Score Normalization in Multimodal Systems using Generalized Extreme Value Distribution

Renu Sharma[1, 2]
renu@cdac.in
Sukhendu Das[2]
sdas@iitm.ac.in
Padmaja Joshi[1]
padmaja@cdac.in

[1]Centre for Development of Advanced Computing, Mumbai, India
[2]Indian Institute of Technology, Madras, India

Abstract

In multimodal biometric systems, human identification is performed by fusing information in different ways like sensor-level, feature-level, score-level, rank-level and decision-level. Score-level fusion is preferred over other levels of fusion because of its low complexity and sufficient availability of information for fusion. However, the scores obtained from different unimodal systems are heterogeneous in nature and hence they all require normalization before fusion. In this paper, we propose a client-centric score normalization technique based on extreme value theory (EVT), exploiting the properties of Generalized Extreme Value (GEV) distribution. The novelty lies in the application of extreme value theory over the tail of the complete score distribution (genuine and impostor scores), assuming that the genuine scores form extreme values (tail) with respect to the entire set of scores. Normalization is then performed by estimating the cumulative density function of GEV distribution, using the parameter set obtained from genuine data. For evaluation, the proposed method is compared with state-of-the-art methods on two publicly available multimodal databases: i) NIST BSSR1 [22] multimodal biometric score database and ii) Database created from Face Recognition Grand Challenge V2.0 [23] and LG4000 iris images [24], to show the efficiency of the proposed method.

1 Introduction

Person identification in a multimodal biometric system is done by fusing cues from different biometric traits. Fusion of information is done roughly at five levels: sensor, feature, score, rank, and decision level. Incompatibility and high complexity issues prevail at the sensor and feature level fusion methods. Scarcity of information makes rank level and decision level fusion infeasible to implement. At the score-level, sufficient information is available for fusion and fusion can be done without increasing the complexity of the system. But the major challenge in information fusion at the score-level is the heterogeneous nature of matching scores obtained from the individual classifiers. The disparate nature of the underlying distribution of scores and failure of the constituent classifiers further complicates the fusion of information at the score-level. To overcome these shortcomings, normalization is necessary to transform the matching/identification scores obtained from the ensemble of classifiers to a common domain before fusion.

Pages 131.1-131.12
DOI: https://dx.doi.org/10.5244/C.29.131

Score normalization is a process of altering location, scale and shape parameters of score distributions obtained from the individual classifiers, so that matching scores of different classifiers fall within a common domain. We have proposed a client-centric score normalization technique using Extreme Value Theory (EVT), where the parameter set (scale, location and shape) is determined by modelling the Generalized Extreme Value (GEV) distribution over the genuine data. We have exploited the fact that all genuine scores are extreme values and can be used to model the extreme value distribution. This has been ignored by the previous EVT-based score normalization techniques [7, 8]. Score transformation is performed by estimating the cumulative density function (CDF) for the given (test) scores using the GEV distribution parameter set. Experimental results are shown on two publicly available multimodal databases: i) NIST BSSR1 [22] multimodal biometric score database and ii) Database created from Face Recognition Grand Challenge Ver2.0 (FRGC v2.0) [23] and LG 4000 iris images [24]. A brief description of the problem and the need of score normalization follow.

Consider that, N identities be stored in a unimodal biometric system, labelled as X, and x is a template (biometric sample). Template x is compared with all N identities, which produces one genuine score and N-1 impostor scores (when one template/identity per subject, is stored in the system) using the scoring function $s_X = \delta_X(x)$, where s_X is the matching score obtained a from unimodal system X. Figure 1(a) shows the genuine and impostor score distributions (synthetic) for a biometric system X. If Δ_x is the local threshold for the system, then the total error of the system [9] is

$$ER_X = \int_{-\infty}^{\Delta_x} p\left(\frac{s_X}{I}\right) p(I)ds + \int_{\Delta_x}^{\infty} p\left(\frac{s_X}{G}\right) p(G)ds \quad (1)$$

where, $p(s_X/I)$, $p(s_X/G)$ are the class-conditional probabilities and $p(I)$, $p(G)$ are the prior probabilities of impostor and genuine classes respectively. The shaded area in Figure 1(a) represents the total error in the biometric system X when Δ_x is taken as a threshold.

Similarly, for another unimodal biometric system labelled Y, total error of the system is

$$ER_Y = \int_{-\infty}^{\Delta_Y} p\left(\frac{s_Y}{I}\right) p(I)ds + \int_{\Delta_Y}^{\infty} p\left(\frac{s_Y}{G}\right) p(G)ds. \quad (2)$$

where, s_Y is the matching score obtained from another unimodal system Y, with $p(s_Y/I)$ and $p(s_Y/G)$ being the class-conditional probabilities. The probability distribution of genuine and impostor scores for unimodal system Y is shown in Figure 1(b) where the shaded area represents the total error. Genuine and impostor distributions of both unimodal systems differ in location, scale and shape. Figure 1(c) exhibits the difference in their local thresholds when their distributions are not aligned. In such scenarios, identifying the global threshold for the final decision becomes a challenging task in a multimodal biometric system.

