Can We Do Better in Unimodal Biometric Systems? A Rank-Based Score Normalization Framework

Panagiotis Moutafis, *Student Member, IEEE*, and Ioannis A. Kakadiaris, *Senior Member, IEEE*

***Abstract*—Biometric systems use score normalization techniques and fusion rules to improve recognition performance. The large amount of research on score fusion for multimodal systems raises an important question: can we utilize the available information from unimodal systems more effectively? In this paper, we present a rank-based score normalization framework that addresses this problem. Specifically, our approach consists of three algorithms: 1) partition the matching scores into subsets and normalize each subset independently; 2) utilize the gallery versus gallery matching scores matrix (i.e., gallery-based information); and 3) dynamically augment the gallery in an online fashion. We invoke the theory of stochastic dominance along with results of prior research to demonstrate when and why our approach yields increased performance. Our framework: 1) can be used in conjunction with any score normalization technique and any fusion rule; 2) is amenable to parallel programming; and 3) is suitable for both verification and open-set identification. To assess the performance of our framework, we use the UHDB11 and FRGC v2 face datasets. Specifically, the statistical hypothesis tests performed illustrate that the performance of our framework improves as we increase the number of samples per subject. Furthermore, the corresponding statistical analysis demonstrates that increased separation between match and nonmatch scores is obtained for each probe. Besides the benefits and limitations highlighted by our experimental evaluation, results under optimal and pessimal conditions are also presented to offer better insights.**

***Index Terms*—Fusion, open-set identification, score normalization, unimodal biometric systems, verification.**

I. Introduction

In THIS paper, we focus on score normalization for verification and open-set identification for unimodal biometric systems. Verification is the task of comparing a probe against one or more previously enrolled samples to confirm a subject's claimed identity. Open-set identification is a two-step process: 1) determine whether a probe is part of the gallery, and if it is and 2) return the corresponding identity. The most common approach is to select the maximum matching score obtained for a given probe and compare it against a threshold. In other words, we compare the probe against the gallery subject that appears to be the most similar to it. As a result, the open-set identification problem can be viewed as a hard verification problem. A more detailed discussion concerning this standpoint is provided by Fortuna *et al.* [1]. The performance of biometric systems is usually degraded by a number of variations during data acquisition. For example, differences in pose, illumination, and other conditions may occur. Consequently, each time that a different probe is compared against a given 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 fusing scores in multimodal systems. Specifically, classifiers from different modalities produce heterogeneous score distributions. Normalizing scores before fusing them is thus crucial for the performance of the system [2]. Even though this paper does not focus on multimodal systems the relevant results of prior research are useful in understanding the intuition behind the proposed framework.

For the rest of this paper we consider the following scenarios unless otherwise indicated: 1) the gallery comprises multiple samples per subject from a single modality and 2) the gallery comprises a single sample per subject from different modalities. We refer to the former scenario as unimodal and to the latter as multimodal. The fusion of scores in the unimodal scenario is an instance of the more general problem of fusing scores in the multimodal scenario [3]. To distinguish between the two we say that we integrate scores for the former, while we combine scores for the latter. A search in Google Scholar returns 369 papers that include the terms biometric and fusion in their title and 17 papers that include the terms biometric and score normalization (eight of which refer to multimodal scenarios). The question that arises is whether there is space for improved score normalization for unimodal biometric systems.

In this paper, we present a rank-based score normalization (RBSN) framework that consists of three algorithms, each of which is innovative in a different way. The first algorithm partitions a set of scores into subsets, when multiple samples per subject are available, and then normalizes the scores of each subset independently. The normalized scores may be integrated using any suitable rule. We invoke the theory of stochastic dominance to illustrate

Manuscript received June 24, 2013; revised February 22, 2014, May 13, 2014, and August 14, 2014; accepted September 1, 2014. Date of publication December 23, 2014; date of current version November 13, 2015. This work was supported in part by the U.S. Army Research Laboratory under Grant W911NF-13-1-0127 and in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. This paper was recommended by Associate Editor X. Li.

The authors are with the Computational Biomedicine Laboratory, Department of Computer Science, University of Houston, Houston, TX 77004 USA (e-mail: ioannisk@uh.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2379174

that our rank-based approach imposes the subsets' score distributions to be ordered as if each subset was obtained by a different modality. Therefore, by normalizing the scores of each subset individually, the corresponding distributions become homogeneous and the performance increases. The second algorithm utilizes the gallery-based information to normalize scores in a gallery-specific manner when the system is confident concerning the probe's estimated identity. The normalized scores obtained are combined with the output of the first algorithm. Finally, the third algorithm dynamically augments the gallery in an online fashion. That is, the system incorporates probes to the gallery when a confidence level about their estimated identity has been reached. Parts of this research have been presented in [4]. However, in this paper: 1) the second algorithm always utilized the gallery-based information and 2) the gallery-based information was augmented without incorporating probes to the gallery. As a result, the framework did not have a mechanism to deal with incorrect identity estimates and it was only applicable to multisample galleries. The proposed version of our framework addresses these problems. In this paper, we also pose questions about uninvestigated problems up to this date and suggest relevant directions for future research. In addition to the proposed extensions we provide better insights and a more detailed experimental evaluation. Specifically, we use the FRGC v2 [5], [6] and UHDB11 face databases [7] that provide more matching scores and a greater variability during the data acquisition conditions, respectively. In addition, we employ the *Z*-score, MAD, and *W*-score [8] normalization techniques in conjunction with the max and sum fusion rules. Moreover, we present comparative results between the proposed extensions of this paper and the algorithms present in our previous work along with results under optimal and pessimal conditions. The purpose of this paper is not to evaluate combination or integration rules nor to assess score normalization techniques. Instead, we present an approach that has the potential to increase the performance of unimodal biometric systems.

