Chapter 14
Exploiting Score Distributions for Biometric Applications

Panagiotis Moutafis and Ioannis A. Kakadiaris

Abstract Biometric systems compare biometric samples to produce matching scores. However, the corresponding distributions are often heterogeneous and as a result it is hard to specify a threshold that works well in all cases. Score normalization techniques exploit the score distributions to improve the recognition performance. The goals of this chapter are to (i) introduce the reader to the concept of score normalization and (ii) answer important questions such as why normalizing matching scores is an effective and efficient way of exploiting score distributions, and when such methods are expected to work. In particular, the first section highlights the importance of normalizing matching scores; offers intuitive examples to demonstrate how variations between different (i) biometric samples, (ii) modalities, and (iii) subjects degrade recognition performance; and answers the question of *why score normalization effectively utilizes score distributions*. The next three sections offer a review of score normalization methods developed to address each type of variation. The chapter concludes with a discussion of why such methods have not gained popularity in the research community and answers the question of *when and how one should use score normalization*.

14.1 Introduction

The goal of biometric systems is to determine whether or not (two or more) biometric samples have been acquired from the same subject. This problem is usually formulated as a verification or an open-set identification task. Regardless of the task or biometric trait used, one matching score is obtained for each pairwise compar-

P. Moutafis (✉) · I.A. Kakadiaris
Computational Biomedicine Lab, University of Houston, 4800 Calhoun Rd,
Houston, TX 77004, USA
e-mail: pmoutafis@uh.edu

I.A. Kakadiaris
e-mail: ioannisk@uh.edu

T. Bourlai (ed.), *Face Recognition Across the Imaging Spectrum*,
DOI 10.1007/978-3-319-28501-6_14

ison of biometric samples. This number reflects how similar the matched samples are. To reach a decision, the matching scores obtained are compared to a threshold. Ideally, the matching score distributions of the match and nonmatch scores would be separable. Hence, a single threshold would always yield a correct classification. In real-life applications, though, these distributions overlap greatly. To address this problem, many algorithms have been and continue to be developed with the goal of yielding more robust feature sets with better discriminative properties. For example, improved landmark detection and illumination normalization can significantly improve face recognition performance. However, such algorithms cannot always produce the desired results. Even worse, they cannot address inherent variations that increase the overlap of the match and nonmatch score distributions. In this section, we identify the sources of these variations and demonstrate how score normalization methods can effectively and efficiently improve recognition performance. The sources of variations reported in the literature can be grouped into three categories, as follows:

1. *Acquisition conditions*: Variations during data acquisition include differences in pose, illumination, and other conditions. For example, let us assume that we have a gallery of biometric samples. Let us further assume that all images in the gallery are frontal facial images captured under optimal illumination conditions. If a probe that is captured under similar conditions is submitted to the matching system, we can expect that the matching scores obtained will be high on average, even if the subject depicted is not part of the gallery. On the other hand, if another probe is captured under different conditions and then compared with the gallery, we can expect that the matching scores obtained will be low on average, even if the subject depicted is part of the gallery. In other words, the matching score distributions obtained for the two probes are heterogeneous. In this scenario, it would be difficult to correctly classify the two probes using the same threshold.
2. *Multimodal systems*: Unimodal systems are usually vulnerable to spoofing attacks [1] and prone to misclassifications for several reasons, such as lack of uniqueness and noisy data [9]. Multimodal biometric systems utilize information from multiple sources to address these challenges. Such sources may include different biometric traits (e.g., face, iris, and fingerprint) or different pipelines that utilize the same input data. However, fusing the information obtained from different modalities is not easy. The reason is that the matching score distributions produced by different modalities are heterogeneous, even if the gallery and probe subjects are the same. This effect complicates the fusion process. To provide visual evidence of this source of variation, we used pairwise matching scores for the face and iris traits (i.e., $2 \times 3{,}296{,}028$ distances) obtained from the CASIA-Iris-Distance database [37]. This dataset comprises 2,567 images obtained from 142 subjects, most of whom are graduate students at CASIA. The purpose of collecting these images was to promote research on long-range and large-scale iris recognition. Specifically, the images were acquired using a long-range multimodal biometric image acquisition and recognition system

developed by the CASIA group. The same samples were used to extract features for the face and iris traits, independently. The CASIA group provided us with the corresponding distances and the formula *score = max(distance) − distance* was used to convert them into scores. The corresponding boxplots are depicted in Fig. 14.1. As illustrated, the corresponding distributions are heterogeneous. In particular, the face matching scores have a higher median value than the iris matching scores. To assess the discriminative properties of the two modalities, we computed the corresponding receiver operating characteristic (ROC) curves. The area under the curve (AUC) obtained for the face matching scores is 93.48 %, while the AUC obtained for the iris matching scores is 94.17 %. Even though the two biometric traits yield comparable performance and the features were extracted using the same images, fusing the corresponding information is not straightforward.

3. *Subject variability*: It has been observed that, when assessing the performance of biometric systems in large populations, some subjects are easier to recognize than others. Similarly, some subjects can easily spoof the system. This phenomenon was first reported in the literature by Doddington et al. [5]. In that paper, the subjects were classified as sheep, goats, lambs, and wolves, depending on the statistical properties of the matching scores obtained for certain groups of subjects. This subject-specific variability of the matching scores is known as *biometric menagerie* and hinders the selection of a threshold that works well for all subjects.

Why do such methods work? As illustrated, biometric systems are vulnerable to inherent variations that increase the overlap of the match and nonmatch score distributions, thus degrading their recognition capability. Regardless of the source of variation, the challenge that needs to be addressed is the same: the matching

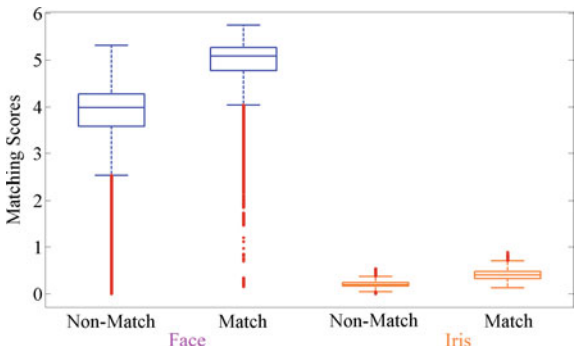

Fig. 14.1 Boxplots of the match and nonmatch scores obtained using two different modalities. The two boxplots on the left correspond to face matching scores, while the two boxplots on the right correspond to iris matching scores

score distributions are heterogeneous. Score normalization methods are techniques that map the matching scores to a common domain where they are directly comparable. In other words, they transform heterogeneous distributions into homogeneous ones.