1.1 Related Work

Poh and Kittler [5] classified the solutions to the above problem as: model-specific thresholding and model-specific normalization. In model-specific thresholding, local thresholds of individual unimodal systems are used to make decisions in the corresponding

systems and the final decision is made by fusion of information from different unimodal systems at rank and decision levels. In model-specific normalization, genuine and impostor score distributions are transformed in such a way that genuine-genuine and impostor-impostor distributions become well aligned and a single global threshold could be determined. Figure 1(d) shows an illustration where genuine and impostor distributions become well aligned after normalization and determining the global threshold becomes an easy task. Model-specific normalization can be of four categories: i) Impostor-centric, ii) Client-centric, iii) Impostor-client centric and iv) neither impostor nor client centric. In case of impostor-centric techniques, statistical information is obtained from the impostor scores and normalization is performed using the same. Z-Norm [1], T-Norm [1] are examples of the impostor-centric techniques. In the same way, client-centric techniques use genuine scores for predicting statistical information, necessary for normalization and the methods proposed in [2, 15, 20] fall into this category. F-Norm [3], EER-Norm [4], MS-LLR, [5] methods fall under the impostor-client centric category of techniques, which utilize information from both the distributions. Poh et *al.* [18] proposed a group-specific score normalization in which users are first categorize into groups and then F-Norm is applied over each group.

Jain et *al.* [6] describe various score normalization techniques for multimodal systems, categorizing them into two broad categories: fixed and adaptive score normalization. In fixed score normalization, the training set is used for fitting a distribution model and their parameters (scale and location) are used for normalization. In adaptive score moralization, parameters are estimated based on test score vectors obtain from the unimodal systems. This estimation has an ability to adapt to the variations in the input data, but data is quite limited for parameter estimation. A few more adaptive score normalization techniques are proposed in [19]. Shi et *al.* [7] modelled unimodal biometric systems using an Extreme Value Theory distribution, termed the Generalized Pareto Distribution (GPD). They have used a non-parametric method for modelling the significant part of the genuine distribution and a parametric GPD for modelling the tail part of the genuine distribution. Later on, Scheirer et *al.* [8, 21] also proposed an EVT-based adaptive score normalization method (W-Score) using the Weibull distribution. The methods proposed in [7, 8] focused on modelling the tail of the impostor (or genuine [7]) distribution. We assume that genuine scores form the tail of a complete (genuine and imposter combined) distribution, and hence we analyze the tail of the complete distribution by considering only the genuine scores. Moutafis et *al.* [17] further improved the W-Score by applying a rank-based scheme on W-Scores. Struc et *al.* [9] proposed an impostor-centric composite normalization, which is a two step process: first step is performed offline, where non-parametric rank transform is used, and the second step is performed parametrically using a log-normal distribution. Poh and Tistarelli [10] proposed a discriminative version of Z-norm (dZ-norm), F-norm (dF-norm) and parametric norm (dp-norm), by computing a weighted sum of the constituent linear terms of the Z-norm, F-norm and parametric norm. Recently, cohort-based score normalizations have been proposed by Merati et *al.* [11] and Tistarelli et *al.* [12]. Since the work presented in our paper deals only with score normalization techniques, methods of fusion [20, 25, 26] of scores are not discussed.

The rest of the paper is organized as follows. Section 2 gives the algorithmic details. Experimental results and analysis are presented in Section 3. Finally, the paper is concludes in Section 4.

2 Algorithmic Description

Our proposed method is based on modelling the extreme value distribution over the genuine scores, assuming that they form the tail of the complete score distribution (impostor and genuine). Then the cumulative density function of the model is used to transform the scores. First part of this section gives a brief overview of the Extreme Value Theory. The second part describes the proposed method in detail.

