Can we do better in Unimodal Biometric Systems?
A Novel Rank-based Score Normalization Framework for Multi-sample Galleries

Panagiotis Moutafis and Ioannis A. Kakadiaris
Computational Biomedicine Lab, Department of Computer Science
University of Houston, 4800 Calhoun Rd. Houston, TX 77004
{pmoutafis,ioannisk}@uh.edu

Abstract

The large amount of research on multimodal systems raises an important question: can we extract additional information from unimodal systems? In this paper, we propose a rank-based score normalization framework that addresses this problem when multi-sample galleries are available. The main idea is to partition the matching scores into subsets and normalize each subset independently. In addition, we present two versions of our framework that: (i) use gallery-based information (i.e., gallery versus gallery scores), and (ii) update available information in an online fashion. We use the theory of Stochastic Dominance to illustrate that the proposed framework can increase the system’s performance. Our approach: (i) does not require tuning of any parameters, (ii) can be used in conjunction with any score normalization technique and any integration rule, and (iii) extends the use of W-score normalization to multi-sample galleries. While our approach is better suited for an Open-set Identification task, we also demonstrate that it can be used for a Verification task. In order to assess the performance of the proposed framework we conduct experiments using the BDCP Face database. Our approach improves the Detection and Identification Rate by 14.87% for Z-score and by 4.82% for W-score.

1. Introduction

In this paper, we focus on the Open-set Identification task for unimodal systems when multiple biometric samples per gallery subject are available. The Open-set Identification is a two-step process: (i) determine whether a probe is part of the gallery, and (ii) return the corresponding identity. The most common approach is to select the maximum matching score for a given probe and compare it against a given threshold. In other words, we match the probe against the gallery sample that appears to be the most similar to it. As a result, the Open-set Identification problem can be considered to be a hard Verification problem. A detailed discussion is presented by Fortuna et al. [6]. This is not the only reason why the Open-set Identification task is a hard problem. Each time that a subject submits its biometric sample to the system there are a number of variations that may occur (e.g., differences in pose, illumination and other conditions during data acquisition). Consequently, each time that a different probe is compared against the gallery, the matching scores obtained follow a different distribution. One of the most efficient ways to address this problem is score normalization. Such techniques map scores to a common domain where they are directly comparable. As a result, a global threshold may be found and adjusted to the desired value. Score normalization techniques are also very useful when combining scores in multimodal systems. Specifically, different classifiers from different modalities produce heterogeneous scores. Normalizing these scores before combining them is thus crucial for the performance of the system [8]. Even though this paper is not focusing on multimodal systems, the relative results are very useful in understanding the intuition behind the proposed approach.

For the rest of this paper, we consider the following scenarios: (i) the gallery set is comprised of multiple samples per subject from a single modality, and (ii) the gallery set is comprised of a single sample per subject from different modalities. We shall refer to the former scenario as unimodal and to the latter as multimodal. We note that the integration of scores in the unimodal scenario is an instance of the more general problem of combining scores in the multimodal scenario [12]. To distinguish between the two we say that we integrate scores for the former while we combine scores for the latter. We notice that a search in Google Scholar for the last ten years (i.e., 2002-2012) returns 322 papers that include the terms multimodal and biometric in their title while only eight entries are found for the unimodal case. The question that arises is whether there is space for improvement in the performance of unimodal systems.

In this paper, we propose a rank-based score normalization framework that is suitable for unimodal systems...
when multi-sample galleries are available. Specifically, the first algorithm partitions the set of scores into subsets and then normalizes each subset independently. The normalized scores may then be integrated using any suitable rule. We use the Stochastic Dominance theory to illustrate that our approach imposes the subsets’ score distributions to be ordered, as if each subset was obtained by a different modality. Therefore, by normalizing each subset individually the corresponding distributions are being aligned and the system’s performance improves. The second algorithm uses the gallery versus gallery scores to normalize the produced scores for a given probe in a gallery-based manner. The obtained normalized scores are then combined with the scores produced by the first algorithm using any combination rule. Finally, the third algorithm uses scores from already referenced probes to augment the gallery versus gallery similarity scores. Thus, it updates the available information in an online fashion. We stress that the purpose of this paper is not to evaluate combination or integration rules, nor to assess score normalization techniques. Instead, we focus on the fact that the proposed approach can increase the performance of unimodal biometric systems. The experimental evaluation is performed using the BEST Development Challenge Problem (BDCP) Face Database [1].