Usually, two or more sources of variations occur at the same time. For example, a multimodal biometric system employed for large-scale open-set identification would be subject to all three types of variations. In the subsequent sections, we focus on one type of variation at a time. Specifically, we constrain our interest to the task for which the effect of the source of variation is more pronounced and review score normalization methods that address it more effectively.

14.2 Acquisition Conditions

The open-set identification task (also known as watch-list) consists of two steps: (i) a probe is matched with the gallery samples, and (ii) a candidate list is returned with the gallery samples that appear to be the most similar to it. This task can thus be viewed as a hard verification problem (see Fortuna et al. [7] for a more detailed discussion). Consequently, the recognition performance of such systems is significantly affected by variations in the acquisition conditions for both the gallery samples and the probes. Specifically, each time a probe is compared with a given gallery, the matching scores obtained follow a different distribution. Score normalization techniques address this problem by transforming the corresponding matching score distributions to homogeneous ones. Hence, a global threshold can be determined that works well for all submitted probes. In the following, we review some of the most popular methods tailored for this task.

Z-score: Due to its simplicity and good performance, this is one of the most widely used and well-studied techniques. In particular, it is expected to perform well when the location and scale parameters of the score distribution can be approximated sufficiently by the mean and standard deviation estimates. When the matching scores follow a Gaussian distribution, this approach can retain the shape of the distribution. The most notable limitations of Z-score are as follows: (i) it cannot guarantee a common numerical range for the normalized scores and (ii) it is not robust because the mean and standard deviation estimates are sensitive to outliers.

Median and median absolute deviation (MAD): This method replaces the mean and standard deviation estimates in the *Z-score* formula with the median value and the median absolute deviation, respectively. Therefore, it addresses the problem of lack of robustness due to outliers. However, it is not optimal for scores that follow a Gaussian distribution.

W-score [36]: Scheirer et al. proposed a score normalization technique that models the tail of the nonmatch scores. The greatest advantage of this approach is

that it does not make any assumptions concerning the score distribution. Also, it appears to be robust and yields good performance. However, to employ W-score the user must specify the number of scores to be selected for fitting. While in most cases it is sufficient to select as few as five scores, selecting a small number of scores may yield discretized normalized scores. Consequently, it is not possible to assess the performance of the system in low false acceptance rates or false alarm rates. On the other hand, selecting too many scores may violate the assumptions required to invoke the extreme value theorem. Another limitation of W-score is that it cannot be applied to multisample galleries unless an integration rule is first employed. As a result, it is not possible to obtain normalized scores for each sample independently. As it will be demonstrated, a recently proposed framework addresses this problem and extends the use of W-score to multisample galleries.

Additional score normalization techniques (e.g., tanh-estimators and double sigmoid function) are reviewed in [8]. Finally, some score normalization methods have been proposed that incorporate quality measures [27, 28, 33]. However, they are tailored to the verification task and have not been evaluated for open-set identification. The aforementioned methods consider the matching scores obtained for a single probe as a single set. This strategy does not fully utilize the available information for galleries with multiple samples per subject. To address this problem, Moutafis and Kakadiaris [15, 16] introduced a framework that describes how to employ existing score normalization methods (and those to be invented) more effectively. First, we review the theory of stochastic dominance, which theoretically supports their framework.

Definition The notation $X \gtrsim_{\text{FSD}} Y$ denotes that X first-order *stochastically dominates Y,* that is

$$\Pr\{X > z\} \geq \Pr\{Y > z\}, \quad \forall z. \tag{14.1}$$

As implied by this definition, the corresponding distributions will be ordered. This is highlighted by the following lemma (its proof may be found in [42]).

Lemma *Let X and Y be any two random variables, then*

$$X \gtrsim_{\text{FSD}} Y \Rightarrow E[X] \geq E[Y]. \tag{14.2}$$

An illustrative example of first-order stochastic dominance is depicted in Fig. 14.1 of Wolfstetter et al. [42] where $\bar{F}(z) \gtrsim_{\text{FSD}} \bar{G}(Z)$. Note that the first-order stochastic dominance relationship implies all higher orders [6]. In addition, this relation is known to be transitive as implicitly illustrated by Birnbaum et al. [4]. Finally, the first-order stochastic dominance may also be viewed as the stochastic ordering of random variables.

Algorithm 1 Rank-Based Score Normalization

Input: $S^p = \bigcup_i \{S_i^p\}$, f

Step 1: Partition S^p into subsets

$C_r = \{\emptyset\}, \forall r$
for $r = 1 : max_i\{|S_i^p|\}$ **do**
  for all $i \varepsilon I$ **do**
    $C_r = C_r \bigcup S_{i,r}^p$
  end for
end for

Step 2: Normalize each subset C_r

$S^{p,N} = \{\emptyset\}$
for $r = 1 : max_i\{|S_i^p|\}$ **do**
  $S^{p,N} = S^{p,N} \bigcup f(C_r)$
end for

Output: $S^{p,N}$

Rank-Based Score Normalization (RBSN): For the case of systems with multi-sample galleries, Moutafis and Kakadiaris [15, 16] proposed a RBSN algorithm that partitions the matching scores into subsets and normalizes each subset independently. An overview of the proposed RBSN framework is provided in Algorithm 1. The notation used is the following:

S^p the set of matching scores obtained for a given probe p when compared with a given gallery,
S_i^p the set of matching scores that correspond to the gallery subject with *identity* = i, $S_i^p \subseteq S^p$,
$S_{i,r}^p$ the ranked-r score of S_i^p,
$S^{p,N}$ the set of normalized scores for a given probe p,
C_r the rank-r subset, $\cup_r C_r = S^p$,
$|d|$ the cardinality of a set d,
I the set of unique gallery identities, and
f a given score normalization technique

An illustrative example of how to apply the proposed approach is provided in Fig. 14.2. Let us assume that there are three subjects in the gallery, namely *X, Y,* and *Z*. Let us further assume that three biometric samples are available for *X* (denoted by X_1, X_2, and X_3), two samples are available for *Y* (denoted by Y_1 and Y_2), and three samples are available for *Z* (denoted by Z_1, Z_2, and Z_3). Finally, let us assume that a probe is submitted to the system (denoted by p_i) and matched with all the gallery samples. Existing approaches would consider the obtained matching scores as a single set and normalize them in a single step. In contrast, the first step of the RBSN framework is to rank the matching scores for each gallery subject independently. For instance, if the matching scores obtained for *X* are $S(X_1, p_i) = 0.7$, $S(X_2, p_i) = 0.8$, and $S(X_1, p_i) = 0.6$, the corresponding ranks are 2, 1, and 3, respectively. If for subject *Y* we obtain $S(Y_1, p_i) = 0.4$, $S(Y_2, p_i) = 0.3$, then the ranks

