}

$$\sigma_L^2 = \frac{2}{\left(\sum_{s_i \in \mathcal{S}_L} q(s_i)\right) - 1} \sum_{s_i \in \mathcal{S}_L} \left(q(s_i)(s_i - s_m)^2\right), \quad (1)$$

$$\mathcal{S}_L = \{s_i \in \mathcal{S} | s_i = s_{min} + d \cdot i \leq s_m, \forall i \in \mathbb{N}_0\}. \quad (2)$$

$$\sigma_R^2 = \frac{2}{\left(\sum_{s_i \in \mathcal{S}_R} q(s_i)\right) - 1} \sum_{s_i \in \mathcal{S}_R} \left(q(s_i)(s_i - s_m)^2\right), \quad (3)$$

$$\mathcal{S}_R = \{s_i \in \mathcal{S} | s_i = s_{max} - d \cdot i \geq s_m, \forall i \in \mathbb{N}_0\}, \quad (4)$$

where

$$q(s_i) = 0.5 - HTER(s_i), \quad (5)$$

$$d = \frac{|s_{max} - s_{min}|}{N}, \quad (6)$$

where the number of samples measured from the HTER curve is N, and have to be chosen large enough to enable accurate estimation of σ_L and σ_R (in our experiments, N=1000). s_{min} and s_{max} are the minimum and maximum comparison scores in the development data of the biometric source.

Taking the previous definitions into account, a transformation should be defined to achieve a common relation between