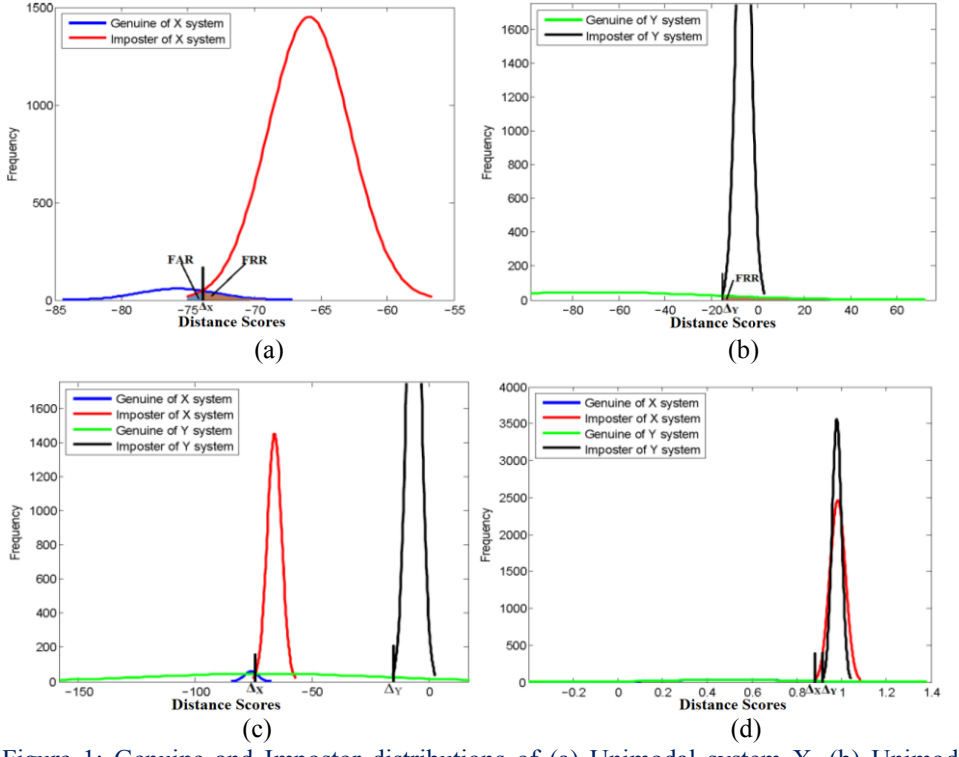


Figure 1: Genuine and Impostor distributions of (a) Unimodal system X, (b) Unimodal system Y, (c) X and Y unimodal systems before normalization and (d) X and Y unimodal systems after normalization.

2.1 Extreme Value Theory

Extreme Value Theory (EVT) [16] focuses on the statistical modelling of extreme and rare values of probability distributions. According to the EVT, there are two basic approaches to analyse or characterize the extreme values: i) Block Maxima and ii) Peak over Threshold. Block maxima is a parametric approach of modelling the maxima/minima taken from large blocks of independently and identically distributed (i.i.d.) random variables. The number and size of the blocks create a trade-off between low variance and bias in the parameter estimates. A large number of blocks lead to the estimate of a low variance, and a large block size leads to low bias in the estimation. Modelling of a sequence of maxima/minima by the parametric distribution is governed by Fisher–Tippett–Gnedenko theorem (Extreme value theorem) [16]. The theorem describes the limiting behaviour of

sequence of extreme values. Let X_1, X_2 ... be a sequence of independent and identically distributed (i.i.d.) random variables and $M_n = \max\{X_1, \ldots, X_n\}$ be the maximum of first n observations. If there exists a sequence of real numbers $a_n > 0$, $b_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) \tag{3}$$

for some non-degenerated distribution function $G(x)$, then the distribution function $G(x)$ belongs to one of three extreme value distributions. Three extreme distributions are Gumbel, Frechet and Weibull, which can together be represented by the Generalized Extreme Value (GEV) distribution [16]. GEV is the only possible limiting distribution for explaining the behaviour of the maxima/minima sequence. The cumulative density function (CDF) of GEV is given as

$$G(x, \mu, \sigma, k) = \begin{cases} exp\left(-\left(1 + k\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/k}\right) & if\ k \neq 0 \\ exp\left(-exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right) & if\ k = 0 \end{cases} \tag{4}$$

where, $1 + k\,(x - \mu)/\sigma > 0$ is such that $1 + kx > 0$. The three parameters μ , σ and k correspond to the mean, standard deviation and shape parameters of the distribution respectively. The value of $k = 0$ corresponds to the Gumbel (Type I), $k > 0$ for Frechet (Type II) and $k < 0$ for the Weibull (Type III) distributions. Weibull is a short-tailed distribution having an upper bound of $(\mu - \sigma/k)$. Frechet has a lower bound $(\mu - \sigma/k)$ and the tail falls off polynomially, whereas Gumbel is an unbounded distribution and the tail decreases exponentially. GEV distribution acquires any form depending upon the underlying distribution without taking any presumption about the boundedness of the distribution.

The peak over threshold approach is used to model large observations which exceed a high threshold. The observations which overshoot are modelled by the Generalized Pareto distribution or Poisson distribution. The method proposed in [7] is based on this approach using Generalized Pareto Distribution (GPD).