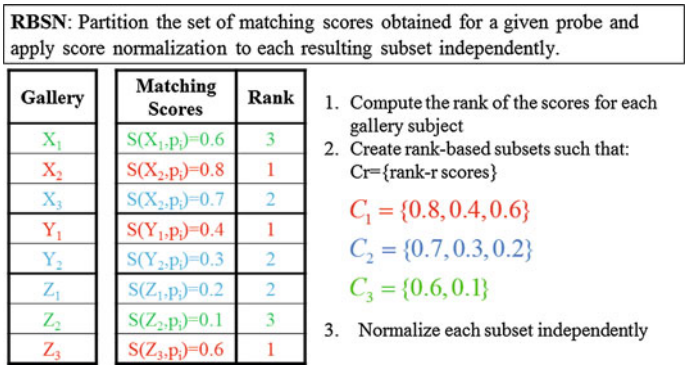

Fig. 14.2 Overview of the rank-based score normalization algorithm. The notation $S(X_1; p_i)$ is used to denote the score obtained by comparing a probe *pi* to the biometric sample 1 of a gallery subject labeled *X*

are 1 and 2, while if for subject *Z* we obtain $S(Y_1, p_i) = 0.2$, $S(Y_2, p_i) = 0.1$, $S(Y_1, p_i) = 0.7$, then the corresponding ranks are 2, 3 and 1 respectively. The second step of RBSN is to use the rank information to partition the matching scores into subsets. Specifically, the matching scores that ranked first comprise the subset $C_1 = \{0.8, 0.4, 0.7\}$, the ranked second scores comprise the subset $C_2 = \{0.7, 0.3, 0.2\}$, and the ranked third scores comprise the subset $C_3 = \{0.6, 0.1\}$. By invoking the theory of stochastic dominance, it is straightforward to demonstrate that the rank-based partitioning imposes the subsets' score distributions to be ordered (i.e., heterogeneous). To illustrate this point, each curve in Fig. 14.3 depicts the probability density estimate that corresponds to such subsets obtained from a gallery with six samples per subject. By normalizing the scores of each subset individually, the corresponding distributions become homogeneous and the system's performance improves. Hence, going back to our example, the matching scores of each set C_1,

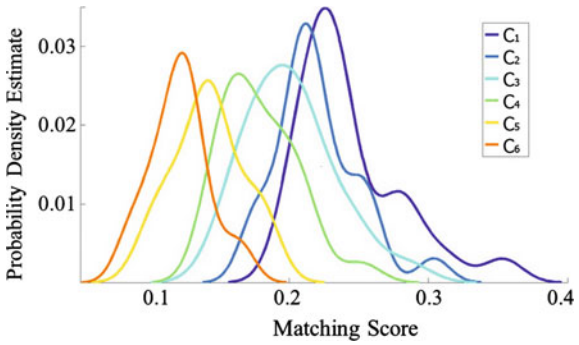

Fig. 14.3 Each curve depicts the probability density estimate corresponding to a C_r subset. Each subset C_r was constructed by *Step 1* of *RBSN* using the set S^p for a random probe *p*

C_2, and C_3 are normalized independently. Finally, the user might choose to fuse the normalized matching scores for each subject to consolidate the corresponding information. The RBNS framework (i) can be used in conjunction with any score normalization technique and any fusion rule, (ii) is amenable to parallel programming, and (iii) is suitable for both verification and open-set identifications. Two of the most important implications of this work are that (i) multiple samples per subject are exploited more effectively compared to existing methods, which yields improved recognition accuracy and (ii) improvements in terms of identification performance on a per-probe basis are obtained. We highlight selected results from [16] to illustrate these two points. First, the impact of the number of same-subject samples on the recognition performance was assessed. To this end, the UHDB11 dataset [38] was used which was designed to offer a great variability of facial data in terms of acquisition conditions. Specifically, 72 light/pose variations are available for 23 subjects, resulting in 2,742,336 pairwise comparisons. Six samples per subject were selected (one for each illumination condition) to form the gallery and the rest samples were used as probes. The matching scores used were provided by Toderici et al. [39]. Random subsets of one, three, and five samples per gallery subject were selected and each time the ROC curve and the corresponding AUC values were computed. This procedure was repeated 100 times using the unprocessed, raw matching scores, Z-score normalized scores, and RBSN:Z-score normalization scores. The obtained results are summarized in Fig. 14.4. As illustrated, RBSN:Z-score utilizes more effectively multiple samples per subject compared to Z-score. Second, the impact on the separation between the match and nonmatch scores on a per-probe basis was assessed. To this end, the FRGC v2 dataset was used that comprises 4,007 samples obtained from 466 subjects under different facial expressions. The 3D face recognition method of Ocegueda et al. [19] was used to extract the signatures and the Euclidean distance to compute the dissimilarity

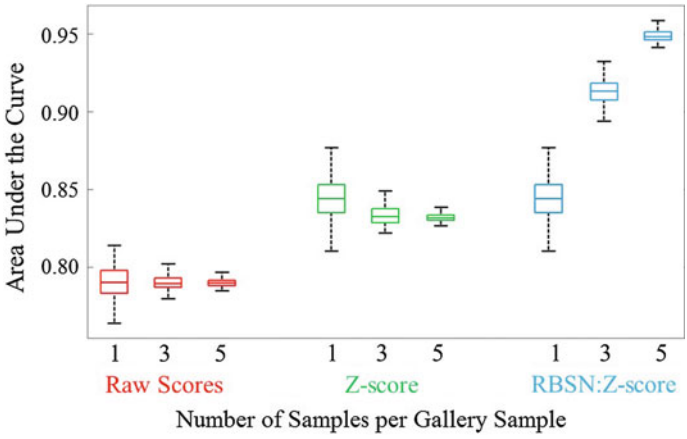

Fig. 14.4 Depicted are the boxplots for: (1) raw scores; (2) Z-score; and (3) RBSN:Z-score, when one, three, and five samples per gallery subject are randomly selected from UHDB11

values. The gallery was formed by randomly selecting 1,893 samples from 350 subjects. The rest were used as probes, resulting in an open-set problem. The Rank-1 errors for probes that belong to the gallery are as follows: (i) raw matching scores 0.74 %, (ii) Z-score normalized scores 0.74 %, and (iii) RBSN:Z-score normalized scores 0.66 %. Z-score and most existing approaches consist mostly of linear transformations, and therefore, they do not alter the order of the matching scores. Hence, the Rank-1 error for the raw matching scores and the normalized ones is the same. The RBSN algorithm, however, addresses this problem and has the potential to improve the accuracy of the rankings as illustrated.

To avoid confusion, we refer the readers to [15, 16] where they can find more implementation details, insights, experimental results, along with two versions of the RBSN algorithm that (i) fully utilizes the gallery versus gallery matching scores matrix and (ii) dynamically augments the gallery in an online fashion.