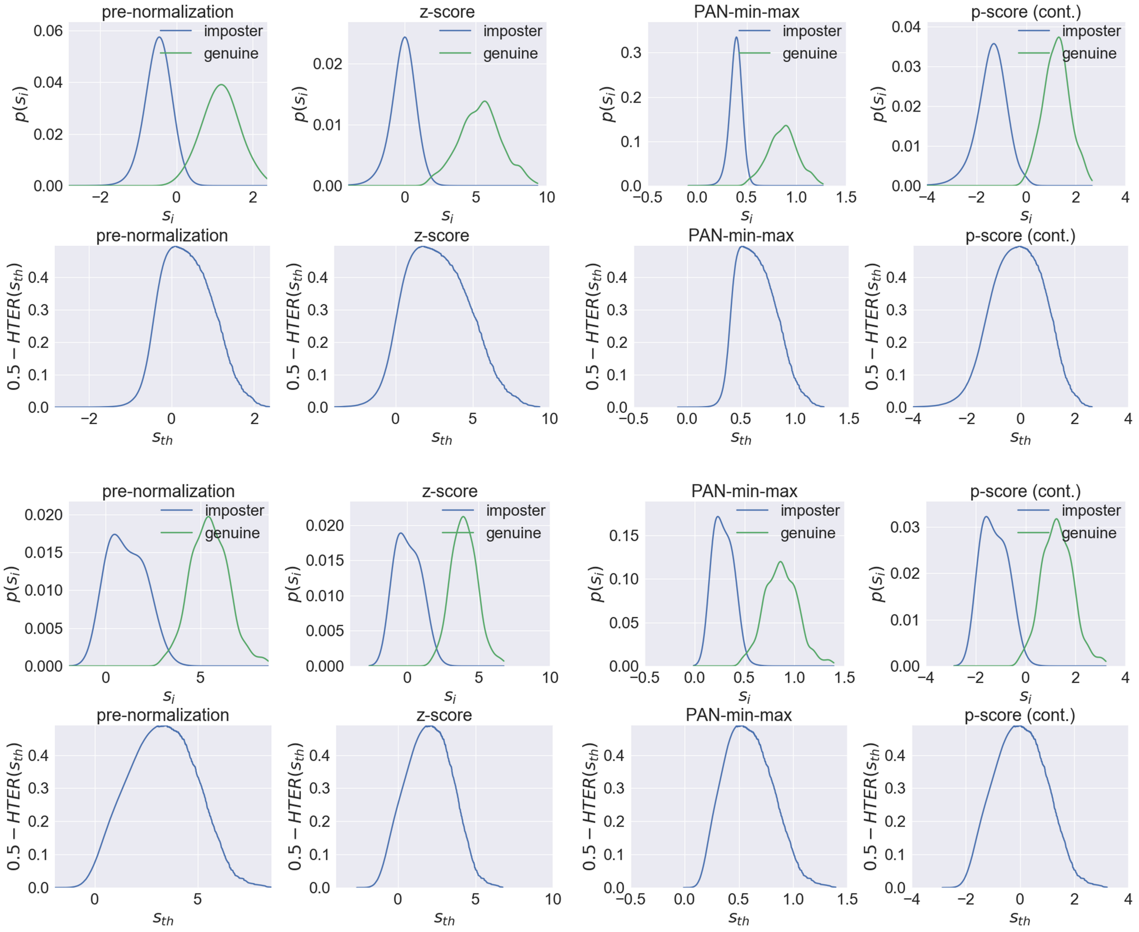

Fig. 3: The genuine and imposter score distributions and the $HTER(s_{th})$ of two biometric sources before normalization, after z-score, PAN-min-max, and the proposed p-score (cont.) normalization. The two biometric sources are from the XM2VTS LP2 database, the top 8 plots are of the XM2VTS-PL2 face expert DCTb-GMM, the bottom eight plots are of the XM2VTS-PL2 voice expert LFCC-GMM [4]. Notice the similarity between the plots after p-score normalization from both sources, which is not the case for other normalizations.

score values and performance (HTER), and thus a common $0.5 - HTER(s_{th})$ curve. A computationally convenient common $0.5 - HTER(s_{th})$ function can be initially based on one with a standard deviation of one and a mean value of $s_m = 0$. To achieve this, a transformation similar to the z-score normalization can be defined to transform any score value $s_i \in \mathcal{S}$ into $s'_i \in \mathcal{S}'$ in the normalized space

$$s'_i = \frac{s_i - s_m}{\sigma_N} \cdot \frac{1}{1 - 2 \cdot minHTER}, \tag{7}$$

where σ_N is a normalization standard deviation derived from the properties of the $HTER(s_{th})$ function. The second term uses minHTER to adapt to sources with different levels of performance and thus different heights of $HTER(s_{th})$ peaks. An obvious choice of σ_N is σ_L for the score values below s_m and σ_R for the values higher than this threshold as follows

$$\sigma_N = \begin{cases} \sigma_R & \text{if } s_i \geq s_m \\ \sigma_L & \text{if } s_i < s_m \end{cases}, \tag{8}$$

which will be referred to as a $step$ transition and it may lead to some irregularities around the performance sensitive area around s_m. To overcome this step change in σ_N values, two additional forms of transition are proposed, a $linear$ one and a $continuous$ one. For a linear transition, σ_N is described as

$$\sigma_N = \begin{cases} \sigma_L & \text{if } s_i \leq s_m - \beta\sigma_L = r_L \\ \frac{(\sigma_R - \sigma_L)(s_i - s_m + \beta\sigma_L)}{\beta(\sigma_R + \sigma_L)} + \sigma_L & \text{if } r_L < s_i < r_R \\ \sigma_R & \text{if } s_i \geq s_m + \beta\sigma_R = r_R \end{cases} \tag{9}$$

here β controls how fast do the vale reach σ_R and σ_L when moving away from s_m. For a smoother transition around s_m a continuous function is defined based on an asymmetrical shifted sigmoid as

$$\begin{aligned} \sigma_N =& g\left(s_i \,\middle|\, \sigma_L,\, \tfrac{\sigma_L + \sigma_R}{2},\, s_m - \beta\sigma_L,\, s_m,\, \epsilon\right) - \tfrac{\sigma_L + \sigma_R}{2} \\ &+ g\left(s_i \,\middle|\, \tfrac{\sigma_L + \sigma_R}{2},\, \sigma_R,\, s_m,\, s_m + \beta\sigma_R,\, \epsilon\right), \end{aligned} \tag{10}$$