The rest of this paper is organized as follows. Section II reviews score normalization techniques and fusion rules. Section III provides an overview of the stochastic dominance theory and describes the proposed algorithms. Section IV presents the experimental results. Section V provides a discussion about possible directions for future research and Section VI concludes this paper with an overview of our findings.

II. Related Work

In this section, we focus on the advantages and limitations of the methods used in our experiments. This is due to the fact that our framework can be implemented in conjunction with any score normalization technique and any fusion rule.

A. Fusion Rules

Kittler *et al.* [9] have studied the statistical background of fusion rules. Such rules address the general problem of fusing evidence from multiple measurements. Hence, they can be used to both integrate and combine scores [3]. Kittler *et al.* [9] refer exclusively to likelihood values. Nevertheless, such rules 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 and max rules. 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 [2] and [9]. 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.

B. Score Normalization Techniques

Score normalization techniques: 1) accommodate for variations between different biometric samples (e.g., probes) and 2) make score distributions from multimodal biometric systems homogeneous before combining them. A comprehensive study of such approaches is offered by Jain *et al.* [2].

1) Z-score: Due to its simplicity and good performance it is one of the most widely used and well examined techniques. In particular, it is expected to perform well when the location and scale parameters of the score distribution can be approximated sufficiently well by the mean and standard deviation estimates. For Gaussian scores this approach can retain the shape of the distribution. The most notable limitations of *Z*-score are: 1) it cannot guarantee a common numerical range for the normalized scores and 2) it is not robust because the mean and standard deviation estimates are sensitive to outliers.

2) 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 MAD, respectively. Therefore, it addresses the problem of robustness attributed to outliers. However, it is not optimal for scores that follow a Gaussian distribution.

3) W-Score: Scheirer *et al.* [8] 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, 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, we have observed that selecting a small number of scores yields discretized normalized scores. Consequently, it is not possible to assess the performance of the system in low false acceptance rates or false alarm rates (FAR). 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, the proposed framework addresses this problem and extends the use of *W*-score to multisample galleries.

III. RBSN Framework

In this section, we first review the theory of stochastic dominance, which is an important part of the theoretical background

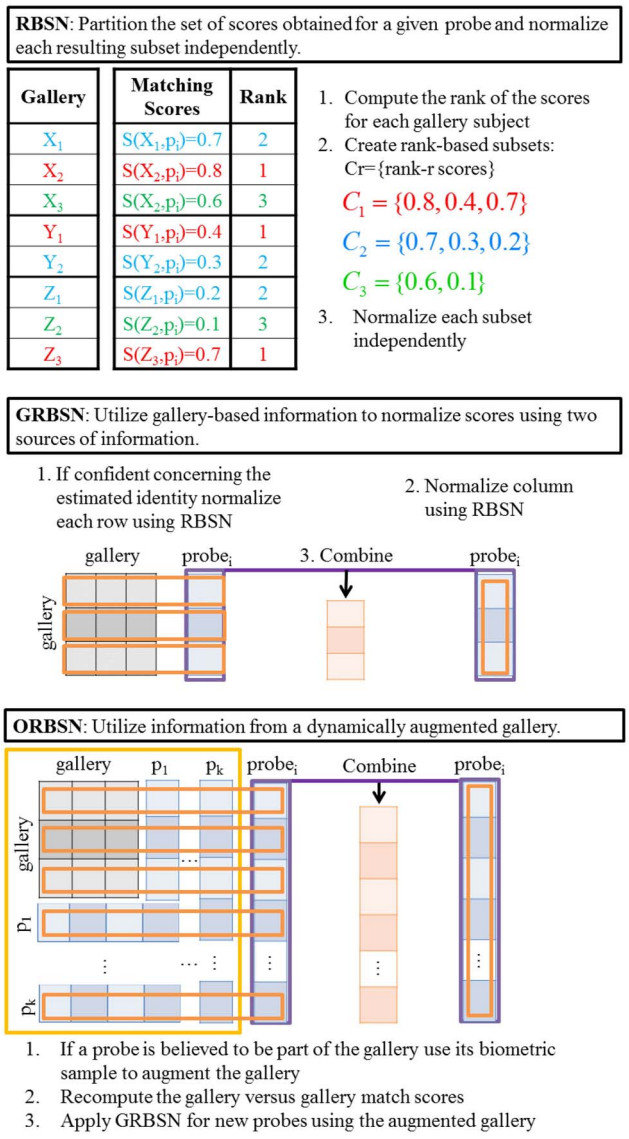

Fig. 1. Overview of the proposed framework (color figure). The notation $S(X_1, p_i)$ is used to denote the score obtained by comparing a probe p_i to the biometric sample 1 of a gallery subject labeled X.

of our approach. Then, we describe the three algorithms that comprise the RBSN. Since each algorithm builds on top of the other we begin from the most general case and build our way to the most restricted one. At the end of each subsection we provide a brief discussion with our insights and implementation details. An overview of the proposed algorithms is presented in Fig. 1.