The rest of this paper is organized as follows: Sec. 2 reviews score normalization techniques and combination rules, Sec. 3 provides an overview of the Stochastic Dominance theory and describes the proposed framework, Sec. 4 presents the experimental results, and Sec. 5 concludes this paper with an overview of our findings.

2. Related Work

In this section, we do not present an extensive overview of the literature because the proposed framework can be implemented in conjunction with any combination rule and any score normalization technique. Therefore, we refer only to those approaches used in our experiments.

2.1. Combination Rules

Kittler et al. [9] have studied the statistical background of combination rules. Such rules address the general problem of fusing evidence from multiple measurements. Hence, they are applicable to both integration and combination tasks [12]. We note that the work by Kittler et al. [9] refers exclusively to likelihood values. These rules however are often applied to scores, even if there is not a clear statistical justification for this case. In this paper, we have used the sum rule which, under the assumption of equal priors, is implemented by a simple addition. Even though this rule makes restrictive assumptions it appears to yield good performance [8, 9].

2.2. Score Normalization Techniques

Score normalization techniques are used in order to: (i) accommodate for the variations between different biometric samples and, (ii) align score distributions before combining them. A comprehensive study of such approaches is offered by Jain et al. [8].

Z-score: Due to its simplicity and good performance in many settings this is one of the most widely used and well examined techniques. Specifically, it is expected to perform well when the location and scale parameters of the score distribution can be sufficiently approximated by the mean and standard deviation estimates, respectively. In addition, for scores following a Gaussian distribution this approach can retain the shape of the distribution. We note that the most notable disadvantages of Z-score normalization are: (i) it cannot guarantee a common numerical range for the normalized scores, and (ii) it is not robust, as the mean and standard deviation estimates are sensitive to outliers.

W-score: Scheirer et al. [11] recently proposed the use of a score normalization technique that models the tail of the non-match scores. The greatest advantage of this approach is that it does not make any assumptions concerning the score distribution. It also appears to be robust and yields good performance. In order to apply W-score normalization the user has to specify the number of scores to be selected for fitting. While in most cases it is sufficient to select as few as five scores, we have observed that selecting a small number of scores often yields discretized normalized scores. As a result, 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 required assumptions needed to invoke the Extreme Value Theorem. Another limitation of W-score is that it cannot be applied to multi-sample galleries, unless an integration rule is used first (e.g., sum). Consequently, it is not possible to obtain normalized scores for each sample independently. As it will be demonstrated, the proposed framework addresses this problem and extends the use of W-score normalization to multi-sample galleries.

3. Rank-based Score Normalization Framework

In this section, we first review the Stochastic Dominance theory which covers the theoretical background of the proposed framework. Then, we describe the three algorithms that comprise the Rank-based Score Normalization Framework. Since each algorithm builds on top of the other we begin from the most general case and build our way through to the most restricted one. An overview of the proposed algorithms is presented in Fig. 1.
3.1. Stochastic Dominance Theory

In this section, we present basic concepts of the stochastic dominance theory which is used to cover theoretical aspects of the proposed framework. We note that the theory of stochastic dominance falls within the domain of decision theory and therefore it is widely used in finance [14].

**Definition:** The notation \( X \succeq_{FSD} Y \) denotes that \( X \) first order stochastically dominates \( Y \), if

\[
Pr\{X > z\} \geq Pr\{Y > z\}, \quad \forall z.
\]  

As it is implied by this definition the corresponding distributions will be ordered. This fact becomes more clear by the following lemma.

**Lemma:** Let \( X \) and \( Y \) be any two random variables, then

\[
X \succeq_{FSD} Y \Rightarrow E[X] \geq E[Y].
\]

A proof of this Lemma may be found in [14].

An illustrative example of first order stochastic dominance is depicted in Fig. 1 of Wolfstetter [14] where \( F(z) \succeq_{FSD} G(z) \). Note that the first order stochastic dominance relationship implies all higher orders [5]. In addition, this relation is known to be transitive as implicitly illustrated by Birnbaum et al. [3]. Finally, the first order stochastic dominance may also be viewed as the stochastic ordering of random variables.