where

$$g(x|A,B,c,d,\epsilon) = \frac{|A-B|}{1+\exp\left[h(x|A-B,c,d,\epsilon)\right]} + \min(A,B), \tag{11}$$

$$h(x|K,c,d,\epsilon) = \mathrm{sign}(K)(2\cdot\frac{x-c}{d-c}-1)\log(\frac{|K|}{\epsilon}-1)\ , \tag{12}$$

holds for for $|A-B|> 2\epsilon$. β controls how fast do the value reachs σ_R and σ_L when moving away from s_m, while ϵ is a small number ($\epsilon > 0$) controls how horizontal (flat) is the function around s_m and is a function of σ_L and σ_R, $\epsilon = \gamma|\sigma_L - \sigma_R|$ to enhance adaptability. In the experiments carried out in this work, these parameters where set so that $\beta = 0.5$ and $\gamma = 0.001$, changing both parameters in the range $\beta = [0.25, 0.75]$ and $\gamma = [0.01, 0.0001]$ didn't produce significant or systematic effect on performance.

Figure 2 visualizes how σ_N develops from σ_L to σ_R around s_m in a $step$, $linear$, and $continuous$ function. Based on these three approaches to calculate σ_N and Equation 7, scores can be normalized so that each score value from any source will have a relatively similar performance influence (if used as a decision threshold) and thus have a similar interpretation. This proposed performance aligned normalization approach will be referred to as p-score.

IV. Experiment and Results

Experiment Settings: for the development and evaluation of the proposed solution, two parts of the XM2VTS score-level fusion Database were used, LP1 and LP2 [4]. LP1 and LP2 contain comparison scores by different face and voice baseline algorithms. LP1 contains eight score sources (5 face algorithms and 3 voice algorithms) while LP2 contains five sources (2 face algorithms and 3 voice algorithms). LP1 used three training captures per subject while LP2 used four. For more details about the face and voice comparison algorithms that produced the scores, one can refer to the work of Poh et al. [4]. The publicly available database contains two parts, developments and evaluation. Score normalization parameters for each biometric source, in both LP1 and LP2, were obtained from the development data. The results and performance evaluation discussed later are obtained from the evaluation data. The evaluation is performed by fusing all available sources in each database under the verification scenario.

Different normalization approaches mentioned in Section II are considered. Namely, these approaches are Min-max [6], MAD [8], z-score [2], TanH [10], PAN-min-max [3], PAN-MAD [3], PAN-TanH [3], AMM [12], and IAMM [11]. After normalization, the comparison scores are fused by simple sum-rule fusion [6] and two weighted-sum fusion approaches, the equal error weighting (EERW) [6], the overlap standard deviation weighting (OLDW) [13]. The weights are calculated based on the development data. A more detailed description of each of the fusion and normalization baselines can be obtained form the corresponding references.

Results: a visual comparison between some of the baseline normalization techniques and the proposed one is presented in Figure 3. This figure shows the genuine/imposter score distributions for two biometric sources before normalization and after normalization by z-score, PAN-min-max and the proposed p-scor (cont.). The performance relation to the score values is also presented as plots of the $HTER(s_{th})$. The $HTER(s_{th})$ plots of both sources after p-score normalization visually similar, unlike other normalization approaches, which confirms the goal of the proposed normalization by bringing different sources into a common mapping between score values and performance. This mapping is reflected on the score distributions that have similar properties after p-score normalization.