2.2 Proposed Score Normalization Technique

Architecture of the proposed method of score normalization is given in Figure 2(a). Input to the system of score normalization comprises of a distance matrix of scores (both genuine and impostor) from a biometric cue, using which the GEV distribution parameters are learned. These input score distance vectors will be termed as probe score vectors. Once the parameters of the distribution are learned, test score vectors (queries) are normalized using a CDF of the distribution formed using the learned parameters. Our proposed method and W-Score [8] technique are both based on the block maximum approach for extreme value analysis. W-Score method is an adaptive impostor-centric technique, which uses a single score vector obtained by comparing the input test (query) template (during query or testing) to the enrolled templates. From the score vector, the W-Score method uses a few top impostor scores excluding the topmost score (assuming that it is a genuine score) to fit a Weibull distribution. In contrast to the W-Score method, our algorithm falls under the category of client-centric [2], and hence for modelling the GEV distribution the number of

probe samples and their corresponding score vectors are utilized. This process occurs offline during training, which produces one genuine score (one template/identity is used) and N-1 impostor scores. A single genuine score is considered as an extreme value in the score vector. A collection of genuine scores forms a set of extreme values with respect to the probe score vectors. If a single probe score vector is considered as a block, genuine scores form a sequence of minimum (or maximum) values. According to the EVT theory, minima (or maxima) of sequences is characterized by the GEV distribution. So, the genuine data from the probe score vectors are modelled by the GEV distribution and the parameter set (mean, scale and location) is computed by the maximum likelihood estimation method. If $S_1, \dots, S_M$ are M genuine scores, then the log-likelihood function to be maximized is, formulated as:

$$LL(\mu, \sigma, k) = -M \log \sigma - \sum_j \left[1 + \frac{k(S_j - \mu)}{\sigma}\right]^{-1/k} - \left(\frac{1}{k} + 1\right) \sum_j \log \left[1 + \frac{k(S_j - \mu)}{\sigma}\right]. \quad (5)$$

As GEV is a parametric distribution and requires the estimation of only three parameters, few genuine values are sufficient to model the GEV distribution, as opposed to the non-parametric [7] [9] techniques which require a large number of genuine scores to estimate the distribution reliably. Hence, the proposed parametric technique has a lower computational complexity than the non-parametric techniques during training (fitting of model parameters) of the normalization process.

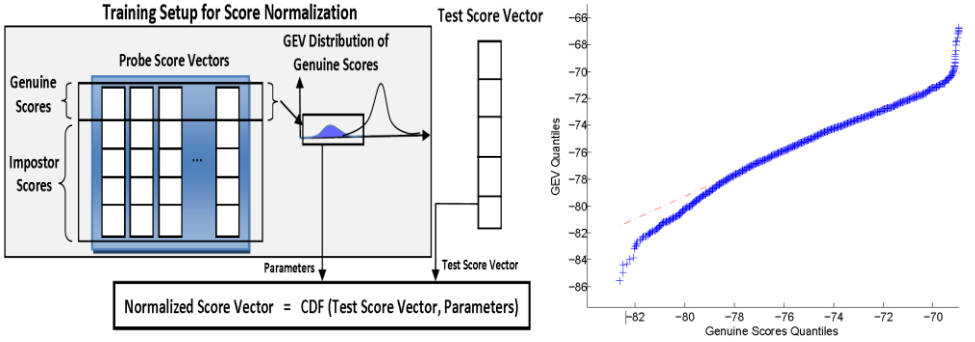
Given a GEV distribution, estimating the probability that a given score is an outlier is computed from the value of CDF of the GEV distribution. So, the normalized data are computed by CDF of the GEV distribution (Equation (4)), using the parameters estimated before, as follows:

$$S'_i = G(S_i, \mu, \sigma, k) \quad (6)$$

where, S'_i is the i^{th} class normalized score. After normalization, scores from the different unimodal systems fall within a common domain as shown in Figure 1(d). These normalized scores are fused by using any score-level fusion technique. Now the estimation of optimal global threshold becomes trivial in case of verification. In identification mode, the user is identified if the enrolled subject corresponds to the top score from the fused score vector. However, in case of a poor genuine score, it becomes an outlier and may appear as an impostor. To verify the correctness of fitting the GEV distribution over the genuine data, we observe a Quantile-Quantile plot of the modelled GEV distribution, as shown in Figure 2(b). The relationship is almost linear, which represents a good fitting of the data.

3 Experimental Results

For evaluation of the proposed method, we have performed experiments using two multi-modal biometric databases i) NIST BSSR1 [22] ii) Database created from Face Recognition Grand Challenge Ver2.0 (FRGC v2.0) [23] and LG 4000 iris images [24]. Experiments for score normalization are done in verification as well as identification modes. All experiments are repeated five times and the average performance is shown in all results. Experiments are performed on a core-*i*5, 2.3GHz, 8GB RAM machine.



(a) (b)

Figure 2: (a) Architecture of the proposed method for score normalization. (b) Almost a linear Quantile-Quantile plot of the genuine scores against GEV distribution, illustrating a proper fit.