### 3.2. Rank-based Score Normalization

In this section, we present the Rank-based Score Normalization (RBSN) that partitions a set of scores into subsets and then normalizes each subset independently. We assume that the gallery consists of multiple samples per subject. An overview is provided in Algorithm 1 below. The notation to be used throughout this paper is as follows:

- \( S^p \): the set of similarity scores for a given probe \( p \) when compared against a given gallery set
- \( S^p_i \): the set of scores that correspond to the gallery subject with \( \text{identity}=i \), \( S^p_i \subseteq S^p \)
- \( S^p_{i,r} \): the \( r^{th} \) ranked score of \( S^p_i \)
- \( S^{p,N} \): the set of normalized scores for a given probe \( p \)
- \( C_r \): the \( r^{th} \) rank subset, \( \bigcup C_r = S^p \)
- \( |d| \): the cardinality of any given set \( d \)
- \( I \): the set of unique gallery identities
- \( f \): a given score normalization technique

**Algorithm 1** Rank-based Score Normalization

1. **procedure** RBSN\((S^p = \bigcup i \{S^p_i\}, f)\)
2. **Step 1:** Partition \( S^p \) into subsets
3. \( C_r = \{\emptyset\}, \forall r \)
4. **for** \( r = 1 : \max_i \{|S^p_i|\} \)** **do**
5. **for all** \( i \in I \)** **do**
6. \( C_r = C_r \cup S^p_{i,r} \)
7. **end for**
8. **end for**

\( (\text{i.e., } C_r = \bigcup_i S^p_{i,r}) \)

**Step 2:** Normalize each subset \( C_r \)
9. \( S^{p,N} = \{\emptyset\} \)
10. **for** \( r = 1 : \max_i \{|S^p_i|\} \)** **do**
11. \( S^{p,N} = S^{p,N} \cup f(C_r) \)
12. **end for**
13. **return** \( S^{p,N} \)

**Step 1 - Partition \( S^p \) into subsets:** The main goal is to partition the set of scores \( S^p \) into subsets \( C_r \). The term \( S^p_i \) denotes the set of scores that correspond to the gallery subject with \( \text{identity}=i \). Each subset \( C_r \) is formed by selecting the \( r^{th} \) highest score from every set \( S^p_i \). This procedure is repeated until all scores in \( S^p \) have been assigned to a subset \( C_r \). Each curve in Fig. 2 depicts the kernel density estimate that corresponds to a subset \( C_r \) obtained from Step 1 of RBSN.
Given the relevant results from Sec. 3.1, it is clear that the corresponding to a \( C_r \) subset. Each subset \( C_r \) was constructed by the Step 1 of RBSN using the set \( S^p \) for a random probe.

We now demonstrate that the ordering of the densities in Fig. 2 is imposed by the rank-based construction of the subsets \( C_r \). By construction we have that

\[
S^p_{x,i} \geq S^p_{x,j}, \forall i \leq j \text{ and } \forall x. \tag{3}
\]

Let \( X_i \) and \( X_j \) be the variables that correspond to \( S^p_{x,i} \) and \( S^p_{x,j} \) (i.e., \( C_i \) and \( C_j \)). As shown by Hadar and Russell [7], this condition is sufficient to conclude that \( X_i \geq_{FSD} X_j \). Given the relevant results from Sec. 3.1, it is clear that the densities \( P_{X_i} \) and \( P_{X_j} \) are ordered if \( i \neq j \).

**Step 2 - Normalize each subset \( C_r \):** The set \( S^{p,N} \) is initially an empty set which is gradually updated by adding normalized scores to it. Specifically, the scores for a given subset \( C_r \) are normalized independently of the other subsets and then added to \( S^{p,N} \). We iterate until the scores of all the subsets \( C_r \) have been normalized and added to \( S^{p,N} \).

### 3.2.1 Key Remarks and Implementation Details

First, we explain why the obtained set \( S^{p,N} \) yields increased performance compared to other approaches. Under the multimodal setting a set of scores is obtained from each modality. Each such set consists of a single sample per gallery subject, and since they are obtained by different classifiers, they are heterogeneous. In order to increase the systems’ performance a score normalization step that maps the scores into a common domain is needed before they are combined [8, 12]. The obtained subsets from Step 1 include by construction at most one sample per gallery subject and they are also ordered. Using exactly the same arguments made in previous works for the multimodal case it is clear that our approach yields increased performance. After all, since we do not normalize scores from different modalities together there is no good reason to do so under the unimodal scenario. One important limitation though is that we cannot make any inferences concerning the score distributions of the subsets \( C_r \). Even if the set of scores for a given probe is known to follow a certain distribution, the resulting subsets might follow a different, unknown distribution. Despite this, our experiments indicate that RBSN yields increased performance in practice. In addition, the use of W-score normalization which does not make any assumptions concerning the scores’ distribution is now feasible as the constructed subsets include at most one score per subject.