14.3 Multimodal Systems

Information fusion in the context of biometrics is a very challenging problem. Therefore, it has been receiving increasing attention over the past few years (Fig. 14.5). The most common approaches employ feature-level or score-level fusion. Methods in the first category (i) utilize the feature representation obtained for each modality to learn a common representation or (ii) learn rules that directly compare the multimodal representations to compute a matching score. Methods in the second category compute one matching score per pairwise comparison for each modality and then they either: (i) learn fusion rules that combine the information into a single matching score, or (ii) transform the scores to a standard form (i.e., score normalization) and then apply fixed fusion rules. In this section, we limit our scope to score-level fusion methods. The selected approaches were identified after conducting a systematic search of the literature that covered the years 2011–2014. To ensure that the latest papers have been included in our search, we focused on

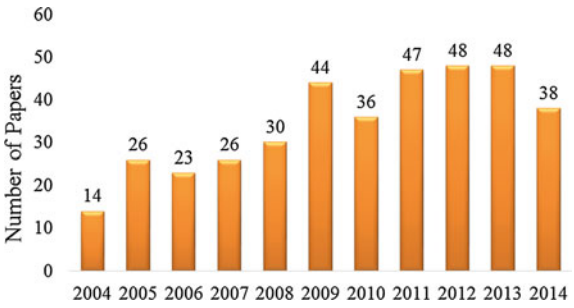

Fig. 14.5 Depiction of the number of papers published during the years 2004–2014 that include the words *biometric* and *fusion* in their title, according to the search engine Google Scholar

selected conferences in the field of biometrics and computer vision. The venues and keywords used are listed in Table 14.1. Papers that include at least one keyword in their title were reviewed in more detail to determine their relevance and interestingness. However, the number of papers selected was relatively small. To address this problem, we expanded the breadth of our search to the citations of the selected papers. We group the reviewed methods in three categories: (i) transformation-based, (ii) classification-based, and (iii) density-based. Methods in the first category normalize the matching scores and then employ fixed rules to combine them. Approaches in the second category usually treat the scores as features and learn a classifier that determines how similar the compared samples are. Finally, approaches in the third category estimate the probability density functions for each class. Such methods can be grouped as generative or discriminative. Generative methods focus explicitly on modeling the matching score distributions using parametric or nonparametric models. Discriminative approaches, on the other hand, focus explicitly on improving the recognition rate obtained by the fused scores. An overview of the categorization of score-level fusion methods is presented in Fig. 14.6, while an overview of the reviewed papers is offered in Table 14.2.

Transformation-based approaches normalize the matching score distributions of each modality independently. Consequently, the corresponding distributions become homogeneous and fixed fusion rules can be applied, which simplifies the fusion process. Kittler et al. [10] have studied the statistical background of fixed fusion rules. Two of the most popular ones are the *sum* and *max* operators. The former is implemented by a simple addition under the assumption of equal priors. Even though this rule makes restrictive assumptions, it appears to yield good performance as demonstrated in the literature [8, 10]. The latter makes less restrictive assumptions and it is also very simple to implement. Specifically, the output of this rule is defined to be the maximum score obtained. Wild et al. [41] employed a median filtering approach for score fusion to increase robustness to outliers. Specifically, this method disregards matching scores for which the distance from the median matching score exceeds a certain threshold. The authors employ

Table 14.1 Conferences used for identifying fusion methods. Papers that include at least one of the keywords in their title were considered in our review

Conferences
Conference on Computer Vision and Pattern Recognition (CVPR)
European Conference on Computer Vision (ECCV)
International Conference on Computer Vision (ICCV)
International Conference of the Biometrics Special Interest Group (BIOSIG)
International Conference on Biometrics: Theory, Applications and Systems (BTAS)
International Conference on Biometrics (ICB)
International Joint Conference on Biometrics (IJCB)
Keywords: Fusion, Information, Multimodal, Score

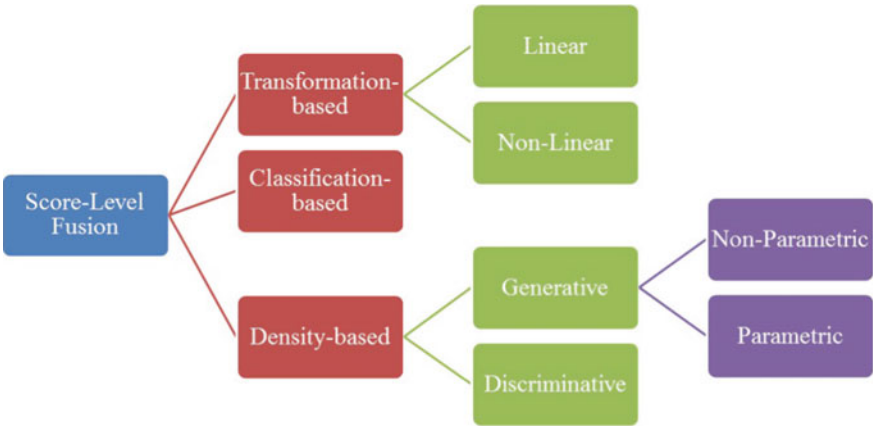

Fig. 14.6 Overview of the categorization of score-level fusion approaches

Table 14.2 Overview of score-level fusion papers. The column "Mapping" denotes whether the operations performed are linear or non-linear, while the column "Learning" denotes whether or not a method relies on offline training. The column "Model" denotes whether a method is Transformation-based, Classification-based, or Density-based (i.e., parametric or non-parametric)

Name	**Year**	Mapping	Learning	Model
Wild et al. [41]	2013	Linear	Adaptive	Transformation
Scheirer et al. [35]	2011	Nonlinear	Offline	Transformation
Nguyen et al. [18]	2014	Linear	Offline	Transformation
Mezai et al. [14]	2011	Linear	Offline	Transformation
Scheirer et al. [34]	2012	Nonlinear	Offline	Transformation
Zuo et al. [44]	2012	Linear	Adaptive	Transformation
Poh et al. [31]	2012	Linear	Offline	Transformation
Makihara et al. [12]	2014	Nonlinear	Offline	Nonparametric
Makihara et al. [13]	2011	Linear	Offline	Parametric
Poh et al. [30]	2011	Linear	Offline	Nonparametric
Liu et al. [11]	2014	Linear	Offline	Classification
Poh et al. [26]	2012	Nonlinear	Offline	Classification
Tyagi et al. [40]	2011	Linear	Offline	Classification