The NIST BSSR1 [22] database consists of scores for two face algorithms (labelled as C and G) of 3000 subjects and two fingerprints (labelled as Li and Ri) scores of 6000 subjects. It also consists of all four (face C, Face G, Fingerprint Li and Fingerprint Ri) scores of 517 subjects. FRGC v2.0 [23] consist of 50,000 recordings comprises of high resolution images, 3D images and multiple images of a person (multiple RGB images are used for experiments). LG 4000 [24] iris dataset is collected for Cross Sensor Iris Recognition Challenge associated with BTAS 2013 dataset. It consists of 27 sessions of data for 676 unique subjects. Four test sets build from these databases are as follows:

Test Set I: Scores of face (Face C, Face G) biometrics for 3000 subjects from the NIST BSSR1 dataset are taken. The database provides scores for two images per subject, and thus has a total of 6000 images. For parameter estimation, 500 genuine scores are used. Results are shown for 5500 genuine and 1,64,94,500 (5500×2999) impostor scores.

Test Set II: Scores of face (Face C, Face G) and fingerprint (Fingerprint Li and Fingerprint Ri) biometrics for 517 common subjects are taken from the NIST BSSR1 dataset. Genuine scores of 200 subjects are used for parameter estimation. Results are shown for the rest: 317 genuine and 1,63,572 (317×516) impostor scores.

Test Set III: A chimeric dataset has been created using face (Face C, Face G) and fingerprint (Fingerprint Li and Fingerprint Ri) scores of 3000 subjects from the NIST BSSR1 dataset. Fingerprint scores of 3000 subjects are randomly selected from the given set of 6000 subjects. For parameter estimation, 500 genuine scores are used. Results are shown for the rest: 2500 genuine and 74,97,500 (2500×2999) impostor scores.

Test Set IV: Another chimeric dataset of 19,079 face images and 19,079 left iris images for 461 subjects, has been created from FRGC v2.0 and LG 4000 datasets. The whole dataset is divided into three sets: Training Set, Evaluation Set and Test Set. Five (5) images of face (and left iris) per subject are selected for the Training Set, also five (5) images of face (and left iris) per subject constitute the Evaluation Set, and the remaining (28,938 = 19,079×2 - 461×10×2) images of face and iris are included in the Test Set. Images are selected at random and do not overlap across sets. All face images were cropped to 131×111 pixels and iris images are resized to 640×480. Gabor filters (8 orientations and 5 scales) are used for feature extraction from face images and a 1-D Log-Gabor feature extraction technique is used for the iris images. For scores generation, SVM [14] and Probabilistic Neural Network [14] classifiers are used for the training the face and iris modalities respectively. Post-training, images from the Evaluation Set are used for parameter estimation of the GEV distribution. Evaluation set used to model the GEV

distribution comprises of 2305 (461×5) genuine scores. From the Test Set, 14,469 (28,938/2) genuine scores and 66,55,740 (460×14,469) impostor scores are used for evaluation of the proposed method.

To observe the impact of only the proposed method of score normalization, on performance of multimodal biometry systems, a simple sum rule is used for fusion. Performance of our proposed method (Figure 2(a)) is compared with the following methods: Without SN (without score normalization), Z-Score [6], Tanh [6], GPD [7], W-Score [8], Non-Parametric [9] and Rank-Based [17]. Verification rate (VR) at an equal error rate (EER) and rank one identification rate (IR) for different methods, are shown in Tables 1-4 on Test Sets I-IV respectively. For verification in W-Score [8] method, all impostor scores of training samples are used to represent the impostor distribution and tail size of 50 has been used, whereas for identification a tail size of 5 is used as given in [8]. Receiver operating characteristic (ROC) and cumulative match characteristic (CMC) curves of different techniques are shown in Figures 3-6(a, b), for the corresponding Test Sets I- IV. Time taken (log-scale) by different normalization techniques is shown in Figure 7. Tanh [6] method outperforms all other methods only in verification mode (see Table 1-4), but drastically fails in identification mode. Tanh [6] emphasizes over the central values of the distribution and reduces the influence of outliers which require additional attention in the identification mode. Our proposed method performs best in identification mode (see Table 1-4) as it emphasizes the extreme values (outliers), and a very close second best to that of Tanh [6] method in verification mode, without any increase the time complexity.

We have also experimented using support vector machine (SVM) with radial basis function (RBF) kernel [20] as a discriminative score-level fusion approach (sum rule was used for fusion in our earlier experiments) only in verification mode. Results of fusion using RBF-SVM are shown in the last two rows of Table 4, where Z-Score+SVM represents Z-Score [6] as a score normalization technique and RBF-SVM as score-level fusion technique. Similarly, Tanh+SVM represents Tanh [6] as a normalization technique and RBF-SVM as fusion technique. Our proposed method of GEV based normalization with RBF-SVM as fusion is termed as 'Our Method+SVM'. Performances (only in verification mode) of all of these four methods based on discriminative score-level fusion using RBF-SVM [20], is inferior to 'Our proposed method' or Tanh [6] method.