Next, we provide a brief discussion concerning implementation details. We note that ties can be broken arbitrarily since they do not affect the final outcome. Also, ranking the scores for each gallery subject can be implemented in parallel. The same applies for the normalization of each subset \( C_r \). Hence, our framework is quite scalable. Finally, we note that gallery sets with different number of samples per subject result in subsets with different number of elements (see Fig. 1 for such an example). Consequently, it is possible to obtain a subset which has too few elements, and therefore a score normalization technique cannot be applied. This is likely to happen for low rank subsets. In such cases, we may substitute or estimate these scores in many ways (e.g., use the corresponding normalized score we would obtain if RBSN was not used). For the purposes of this paper, we replace such scores with Not a Number (NaN) and we do not consider them at a decision level.

### 3.3. Rank-based Score Normalization aided by Gallery-based Information

In this section, we present the Rank-based Score Normalization aided by Gallery-based Information (GRBSN) that exploits additional information to improve the performance even further. Specifically, we compare the gallery against itself and we organize the produced scores using a symmetric matrix \( G \). Each element \( g_{i,j} \) corresponds to the similarity score obtained by comparing the \( i^{th} \) and \( j^{th} \) samples of the gallery. We summarize the proposed approach in Algorithm 2 below.

The additional notation to be used is as follows:

- \( G \): the gallery versus gallery similarity scores matrix
- \( g_{i,j} \): the similarity score obtained by comparing the \( i^{th} \) and \( j^{th} \) elements of the gallery, \( g_{i,j} \in G \)
- \( n \): the number of columns of \( G \)
- \( S^{p,N} \): the set of normalized scores \( S^p \)
- \( h \): a given integration rule

Note that there is a correspondence between \( G \) and \( S^p \). That is, the \( i^{th} \) row/column of \( G \) refers to the same gallery sample as the \( i^{th} \) score of \( S^p \).

**Step 1 - Augment \( G \):** One more column is added to the matrix \( G \) that is comprised of the scores in \( S^p \).

**Step 2 - Normalize the Augmented \( G \):** Each row of the augmented matrix \( G \) is treated independently and normalized using RBSN. The probe \( p \) is unlabeled and thus the last score of each row of \( G \) is not associated with any identity.
Algorithm 2 Rank-based Score Normalization aided by Gallery-based Information

1: procedure GRBSN($G, S^p = \bigcup_i \{S^p_i\}, f, h$)  
   Step 1: Augment $G$
2: \quad \{g_{n+1}\} = S^p \triangleright n \rightarrow n + 1
   Step 2: Normalize the Augmented $G$
3: \quad $w = h(RBSN(S^p, f))$
4: \quad Associate the $n$th column of $G$ with the gallery identity that corresponds to the Rank $- 1$ score of $w$
5: \quad for $i = 1 : |g_{n+1}|$ do
6: \quad \quad $g_{i,n+1} = RBSN(g_{i,,f})$
7: \quad end for
   Step 3: Compute $S^{p,N}$
8: \quad $S^{p,N} = h(RBSN(S^p, f), g_{n+1})$
9: \quad return $S^{p,N}$
10: end procedure

To address this problem, $RBSN(S^p, f)$ is computed and an integration rule $h$ is applied. The gallery identity that is associated with the resulting Rank $- 1$ score is used to label the scores of the last column of the augmented matrix $G$. Normalizing each row of $G$ using $RBSN$ is now feasible and the row-wise normalized matrix $G$ is obtained. In practice, each time that $RBSN$ is applied only the subset $G_r$ that corresponds to the score of the last column of $G$ needs to be normalized. Also, the normalization of each row of $G$ can be implemented in parallel.

Step 3 - Compute $S^{p,N}$: The last column of the augmented matrix $G$ contains the Gallery-specific normalized scores that correspond to the probe $p$. The $RBSN(S^p, f)$ corresponds to the Probe-specific normalized scores for the same probe $p$. Hence, the two vectors are combined using the relevant rule $h$.

In this approach, each score in $S^p$ is normalized in relation to: (i) scores in $S^p$, and (ii) scores contained in each row of $G$ (see Fig. 1). Thus, we obtain both Probe- and Gallery-specific normalized scores using each time two different sources of information. Since the same rules are suitable when combining evidence for multiple measurements in general [12], combining the two vectors is reasonable and results in increased performance.