The verification performance of the proposed solution is presented as FRR values at fixed FARs. This allows performance comparison at different operation points and might be of interest for users with different performance needs. The minHTER is also presented as a general and comparable one-value measure of the evaluation performance, lower minHTER values corresponds to higher performance. These values are listed in Table I for all the experiment setups discussed earlier, including the two databases, three fusion approaches, nine baseline normalizations, and the three versions of the proposed p-score normalization. The table shows that the proposed p-score (cont.) achieved the lowest FRR and lowest minHTER consistently, coming second only in one out of eighteen measures. Other normalization approaches had mixed performances across databases and experimental settings such as the PAN-TanH performing good on the LP1 database where the AMM performed relatively poorly, but vice versa on the LP2. Without a weighting influence, the proposed p-score (cont.) reduced the FRR at 0.01%FAR for the LP1 and LP2 databases by 71% and 60% compared to PAN-min-max, 75% and 77% compared to TanH, and 60% and 75% compared to z-score. Weights induce more information into the fusion process and thus bringing the performance to its limits, such as in the OLDW case, which leads to a closer performance by different normalizations. The behavior of σ_N have to adhere to two criteria, first is to be smooth close to s_m to assure no sudden change in this performance sensitive area, and second, to move fast to σ_R and σ_L when moving away from s_m to achieve the goal of a common score value-performance relation. The continuous solution satisfies the first criteria best, the step satisfies the second one, and the linear satisfies both criteria poorly. This can explain the results where the continuous solution performed best, closely followed by the step and the linear ones.

V. Conclusion

This work presented a new outlook on the normalization problem essential to prepare information for further processing in fusion or classification. Assuming that the normalized values will take part in a binary decision making process, the normalization approach we presented here transfers the normalized values from different sources into a space where

		sum-rule			EERW-sum			OLDW-sum		
		0.01%FAR	0.001%FAR	minHTER	0.01%FAR	0.001%FAR	minHTER	0.01%FAR	0.001%FAR	minHTER
XM2VTS-LP1	min-max	3.50%	8.00%	0.540%	3.50%	4.75%	0.432%	**1.00%**	2.25%	0.178%
	MAD	11.25%	13.50%	1.568%	12.50%	14.75%	1.584%	**1.00%**	2.25%	0.178%
	TanH	4.00%	8.50%	0.492%	2.50%	4.50%	0.379%	4.00%	6.75%	0.344%
	z-score	2.50%	5.75%	0.473%	1.25%	3.25%	0.346%	**1.00%**	2.25%	0.178%
	PAN-min-max	3.50%	6.25%	0.515%	1.50%	4.25%	0.405%	**1.00%**	**2.00%**	0.190%
	PAN-TanH	**1.00%**	**3.75%**	0.355%	**0.75%**	**2.00%**	0.277%	**1.00%**	2.25%	0.178%
	PAN-MAD	3.00%	6.25%	0.531%	3.75%	4.50%	0.483%	**1.00%**	2.25%	0.178%
	AMM	2.25%	6.00%	0.447%	1.25%	3.50%	0.392%	**1.00%**	2.25%	0.176%
	IAMM	2.50%	5.75%	0.449%	1.25%	3.50%	0.374%	**1.00%**	2.25%	0.179%
	p-score (step)	**1.00%**	**4.50%**	**0.326%**	**0.75%**	**2.00%**	**0.248%**	**1.00%**	**2.00%**	**0.169%**
	p-score (linear)	**1.00%**	6.25%	0.337%	**0.75%**	3.50%	0.267%	1.25%	2.50%	0.172%
	p-score (cont.)	**1.00%**	**4.50%**	**0.325%**	**0.75%**	**2.00%**	**0.247%**	**1.00%**	**2.00%**	**0.168%**
XM2VTS-LP2	min-max	2.75%	3.75%	0.166%	3.75%	4.25%	0.264%	**0.25%**	**0.25%**	0.022%
	MAD	12.00%	19.00%	1.518%	12.00%	17.50%	1.516%	**0.25%**	**0.25%**	0.022%
	TanH	2.25%	2.50%	0.409%	2.50%	3.00%	0.209%	1.00%	2.25%	0.125%
	z-score	2.00%	3.50%	0.145%	2.75%	3.75%	0.220%	**0.25%**	**0.25%**	0.022%
	PAN-min-max	1.25%	2.75%	0.215%	2.00%	3.50%	0.265%	**0.25%**	**0.25%**	0.036%
	PAN-TanH	1.25%	3.25%	0.071%	1.50%	3.50%	0.188%	**0.25%**	**0.25%**	0.022%
	PAN-MAD	2.75%	3.75%	0.179%	3.25%	4.25%	0.267%	**0.25%**	**0.25%**	0.022%
	AMM	1.00%	3.00%	0.112%	1.75%	3.00%	0.193%	**0.25%**	**0.25%**	0.025%
	IAMM	1.00%	3.25%	0.063%	1.50%	3.50%	0.192%	**0.25%**	**0.25%**	**0.021%**
	p-score (step)	**0.75%**	**1.75%**	**0.017%**	**0.50%**	**1.00%**	**0.102%**	**0.25%**	**0.25%**	**0.021%**
	p-score (linear)	**0.50%**	**1.50%**	**0.017%**	**0.50%**	1.25%	0.107%	**0.25%**	**0.25%**	0.022%
	p-score (cont.)	**0.50%**	**1.75%**	**0.017%**	**0.50%**	**1.00%**	**0.096%**	**0.25%**	**0.25%**	**0.019%**