Methods	VR (at EER)	IR	Methods	VR (at EER)	IR
Face G	93.72 (0.0628)	77.50	GPD [7]	94.71 (0.0528)	81.11
Face C	94.70 (0.0529)	81.01	W-Score [8]	96.75 (0.0324)	81.75
Without SN	93.81 (0.0619)	78.50	Non-Parametric [9]	95.72 (0.0428)	84.5
Z-Score [6]	95.78 (0.0422)	83.16	Rank-Based [17]	95.68 (0.0431)	84.53
Tanh [6]	**99.99** (0.00002)	62.96	Our Method	96.89 (0.0311)	**85.11**

Table 1: VR (Verification Rate) with the corresponding value of EER (Equal Error Rate) specified within parenthesis, and IR (Identification Rate) obtained for different methods using Test Set I.

Methods	VR (at EER)	IR	Methods	VR (at EER)	IR
Face C	95.59 (0.0440)	89.16	Tanh [6]	**99.99** (0.000003)	87.81
Face G	94.19 (0.0580)	84.33	GPD [7]	95.74 (0.0425)	89.55
Fingerprint Li	91.58 (0.0830)	86.46	W-Score [8]	99.80 (0.0019)	98.06
Fingerprint Ri	95.11 (0.0489)	92.64	Non-Parametric [9]	99.43 (0.0056)	99.80
Without SN	98.25 (0.0174)	98.06	Rank-Based [17]	99.94 (0.0005)	99.61
Z-Score [6]	99.99 (0.0029)	100	Our Method	99.99 (0.00001)	**100**

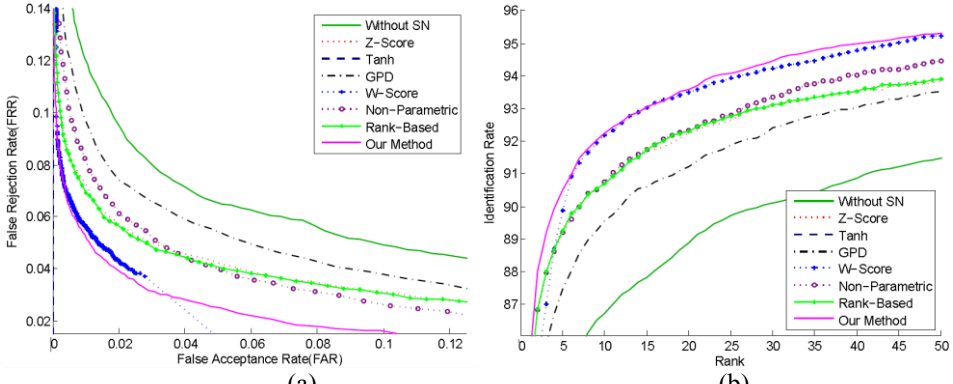
Table 2: VR (corresponding EER) and IR for different methods on Test Set II.

Methods	VR (at EER)	IR	Methods	VR (at EER)	IR
Face C	94.87 (0.0512)	82.33	Tanh [6]	**99.99** (0.000017)	85.03
Face G	93.91 (0.0609)	78.73	GPD [7]	94.86 (0.0514)	82.50
Fingerprint Li	92.22 (0.0778)	81.50	W-Score [8]	98.23 (0.0177)	87.03
Fingerprint Ri	94.75 (0.0525)	88.70	Non-Parametric [9]	99.27 (0.0073)	98.30
Without SN	98.63 (0.0136)	96.83	Rank-Based [17]	99.46 (0.0053)	98.46
Z-Score [6]	99.42 (0.0058)	99.03	Our Method	99.86 (0.0013)	**99.43**

Table 3: VR (corresponding EER) and IR for different methods on Test Set III.