3.4. Online Rank-based Score Normalization

In this section, we build upon the GRBSN algorithm and present an online version of the proposed framework (ORBSN). This version uses information from probes that have been submitted to the system in the past in order to enhance the available information. We provide an overview of the Online Rank-based Score Normalization in Algorithm 3 below. As in Sec. 3.3, we use information from the gallery versus gallery similarity scores. The additional notation to be used is as follows:

$P$: the set of probes presented to the system
$S$: the set of scores for all probes, $\bigcup_p S^p = S$
$S^N$: the set of normalized scores $S$
t: the threshold used to reach a decision

Algorithm 3 Online Rank-based Score Normalization

1: procedure ORBSN($G, S = \bigcup_p \{S^p\}, f, h, t$)  
   Step 1: GRBSN
2: \quad for all $p = 1 : |P|$ do
3: \quad \quad $S^{p,N} = GRBSN(G, S^p, f, h)$
4: \quad \quad Step 2: Augment $G$
5: \quad \quad \quad $w = h(S^{p,N})$
6: \quad \quad \quad if \ max(w) \geq t then
7: \quad \quad \quad \quad \{g_{n+1}\} = S^p \triangleright n \rightarrow n + 1
8: \quad \quad \quad end if
9: \quad \quad return $[S^N = \bigcup_p \{S^{p,N}\}, G]$
10: end procedure

Step 1 - GRBSN: In this step, GRBSN is applied using the corresponding inputs.

Step 2 - Augment $G$: If the system is confident that the probe is part of the gallery the input matrix $G$ is augmented. Specifically, if the Rank-1 score of the $h(S^{p,N})$ is above a given threshold $t$ a new column is added to $G$ comprised of the scores in $S^p$. Hence, the gallery-based information is enhanced and the augmented matrix $G$ will be used when the next probe is submitted to the system. Note that the matrix $G$ contains raw scores at all times and it is only normalized implicitly when GRBSN is invoked.

The intuition of ORBSN is very similar with the idea presented in Sec. 3.3 in the sense that we apply both Probe- and Gallery-specific normalization. Though, when the system is confident about the identity of an unknown probe the corresponding scores are incorporated to the gallery-based information (i.e., matrix $G$). This way, the available information is enhanced in an online fashion in order to improve the performance even further. However, we note that the performance of both GRBSN and ORBSN is sensitive to the identity estimation of the submitted and referenced probes.

4. Experimental Results

In this section, we provide information about: (i) the database used, (ii) implementation details, (iii) evaluation measures, and (iv) experimental results.

The BDCP Database: The BDCP Face database consists of data from 100 subjects [1]. The gallery set is formed by 95 subjects for which 381 3D images have been captured using the Minolta VIVID 900/910 sensor. The number of samples per gallery subject varies from one to six. The probe set is comprised of data from all 100 subjects. Specifically, 2,357 2D images are used which have been captured
Table 1: Summary of the results from the Experiments 1 & 2. The OSI-E and max-DIR refer to absolute values, where for the latter the maximum performance is reported. The osi-$\Delta$AUC and vr-$\Delta$AUC refer to the relative improvement of the raw scores performance. The relative experimental protocol is described in Sec. 4.

<table>
<thead>
<tr>
<th>BDCP Database</th>
<th>Open-set Identification</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methodology</td>
<td>OSI-E (%)</td>
<td>max-DIR (%)</td>
</tr>
<tr>
<td>Z-score</td>
<td>70.93</td>
<td>20.90</td>
</tr>
<tr>
<td>RBSN:Z-score</td>
<td>68.88</td>
<td>22.03</td>
</tr>
<tr>
<td>GRBSN:Z-score</td>
<td>66.06</td>
<td>23.73</td>
</tr>
<tr>
<td>ORBSN:Z-score</td>
<td>60.54</td>
<td>35.77</td>
</tr>
<tr>
<td>W-score</td>
<td>74.53</td>
<td>25.47</td>
</tr>
<tr>
<td>RBSN:W-score</td>
<td>74.88</td>
<td>25.12</td>
</tr>
<tr>
<td>GRBSN:W-score</td>
<td>72.01</td>
<td>27.99</td>
</tr>
<tr>
<td>ORBSN:W-score</td>
<td>69.71</td>
<td>30.29</td>
</tr>
</tbody>
</table>

By four different cameras: (i) Nikon D90, (ii) Canon Powershot Elph SD1400, (iii) Olympus FE-47, and (iv) Panasonic Lumix FP1. The composition of the probe set is: (i) 876 Far Non-Frontal, (ii) 880 Far Frontal, (iii) 305 Near Non-Frontal, and (iv) 296 Near Frontal Faces. The scores used in this paper have been provided by Toderici et al. [13].