TABLE I: FRR values achieved at fixed FAR, and minHTER values for the different experiment settings on two databases. The best rates across normalization techniques (columns) are in bold, the second best values also in bold (if the first one occurred less than three times). Notice the consistency of the proposed p-score performance.

they have a common relation between their values and their relative induced performance. We evaluated our approach under a multi-biometric score-level fusion scenario, where score normalization plays a major role in regulating the influence of the multiple scores in the final verification decision. We conducted evaluation on two publicly available databases and the results showed consistent high performance of our proposed p-score normalization over baseline and state-of-the-art solutions, with the FRR at 0.01%FAR reduced on the two databases by 71% and 81% in comparison to min-max and by 60% and 75% in comparison to z-score under the sum-rule fusion.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within CRISP.

References

[1] N. Damer, A. Opel, and A. Shahverdyan, "An overview on multi-biometric score-level fusion - verification and identification," in *ICPRAM 2013 - Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods, Barcelona, Spain, 15-18 February, 2013*. SciTePress, 2013, pp. 647–653.

[2] E. Kreyszig, *Advanced Engineering Mathematics: Maple Computer Guide*, 8th ed. New York, NY, USA: John Wiley & Sons, Inc., 2000.

[3] N. Damer, A. Opel, and A. Nouak, "Performance anchored score normalization for multi-biometric fusion," in *Advances in Visual Computing - 9th International Symposium, ISVC 2013, Rethymnon, Crete, Greece, July 29-31, 2013. Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 8034. Springer, 2013, pp. 68–75.

[4] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, 2006.

[5] N. Damer and A. Opel, "Multi-biometric score-level fusion and the integration of the neighbors distance ratio," in *Image Analysis and Recognition - 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 8815. Springer, 2014, pp. 85–93.

[6] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270 – 2285, 2005.

[7] N. Damer, A. Opel, and A. Nouak, "CMC curve properties and biometric source weighting in multi-biometric score-level fusion," in *17th International Conference on Information Fusion, FUSION 2014, Salamanca, Spain, July 7-10, 2014*. IEEE, 2014, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6916112/

[8] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383–393, 1974.

[9] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764 – 766, 2013.

[10] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, 1st ed., ser. Wiley Series in Probability and Statistics. Wiley, Jan. 1986.

[11] W. Kabir, M. O. Ahmad, and M. N. S. Swamy, "A new anchored normalization technique for score-level fusion in multimodal biometrie systems," in *IEEE International Symposium on Circuits and Systems, ISCAS 2016, Montréal, QC, Canada, May 22-25, 2016*. IEEE, 2016, pp. 93–96.

[12] ——, "A novel normalization technique for multimodal biometric systems," in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2015, pp. 1–4.

[13] N. Damer, A. Opel, and A. Nouak, "Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, Sept 2014, pp. 1382–1386.