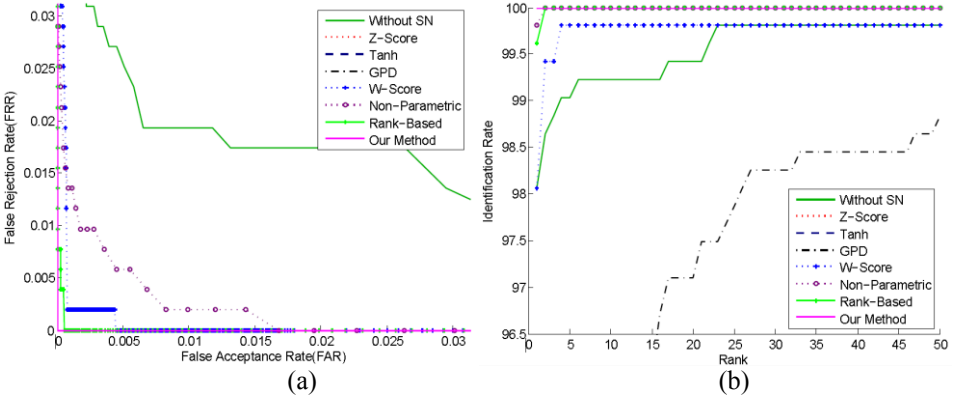
Methods	VR (at EER)	IR	Methods	VR (at EER)	IR
Face	93.53 (0.0647)	77.28	GPD [7]	98.91 (0.0109)	97.49
Iris	97.37 (0.0263)	92.92	W-Score [8]	99.35 (0.0065)	92.73
Without SN	98.80 (0.0120)	97.16	Non-Parametric [9]	98.64 (0.0136)	96.93
Z-Score [6]	98.31 (0.0169)	95.72	Rank-Based [17]	99.57 (0.0042)	93.63
Tanh [6]	**99.99** (0.0001)	56.71	Our Method	99.59 (0.0040)	**98.00**
Z-Score+SVM	99.20	-	RHE+SVM [20]	97.96	-
Tanh+SVM	66.49	-	Our Method+SVM	99.37	-

Table 4: VR (corresponding EER) and IR for different methods on Test Set IV.



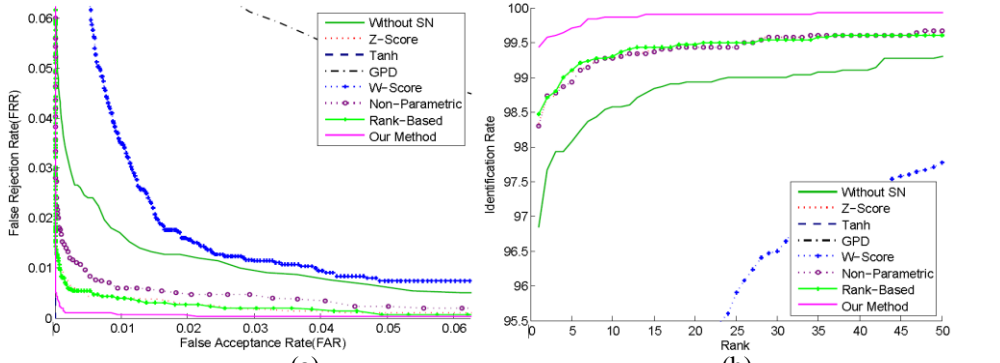
(a) (b)

Figure 3: (a) ROC and (b) CMC curves of different techniques on Test Set I.



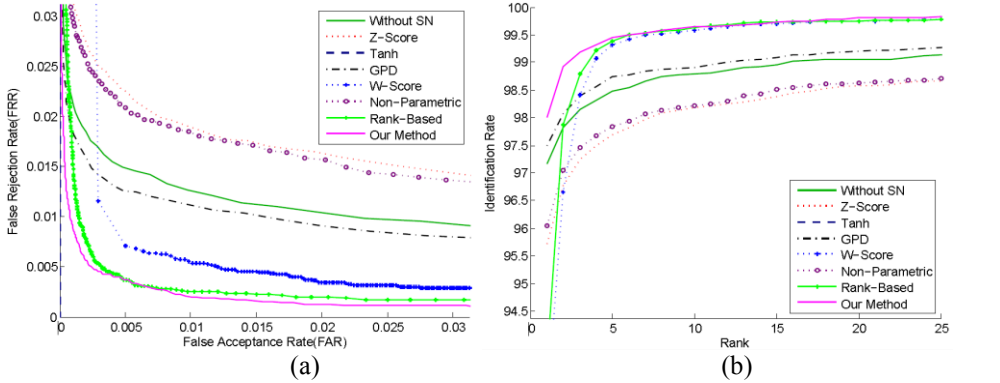
(a) (b)

Figure 4: (a) ROC and (b) CMC curves of different techniques on Test Set II.



(a) (b)

Figure 5: (a) ROC and (b) CMC curves of different techniques on Test Set III.



(a) (b)

Figure 6: (a) ROC and (b) CMC curves of different techniques on Test Set IV.

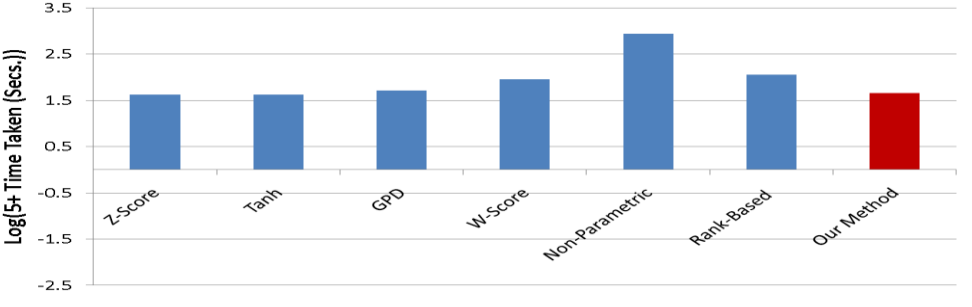


Figure 7: Time taken (in secs) by different normalization techniques, shown in log-scale.