**Implementation Details:** As mentioned before, it is not possible to use W-score normalization directly to multi-sample galleries. Consequently, when reporting results for W-score it is implied that the scores have been integrated before they are normalized. We note that: (i) five scores are used to fit a Weibull distribution as in the original work, and (ii) the sum rule is used whenever it is needed to combine or integrate scores. Therefore, we use the notation $RBSN:Z$-score to indicate that Z-score normalization has been used as an input to Algorithm 1 and the resulting normalized scores have been integrated using the sum rule. As mentioned, normalizing sets with small number of elements is usually problematic. Therefore, we do not normalize subsets $C_i$ that include less that six scores; instead, we replace these values by NaN. We do the same for subsets that have a standard deviation less than $10^{-5}$ because in such cases Z-score becomes unstable. Finally, in order to assess the performance of ORBSN we have followed a Leave-One-Out approach. Each time we assume that $|P| - 1$ probes have been submitted to the system and a decision has to be made for the remaining one. The system uses a threshold to select a subset of the $|P| - 1$ probes for which it is confident concerning the estimated identity. Then, the corresponding scores are incorporated to matrix $G$ and ORBSN is implemented as in Algorithm 3. In our implementation, the system used the gallery versus gallery scores to select a threshold that results in a False Alarm Rate of $10^{-3}$. The reasoning behind this decision is that, even if only a few probes will be selected to update the matrix $G$, the corresponding information will be reliable. In other words, we prefer selecting a small number of probes that add information of a good quality rather than selecting many probes that induce noise.

**Performance Measures:** In this paper, we use the osi-ROC that compares the Detection and Identification Rate (DIR) against the False Acceptance Rate (FAR) to provide an overview of the system’s performance for the Open-set Identification task [2]. For the same task we also use the Open-set Identification Error (OSI-E). That is, the rate at which the max score for a probe corresponds to the wrong identity given that the subject depicted in the probe is part of the gallery [6]. Specifically, the OSI-E is inversely proportional to the percentage of correct classifications based on the rank-1 scores for probes in the gallery, a metric that is usually reported for the closed-set identification task. Moreover, we use the ROC that compares Verification Rate (VR) against the False Acceptance Rate (FAR) to assess the discriminability of the scores. Note that different methods result in FARs with different ranges and hence we cannot compare directly quantities such as Area Under the Curve (AUC) and max-DIR. Therefore, we select a common FAR range by setting the universal lower and upper bound to be equal to the $\inf$ and $\sup$, respectively. Finally, we denote by $\Delta AUC$ the relative improvement of the raw scores performance (e.g., $\Delta AUC = (AUC_{RBSN} - AUC_{raw})/AUC_{raw}$).

**Experiment 1:** The objective of this experiment is to assess the performance of the proposed framework under the Open-set Identification task. Specifically, each of the 2,357 probes is compared against the gallery to obtain the corresponding similarity scores. Based on Fig. 3 and Table 1, it appears that the proposed framework improves the overall DIR performance. Note that the performance in most of the evaluation measures gradually improves when RBSN, GRBSN, and ORBSN are used. In addition, we note that most of the score normalization approaches are linear transformations. Hence, they cannot change the order of scores for a given probe. In other words, they cannot effect the OSI-E [6]. However, the proposed framework appears to significantly reduce the OSI-E (Table 1). This is attributed to the fact that it normalizes each subset $C_i$ independently and the final ordering of the scores is thus changed. In order
to investigate this fact even further, we used statistical hypothesis tests to show that the discriminability of the scores for a given probe increases when our framework is used. Specifically, for each probe we computed the ROC curves and the corresponding AUCs. We used these AUC values to perform a non-parametric Wilcoxon Signed-Rank test in order to determine whether the relative median values are equal (i.e., null hypothesis) or the proposed framework performs better (i.e., one sided alternative hypothesis). The Bonferroni correction was used in order to ensure that the overall statistical significance level (i.e., $\alpha = 5\%$) is not overestimated due to the multiple tests performed. That is, the statistical significance of each individual test is set to $\frac{\alpha}{m}$, where $m$ is the number of tests performed. The corresponding p-values are: (i) Z-score Vs. Raw Scores: 1, (ii) RBSN:Z-score Vs. Z-score: $9.5 \cdot 10^{-92}$, and (iii) RBSN:W-score Vs. Raw Scores: 0.0025. The test W-score Vs. Raw Scores cannot be performed due to the fact that W-score is not applicable to multi-sample galleries. According to these results, we conclude that the discriminability of scores for each probe increases significantly when the proposed framework is used (i.e., the null hypothesis is rejected). However, this is not the case for Z-score, which yields identical AUC values to the raw scores.