the proposed method to fuse matching scores obtained from fingerprints with liveliness values. These values denote the likelihood that the submitted sample is genuine and does not belong to an attacker seeking to spoof the system. As a result, this 1-median outlier detection approach alleviates negative effects to the recognition performance due to matching score anomalies, while it increases security. Scheirer et al. [35] proposed a statistical meta-recognition approach that relies on Weibull distribution. Specifically, the proposed approach models the tail of the

nonmatch scores obtained for a single probe and invokes the extreme value theorem to estimate the probability that the top-K matching scores contain an outlier (i.e., a match score). The decision-making process relies on the rejection rate of the null hypothesis, which states that a match score is contained in the top-K scores. Nguyen et al. [18] proposed a new approach based on the Dempster–Shafer theory. The basic belief assignment (BBA) function is represented as the hypothesis that the query and template belong (i) to the same class, (ii) to a different class, or (iii) that the relationship of the two cannot be defined. This model can naturally incorporate uncertainty measures into the model, which are related to the quality of the data and other factors. Mezai et al. [14] also proposed a Dempster–Shafer based algorithm. The fused scores are assigned into three categories: genuine, impostors, and unclassified. The authors argue that this approach reduces the half total error rate defined as the average of the false acceptance and false rejection rates. However, it does not consider that these metrics are affected by the unclassified data. Scheirer et al. [34] proposed a multiattribute calibration method for score fusion. Specifically, their approach fits a Weibull distribution to the flipped negative decision scores of an SVM classifier. Next, it normalizes the transformed scores using the cumulative density function. The multiattribute fusion is performed using the L_1 norm. That is, for a given query, the target samples that maximize the L_1 norm for each of the attributes are found. Unlike existing approaches that weigh all attributes equally, the proposed method finds the target samples that are most similar to most but not all the attributes. Zuo et al. [44] proposed a new approach for matching short-wave infrared (SWIR) and visible data. The images are first filtered and encoded using well-known filters. The encoded responses are then split into multiple nonoverlapping blocks and bin histograms are generated. The authors observed that the zero values obtained for SWIR and visible images are highly correlated. Hence, they proposed a score normalization method that addresses this problem. Specifically, the symmetric divergence between a visible template and a SWIR template is first computed. Then, a normalization factor is defined as the average difference of the computed divergence and the matching similarity scores obtained for an SWIR probe template. Finally, the normalized scores are computed as the divergence of a given visible image and a given SWIR template, minus the normalization factor and the symmetric divergence computed in the previous two steps. Poh et al. [31] proposed a client-specific score normalization approach. Specifically, the authors proposed three discriminative strategies: (i) dF-norm, (ii) dZ-norm, and (iii) dp-norm. These are defined as the probability of the subject being a client given the corresponding class mean and variance for the client and impostor. To address the problem of few client samples, the client-specific mean score is computed as a weighted average of the client and the global client mean scores. Moutafis and Kakadiaris [17] proposed a RBSN framework for multibiometric score fusion. Unlike existing approaches that normalize the matching scores from each modality independently, the multi rank-based score normalization (MRBSN) framework takes into consideration inherent correlations between the data. The first step is to normalize the matching scores of each modality

independently as usual. The second step, though, is to join the normalized scores to form a single set. Finally, the joined set of scores is processed using RBSN. The implementation is summarized in Algorithm 2. The additional notation is the following:

S^J the set of matching scores obtained for a given probe using the modality denoted by J,
$S^{J,N}$ the set of normalized scores for a given probe,
S the set of joined normalized score sets, $S = \cup_J S^{J,N}$,
S^{N_2} the set of "twice" normalized scores, and
R a given fusion rule.

An illustrative example of how to apply MRBSN is provided in Fig. 14.7. Let us assume that facial and iris data are available for three subjects, namely *X, Y,* and *Z*. The superscript *F* denotes that the biometric sample at hand was derived from face data, while the superscript *I* is used for the iris data. Let us further assume that a probe comprising face p_i^F and iris p_i^I data is submitted to the system. The matching score obtained for the face modality for subject *X* is denoted by $S(X^F, p_i^F)$, while the matching score obtained for the iris modality is denoted by $S(X^I, p_i^I)$. After normalizing the matching scores for the two modalities independently, we obtain the normalized scores. The normalized scores for the face and iris modalities for subject *X* are denoted by $S^N(X^F, p_i^F)$ and $S^N(X^I, p_i^I)$, respectively. The normalized scores

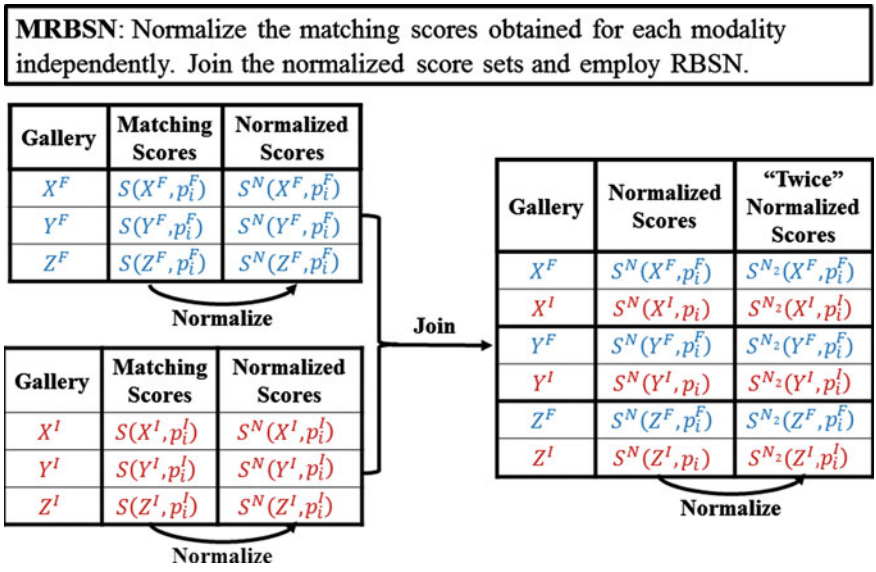

Fig. 14.7 Overview of the rank-based score normalization algorithm. The notation $S(X_1; p_i)$ is used to denote the score obtained by comparing a probe *pi* to the biometric sample 1 of a gallery subject labeled *X*

for the two modalities are then joined to form a single set. Since the scores are no longer distinguished based on the modality, the RBSN algorithm can be employed to leverage the multiple scores per subject. Experimental results using the CASIA dataset illustrate the benefits of this approach. Specifically, one sample from each modality was used for 71 subjects to define the gallery. The rest were used as probes. This process was repeated 50 times and the matching scores were normalized using Z-score, W-score, MRBSN:Z-score, and MRBSN:W-score. The mean values of the verification performance obtained at false acceptance rate equal to 10^{-2} are 90.90, 91.08, 85.46, and 86.29 %, respectively. For a more detailed analysis of the implementation and complete results, we refer the readers to [17].