4 Conclusions

We have analyzed the importance of score normalization in multimodal biometric systems when fusion is done at the score-level. We have proposed a normalization technique based on extreme value theory and evaluations are done on two multimodal databases. Extensive experiments on challenging multimodal databases with few thousands of subjects show the efficiency of our proposed normalizing method against the state-of-the-art in terms of performance. The proposed method outperforms all others in identification mode, and performs a very close second to the Tanh normalization technique in verification mode.

References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing (DSP) Journal*, 10:42–54, 2000.

[2] J. R. Saeta and J. Hernando. On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation. In *The Speaker and Language Recognition Workshop (Odyssey)*, 215–218, 2004.

[3] N. Poh and S. Bengio. F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 721–724, 2005.

[4] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target Dependent Score Normalisation Techniques and Their Application to Signature Verification. In *LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, 498–504, 2004.

[5] N. Poh and J. Kittler. On the use of log-likelihood ratio based model-specific score normalization in biometric authentication. In *Proc. IEEE Int. Conf. on Biometrics (ICB)*, 614–624, 2007.

[6] A. Jain, K. Nandakumar, A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270-2285, 2005.

[7] Z. Shi, F. Kiefer, J. Schneider, and V. Govindaraju. Modeling Biometric Systems Using the General Pareto Distribution (GPD). In *SPIE*, 6944, 2008.

[8] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. Robust Fusion: Extreme Value Theory for Recognition Score Normalization. In *ECCV*, 481-495, 2010.

[9] V. Štruc, J. Žganec-Gros, B. Vesnicer, N. Pavešic´. Beyond parametric score normalisation in biometric verification systems. *IET Biometrics,* 3(2):62-74, 2014.

[10] N. Poh and M. Tistarelli. Customizing Biometric Authentication Systems via Discriminative Score Calibration. In *CVPR,* 2681-2686, 2012.

[11] A. Merati, N. Poh, and J. Kittler. User-Specific Cohort Selection and Score Normalization for Biometric Systems. *IEEE Transactions on Information Forensics and Security,* 7(4), 2012.

[12] M. Tistarelli, Y. Sun, and N. Poh. On the Use of Discriminative Cohort Score Normalization for Unconstrained Face Recognition. *IEEE Transactions on Information Forensics and Security,* 9(12), 2014.

[13] N. Poh and S. Bengio. Why Do Multi-Stream, Multi-Band and Multi- Modal Approaches Work on Biometric User Authentication Tasks? *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 893–896, 2004.

[14] R. O. Duda, P. E. Hart and D. G. Stork. Pattern Classification (2nd Edition). *Wiley-Interscience,* 2000.

[15] N. Poh and J, Kittler. Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems. *IEEE. TASLP*, 16:594-606, 2008.

[16] S. Kotz and S. Nadarajah. Extreme Value Distributions: Theory and Applications. 1 Edn. *World Scientific Publishing Co*., 2001.

[17] P. Moutafis and I. A. Kakadiaris. Can We Do Better in Unimodal Biometric Systems? A Rank-Based Score Normalization Framework. *IEEE Transactions on Cybernetics* , 2014.

[18] N. Poh and A. Rattani and M. Tistarelli and J. Kittler. Group-specific Score Normalization for Biometric Systems. In *CVPR*, 38-45, 2010.

[19] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large-Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. *IEEE Transactions on Pattern Analyis and Machine Intelligence*, 27(3):450-455, 2005.

[20] M. He, S.-J. Horng, P. Fan, R.-S. Run, R.-J. Chen, J.-L. Lai, M. K. Khan and K. O. Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition,* 43(5):1789–1800, 2010.

[21] W. J. Scheirer, A. Rocha, R.J. Micheals, and T. E. Boult. Meta-Recognition: The Theory and Practice of Recognition Score Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1689-1695, 2011.

[22] NIST: Biometric Scores Set (2004) www.itl.nist.gov/iad/894.03/biometricscores.

[23] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Wore. Overview of the face recognition grand challenge. *Computer Vision and Pattern Recognition*, 1:947-954, 2005.

[24] LG 4000 Iris Image Database: http://biometrics.idealtest.org/

[25] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and Javier Ortega-Garcia. Dealing with sensor interoperability in multi-biometrics: the UPM experience at the Biosecure Multimodal Evaluation 2007. *Proc. of SPIE 6994, Biometric Technologies for Human Identification V*. 1-12, 2008.

[26] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury. A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. *IETE Technical Review*. 27(4):293-307, 2010.