Experiment 2: The objectives of this experiment are to illustrate: (i) the use of the proposed framework for a Verification task, and (ii) the increased discriminability that it yields. Specifically, we assume that each probe targets all the gallery samples, one at a time. We further assume that each time this happens the rest of the gallery samples may be used as cohort information. Although the experimental protocol employed is an unlikely scenario, it provides rich information for evaluating the performance of the proposed framework under a Verification task. Based on the relative experimental results it appears that the proposed framework improves the verification performance. The corresponding ROC curves for Z-score are depicted in Fig. 4. Note that RBSN outperforms both GRBSN and ORBSN when Z-score is used. We identify this cause to be the poor identity estimates that the corresponding algorithms use. Specifically, we observe that ORBSN appears to be more robust, even though it aggregates the GRBSN errors. This is due to two reasons: (i) it uses more information, and (ii) it uses a threshold to filter the information to be used. Consequently, one obvious modification of Algorithm 2 would be to use the gallery-based normalization only when the system is confident concerning the estimated identity (i.e., use a threshold). This approach though is not followed in our implementation in an effort to minimize the number of parameters used in each algorithm. An alternative solution would be to normalize the scores in a Gallery-based manner without implementing RBSN. This way it would not be needed to estimate the probe’s identity but on the other hand the quality of the normalized scores obtained would be of a lower quality.

Experiment 3: The objective of this experiment is to assess the effect that different number of samples per gallery subject has to the performance of RBSN. To this end, we randomly selected one, three, and five number of samples per subject and computed the ROC and the corresponding AUCs. This procedure was repeated 100 times and the obtained values are depicted in the corresponding boxplots in Fig. 5. For the statistical evaluation of these results we used the non-parametric Mann-Whitney U test. Specifically, the null hypothesis assumes that the median values are equal when three and five samples are used while the alternative assumes that the median AUC value is increased when five samples are used. The obtained p-values are: (i) Raw Scores: 0.3511, (ii) Z-score: 0.2182, (iii) RBSN:Z-score: $2.1 \cdot 10^{-33}$, and (iv) RBSN:W-score: $1.3 \cdot 10^{-34}$. The statistical evidence indicates that an increase in the number of samples per gallery subject results in improved discriminability when the proposed framework is used. As in Experiment 2, the Bonferroni correction ensures that the overall statistical significance is $\alpha = 5\%$. Finally, we note that...
by repeating the same tests for five samples per subject versus one sample per subject the null hypothesis is rejected for all the methods. However, the former set of tests illustrates that an increase in the number of samples per subject results in greater increase of performance when our framework is employed compared to when our framework is not used for score normalization.

5. Conclusions

In this paper, we proposed a rank-based score normalization framework for multi-sample galleries that consists of three different algorithms. Unlike other approaches (e.g., [4, 10]), our first algorithm uses the rank of the scores for each gallery subject to partition the gallery set and normalizes each subset individually. The second algorithm exploits gallery-based information, while the third algorithm updates the available information in an online fashion. The experimental evaluation indicates that the use of our framework for unimodal systems results in increased performance when multiple samples per subject are available. Specifically, according to the relevant statistical tests the performance of our framework improves as we increase the number of samples per subject. Also, it yields increased discriminability within the scores of each probe.

Acknowledgments

This research was funded in part by the National Science Foundation (DMS-0915242) and the UH Hugh Roy and Lilie 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. Finally, the authors would like to thank Prof. M. Papadakis, Prof. P. Tsiamyrtsis and Dr. G. Evangelopoulos for the fruitful conversations and the reviewers for their insightful comments.

References

[1] University of Notre Dame, Computer Vision and Research Laboratory: Biometrics Data Sets.