Algorithm 2 Multi-Rank-Based Score Normalization

Input: S^J, Z, R
Step 1: Normalize each S^J *independently*
for all J **do**
$S^{J,N} = Z(S^J)$
end for
Step 2: Join $S^{J,N}$
$S = \bigcup_J S^{J,N}$
Step 3: Employ RBSN
$S^{N_2} = RBSN(S, Z)$
Step 4: Fuse the "twice" normalized scores
$S^{N_2} = R(S^{N_2})$
Return S^{N_2}

Density-based methods estimate the parameters of the probability density functions for each class by modeling a function of the likelihood ratio, or represent the distributions using histogram bins. Methods in the former category yield better results when the assumed model is correct. However, they fail when this assumption does not hold. On the other hand, methods in the second category can handle any type of distribution. Nevertheless, they do not scale well because the fitting process is computationally expensive. Makihara et al. [12] proposed a method that uses floating control points (FCP) for binary classification. A stratified sampling is employed multiple times to initialize a k-means clustering algorithm. Then, the generalized Delaunay triangulation is applied on the FCP (i.e., the cluster means) and the posterior distribution (PD) for the FCP is estimated. The PD is estimated by minimizing an energy function, which includes a smoothness constraint. Finally, the PD of the data is represented as an interpolation or extrapolation of the FCP PD based on the triangulation mesh. Makihara et al. [13] proposed another method that relies on the Bayes error gradient (BEG) distribution. The energy function for BEG distributions relies on the data fitness of a multilinear interpolation for each of the lattice-type control points. Furthermore, the authors incorporate prior knowledge into the model by strengthening the smoothness parameters and by adding monotonically increasing constraints upon the BEG

distribution. The experimental results indicate that the BEG with prior information is competitive with the sum-rule, even when the size of the client training samples decreases. Poh et al. [30] proposed a heterogeneous information fusion approach for biometric systems. Depending on the information sources used, the authors distinguish two cases: (i) independent and (ii) dependent score-level fusion. For the first case, the authors proposed a homogeneous fusion scheme (i.e., Naive Bayes), which is defined as the sum of the logit conditional probabilities of a genuine matching score given the source information. For the second case, the authors used the sum of the logit bind probabilities. The conducted experiments demonstrate that greater performance gains are obtained for the heterogeneous case.

Classification-based approaches do not model the distribution of the matching scores. Instead, they consider the matching scores as features and use them to train classifiers that discriminate each class. As a result, they provide a trade-off between accurate recognition and low time complexity. Liu et al. [11] demonstrated that the variance reduction equal error rate (VR-EER) model proposed by Poh and Bengio [22] is theoretically incomplete. To address this limitation, they proposed a new theoretical approach for score-level fusion. In particular, they demonstrated that under certain assumptions optimal fusion weights can be derived that maximize the F-ratio. Hence, the proposed approach can always perform at least as well as the best expert. Poh et al. [26] proposed a temporal fusion bimodal methodology for video and audio fusion. The audio is processed using Gaussian mixture model with maximum aposteriori adaptation (MAP-GMM). The video is processed in two ways. First, features are extracted from each face and each frame using a discrete cosine transform. Then, the MAP-GMM is applied to compute matching scores, which are fused using the mean rule. Second, nonuniform local binary pattern features are extracted followed by Fisher discriminant projection. The corresponding matching scores obtained are fused using the max rule. The first approach yields multiple scores, which are used to compute descriptive statistics. A logistic regression model is then learned that uses these descriptive statistics in conjunction with the scores obtained from the second approach. Finally, the sound and video modalities scores are merged using Naive Bayes. This pipeline allows temporal fusion, improves recognition performance, and increases robustness to spoof attacks. Tyagi et al. [40] proposed a new method to estimate the Gaussian mixture models using the maximum accept and reject criteria. The motivation behind this decision is that, by using the maximum accept and reject criteria instead of the likelihood, the optimization process focuses more on the classification itself rather than the fitting of a density model. As a result, increased recognition performance is achieved.

In summary, transformation-based methods are intuitive, simple, and efficient, but they do not utilize training data. Density-based approaches can be optimal if the assumptions made hold (i.e., parametric) or can fit the data relying on computationally expensive operations (i.e., nonparametric). Finally, classification-based approaches provide a trade-off between accurate recognition and efficiency. However, they require vast training data to ensure good generalization properties.

14.4 Subject Variability

Even when the acquisition conditions are controlled, there are still variations in the matching score distributions. Specifically, the matching scores obtained for different subjects exhibit different statistical properties. Several papers have studied this phenomenon and different groupings of the subjects have been proposed. The most popular are the Doddington's Zoo [5] and Yager and Dunstone's [43] classifications. There are different ways to classify subjects into different groups. For instance, some methods rely on criteria such as the F-ratio, the Fisher ration, and the d-prime metric [24], while other methods rely on the training matching scores dataset to rank and order the subjects [24]. Finally, a biometric menagerie index has been proposed by Poh and Kittler [25] to assess the severity of the biometric menagerie.

Two of the most common ways to address the problem of subject variability are (i) user-specific threshold and (ii) user-specific score normalization. In this section, we review relevant score normalization approaches that work well in a variety of datasets. Such methods can be grouped into two categories: (i) parametric and (ii) learning-based.

14.4.1 Parametric-Based Normalization

Parametric approaches make assumptions concerning the matching score distributions of each subject (or groups of subjects). That is, they model the corresponding distributions and then transform them into a standard form.

Z-norm: This method focuses on the nonmatch score distribution. Specifically, it assumes that the corresponding matching scores follow a Gaussian distribution. Hence, it estimates the corresponding mean and standard deviation values (e.g., using a training set) and uses them to standardize each score obtained for that subject. The distribution of the normalized nonmatch scores has a mean value equal to 0 and standard deviation equal to 1.

F-Norm: This approach extends the Z-norm method in the sense that it models both the match and nonmatch score distributions. It relies on the assumption that the corresponding distributions are Gaussian. Unlike Z-norm, though, it estimates the mean values for the match and nonmatch scores, which are then used to normalize the scores. However, the scarce availability of match scores can yield poor estimates for the mean. To address this problem, the corresponding value is estimated by interpolating the subject-specific match scores mean and the global match scores mean. The distribution of the normalized nonmatch scores has a mean value equal to 0, while the distribution of the normalized match scores has a mean value equal to 1. A more in-depth analysis is offered by Poh and Bengio [21].

The Test Normalization (*T-Norm)* [2] It is a variation of the Z-norm method. However, it is implemented in an online fashion. That is, the nonmatch mean and standard deviation estimates are computed at test time using an additional cohort of impostor samples.

Group-Based Normalization: Unlike existing approaches that normalize the matching scores on a per subject basis, Poh et al. [29] proposed a group-based normalization scheme. In particular, they cluster the subjects into groups and use the corresponding information to normalize the matching scores. As a result, the paucity of match scores is addressed.

14.4.2 Learning-Based Normalization

Learning-based methods employ statistical models with the goal of decreasing the overlap of the match and nonmatch score distributions.

Model-Specific Log Likelihood Ratio (MS-LLR): The proposed approach seeks a transformation that optimizes a likelihood ratio test that relies on the match and nonmatch score distributions [23]. The resulting score normalization method utilizes both match and nonmatch scores. Under the assumption that the standard deviation of the two populations is the same, the MS-LLR is equal to Z-norm, shifted by a constant value that is computed on a per subject basis.

Logistic Regression: One way of normalizing scores is to employ logistic regression. That is, a training set of match and nonmatch scores can be used to train a logistic regression model such that the output approximates the posterior probability of an input being a match score. Another way to utilize the logistic regression is to decompose the Z-norm or F-norm formulas to different terms. Then, the regression model is employed to learn optimal weights [32] for each of the terms.

14.5 Conclusion

Utilizing score distributions has the potential to significantly improve the recognition performance. However, methods such as score normalization must be used carefully and with discretion because inappropriate use may lead to severely degraded recognition performance. To determine whether it is suitable to exploit matching score distributions for a certain application, the first step should be to investigate whether or not there are inherent variations as described in Sect. 14.1. Depending on the results obtained from this analysis and the application at hand, the most appropriate score normalization method should be selected. For example, in the case of multimodal systems, score normalization methods tailored for fusion should be used. Nevertheless, regardless of the method selected, the validity of the corresponding assumptions should be checked. For example, before applying

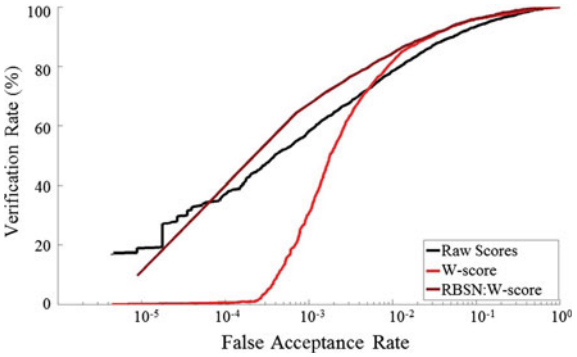

Fig. 14.8 ROC curves obtained for PaSC using the PittPatt face recognition system

Z-score, the user should ensure that the matching scores are approximately Gaussian distributed, and W-score is applicable only to single-sample galleries. To illustrate the importance of checking the necessary assumptions, we used the Point and Shoot Challenge (PaSC) dataset [3] and the face recognition system PittPatt [20]. The PaSC dataset was designed to assess the performance of biometric systems when inexpensive camera technologies are used to capture images from everyday life situations. Specifically, it includes 9,376 images from 293 subjects. For our experiment, we used 659 samples obtained from 117 subjects as a gallery and 2,739 samples from 122 subjects as probes. That is, images for five subjects are not included in the gallery, resulting in an open-set problem. The scores were normalized with W-score and RBSN:W-score using 30 scores to fit the tail of the distribution. Since there are multiple samples per gallery, the extreme value theorem requirements are violated for W-score but not for RBSN:W-score. The obtained ROC curves are depicted in Fig. 14.8. As illustrated, W-score results in degraded verification performance when compared with the verification performance obtained using raw scores. The RBSN:W-score, on the other hand, yields improvements.

As illustrated in this chapter, appropriate utilization of the matching score distributions can increase recognition performance of biometric systems in a reliable manner at a relatively low computational cost.

Acknowledgments The authors would like to thank Prof. Z. Sun and his students for sharing their data. Portions of the research in this paper use the CASIA-IrisV4 collected by the Chinese Academy of Sciences Institute of Automation (CASIA). This research was funded in part by the US Army Research Laboratory (W911NF-13-1-0127) and the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

1. Akhtar, Z., Fumera, G., Marcialis, G., Roli, F.: Evaluation of serial and parallel multibiometric systems under spoofing attacks. In: Proceedings of 5th International Conference on Biometrics: Theory, Applications and Systems, pp. 283–288, New Delhi, India, March 29–April1 2012
2. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digit. Signal Proc. **10**(1), 42–54 (2000)
3. Beveridge, J., Phillips, P., Bolme, D., Draper, B., Givens, G., Lui, Y., Teli, M., Zhang, H., Scruggs, W., Bowyer, K., Flynn, P., Cheng, S.: The challenge of face recognition from digital point-and-shoot cameras. In Proceedings of 6th International Conference on Biometrics: Theory, Applications and Systems, pp. 1–8, Washington DC, 4–7 June 2013
4. Birnbaum, M., Patton, J., Lott, M.: Evidence against rank-dependent utility theories: tests of cumulative independence, interval independence, stochastic dominance, and transitivity. Organ. Behav. Hum. Decis. Process. **77**(1), 44–83 (1999)
5. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In Proceedings of the International Conference on Spoken Language Processing, vol. 4, pp. 1–4, Sydney, Australia, Nov 30–Dec 4 1998
6. Durlauf, S., Blume, L.: The New Palgrave Dictionary of Economics. Palgrave Macmillan, Basingstoke (2008)
7. Fortuna, J., Sivakumaran, P., Ariyaeeinia, A., Malegaonkar, A.: Relative effectiveness of score normalisation methods in open-set speaker identification. In: Proceedings of the Speaker and Language Recognition Workshop, Toledo, Spain, May 31 June 3 2004
8. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **38**(12), 2270–2285 (2005)
9. Jain, A., Ross, A.: Multibiometric systems. Commun. ACM **47**(1), 34–40 (2004)
10. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
11. Liu, Y., Yang, L., Suen, C.: The effect of correlation and performances of base-experts on score fusion. Trans. Syst. Man Cybern. Syst. **44**(4), 510–517 (2014)
12. Makihara, Y., Muramatsu, D., Iwama, H., Ngo, T., Yagi, Y.: Score-level fusion by generalized Delaunay triangulation. In: Proceedings of 2nd International Joint Conference on Biometrics, pp. 1–8, Clearwater, FL, Sep 29 Oct 2 2014
13. Makihara, Y., Muramatsu, D., Yagi, Y., Hossain, A.: Score-level fusion based on the direct estimation of the bayes error gradient distribution. In: Proceedings of the International Joint Conference on Biometrics, pp. 1–8, Washington DC, 11–13 Oct 2011
14. Mezai, L., Hachouf, F., Bengherabi, M.: Score fusion of face and voice using DempsterShafer theory for person authentication. In: Proceedings of 11th International Conference on Intelligent Systems Design and Applications, pp. 894–899. Cordoba, Spain, 22–24 Nov 2011
15. Moutafis, P., Kakadiaris, I.: Can we do better in unimodal biometric systems? A novel rank-based score normalization framework for multi-sample galleries. In: Proceedings of 6th IARP International Conference on Biometrics, Madrid, Spain, 4–7 June 2013
16. Moutafis, P., Kakadiaris, I.: Can we do better in unimodal biometric systems? A novel rank-based score normalization framework. Trans. Cybern. **99**, 1–14 (2014, In Press)
17. Moutafis, P., Kakadiaris, I.: Rank-based score normalization for multi-biometric score fusion. In: Proceedings of 8th International Symposium on Technologies for Homeland Security, Waltham, MA, 14–15 April 2015
18. Nguyen, K., Denman, S., Sridharan, S., Fookes, C.: Score-level multibiometric fusion based on Dempster-Shafer theory incorporating uncertainty factors. Trans. Hum. Mach. Syst. **99**, 1–9 (2014)
19. Ocegueda, O., Passalis, G., Theoharis, T., Shah, S., Kakadiaris, I.: UR3D-C: linear dimensionality reduction for efficient 3D face recognition. In: Proceedings of the International Joint Conference on Biometrics, pp. 1–6, Washington DC, Oct 11–13 2011

20. Pittsburgh Pattern Recognition.: PittPatt face recognition software development kit (PittPatt SDK) v5.2, March 2011
21. Poh, N., Bengio, S.: An investigation of f-ratio client-dependent normalisation on biometric authentication tasks. Technical report, 04-46, IDIAP, Martigny, Switzerland (2004)
22. Poh, N., Bengio, S.: How do correlation and variance of base-experts affect fusion in biometric authentication tasks? Trans. Signal Process. **53**(11), 4384–4396 (2005)
23. Poh, N., Kittler, J.: Incorporating variation of model-specific score distribution in speaker verification systems. IEEE Trans. Audio Speech Lang. Process. **16**(3), 594–606 (2008)
24. Poh, N., Kittler, J.: A methodology for separating sheep from goats for controlled enrollment and multimodal fusion. In: Proceedings of 6th Biometrics Symposium, pp. 17–22, Tampa, FL, 23–25 Sept 2008
25. Poh, N., Kittler, J.: A biometric menagerie index for characterising template/model specific variation. In: Proceedings of 3rd International Conference on Biometrics, pp. 1–10, Sassari, Italy, 2–9 June 2009
26. Poh, N., Kittler, J., Alkoot, F.: A discriminative parametric approach to video-based score-level fusion for biometric authentication. In: Proceedings of 21st International Conference on Pattern Recognition, vol. 3, pp. 2335–2338 (2012)
27. Poh, N., Kittler, J., Bourlai, T.: Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization. In: Proceedings of 1st International Conference on Biometrics: Theory, Applications, and Systems, pp. 1–5, Washington DC, 27–29 Sept 2007
28. Poh, N., Kittler, J., Bourlai, T.: Quality-based score normalization with device qualitative information for multimodal biometric fusion. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(3), 539–554 (2010)
29. Poh, N., Kittler, J., Rattani, A., Tistarelli, M.: Group-specific score normalization for biometric systems. In: Proceedings of 23rd IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 38–45, San Francisco, CA, 13–18 June 2010
30. Poh, N., Merati, A., Kittler, J.: Heterogeneous information fusion: a novel fusion paradigm for biometric systems. In: Proceedings of the International Joint Conference on Biometrics, pp. 1–8, Washington, DC, 10–13 Oct 2011
31. Poh, N., Tistarelli, M.: Customizing biometric authentication systems via discriminative score calibration. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Providence, RI, June 16–21 2012
32. Poh, N., Tistarelli, M.: Customizing biometric authentication systems via discriminative score calibration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2681–2686, Providence, RI, 16–21 June 2012
33. Rua, E., Castro, J.L., Mateo, C.: Quality-based score normalization for audiovisual person authentication. In: Proceedings of the International Conference on Image Analysis and Recognition, pp. 1003–1012, Povoa de Varzim, Portugal, 25–27 June 2008
34. Scheirer, W., Kumar, N.: Multi-attribute spaces: calibration for attribute fusion and similarity search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2933–2940, Providence, RI, 16–21 June 2012
35. Scheirer, W., Rocha, A., Michaels, R., Boult, T.: Meta-recognition: the theory and practice of recognition score analysis. Trans. Pattern Anal. Mach. Intell. **33**, 1689–1695 (2011)
36. Scheirer, W., Rocha, A., Micheals, R., Boult, T.: Robust fusion: extreme value theory for recognition score normalization. In: Proceedings of the European Conference on Computer Vision, vol. 6313, pp. 481–495. Crete, Greece, 5–11 Sept 2010
37. Sun, Z., Tan, T.: CASIA iris image database, 14 Aug 2014 (2012)
38. Toderici, G., Evangelopoulos, G., Fang, T., Theoharis, T., Kakadiaris, I.: UHDB11 database for 3D-2D face recognition. In: Proceedings of the Pacific-Rim Symposium on Image and Video Technology, pp. 1–14, 28 Oct 2013

39. Toderici, G., Passalis, G., Zafeiriou, S., Tzimiropoulos, G., Petrou, M., Theoharis, T., Kakadiaris, I.: Bidirectional relighting for 3D-aided 2D face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2721–2728, San Francisco, CA, 13–18 June 2010
40. Tyagi, V., Ratha, N.: Biometric score fusion through discriminative training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 145–149, Colorado Springs, CO, 20–25 June 2011
41. Wild, P., Radu, P., Chen, L., Ferryman, J.: Towards anomaly detection for increased security in multibiometric systems: spoofing-resistant 1-median fusion eliminating outliers. In: Proceedings of the 2nd International Joint Conference on Biometrics, pp. 1–6, Clearwater, FL, Sept 29 Oct 2 2014
42. Wolfstetter, E., Dulleck, U., Inderst, R., Kuhbier, P., Lands-Berger, M.: Stochastic dominance: theory and applications. Humboldt University of Berlin, School of Business and Economics, Berlin (1993)
43. Yager, N., Dunstone, T.: The biometric menagerie. IEEE Trans. Pattern Anal. Mach. Intell. **32** (2), 220–230 (2010)
44. Zuo, J., Nicolo, F., Schmid, N., Boothapati, S.: Encoding, matching and score normalization for cross spectral face recognition: matching SWIR versus visible data. In: Proceedings of 5th International Conference on Biometrics Theory, Applications and Systems, pp. 203–208, Washington DC, 23–26 Sept 2012