A. 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. The theory of stochastic dominance falls within the domain of decision theory and therefore it is widely used in finance.

Definition 1: The notation $X \gtrsim_{FSD} Y$ denotes that X first order stochastically dominates Y, that is

$$Pr\{X > z\} \geq Pr\{Y > z\}, \quad \forall z. \tag{1}$$

Algorithm 1 RBSN

procedure RBSN($S^p = \bigcup_i \{S_i^p\}, f$)
Step 1: Partition S^p into subsets
  $C_r = \{\emptyset\}, \forall r$
  for $r = 1{:}\max_i\{|S_i^p|\}$ **do**
    for all $i \epsilon I$ **do**
      $C_r = C_r \bigcup S_{i,r}^p$
    end for
  end for ▷ (i.e., $C_r = \bigcup_i S_{i,r}^p$)
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
  return $S^{p,N}$
end procedure

As implied by this definition the corresponding distributions will be ordered. This fact becomes more clear by the following lemma (its proof may be found in [10]).

Lemma 1: Let X and Y be any two random variables, then

$$X \gtrsim_{FSD} Y \Rightarrow E[X] \geq E[Y]. \tag{2}$$

An illustrative example of first order stochastic dominance is depicted in Fig. 1 of Wolfstetter *et al.* [10], where $\bar{F}(z) \gtrsim_{FSD} \bar{G}(z)$. Note that the first order stochastic dominance relationship implies all higher orders [11]. In addition, this relation is known to be transitive as implicitly illustrated by Birnbaum *et al.* [12]. Finally, the first order stochastic dominance may also be viewed as the stochastic ordering of random variables.

1) Key Remarks: Score normalization techniques are used to make score distributions of different probes and different modalities homogeneous. In Section III-B, we invoke the theory of stochastic dominance to illustrate that a rank-based partitioning of the scores obtained for a single probe yields subsets with ordered score distributions. Therefore, by normalizing the scores of each resulting subset the corresponding score distributions become homogeneous and the separation between match and nonmatch scores increases on a per probe basis.

B. RBSN

In this section, we present the RBSN that partitions a set of scores into subsets and then normalizes the scores of each subset independently. This algorithm can be employed under this assumption that multiple biometric samples per subject are available. An overview is provided in Algorithm 1 below. The notation to be used throughout this paper is as follows.

- 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, $\bigcup_r C_r = S^p$.

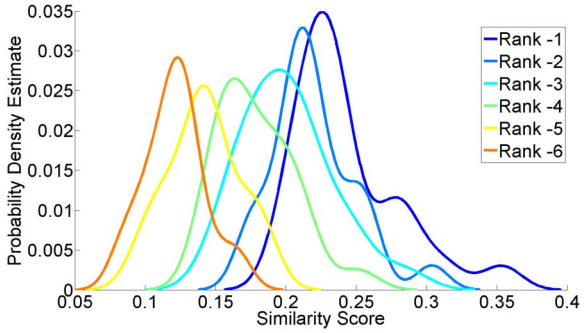

Fig. 2. Each curve depicts the probability density estimate corresponding to a C_r subset. Each subset C_r was constructed by the step 1 of RBSN using the set S^p for a random probe p.

$|d|$ The cardinality of a set d.
I The set of unique gallery identities.
f A given score normalization technique.

1) Step 1—Partition S^p Into Subsets: The main goal is to partition the set of scores S^p into subsets C_r. The symbol S_i^p denotes the set of scores that correspond to the gallery subject with identity equal to i. Each subset C_r is formed by selecting the rank-r score from each set S_i^p. 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 probability density estimate that corresponds to a subset C_r obtained from step 1 of RBSN.

We now demonstrate that the rank-based construction of the subsets C_r imposes the densities in Fig. 2 to be ordered. By construction we have that

$$S_{x,i}^p \geq S_{x,j}^p, \forall i \leq j \quad \text{and} \quad \forall x. \tag{3}$$

Let X_i and X_j be the variables that correspond to $S_{x,i}^p$ and $S_{x,j}^p$ (i.e., C_i and C_j), respectively. As demonstrated by Hadar and Russell [13], this condition is sufficient to conclude that $X_i \gtrsim_{FSD} X_j$. Given the relevant results from Section III-A it is clear that the densities P_{X_i} and P_{X_j} are ordered if $i \neq j$. The subsets can be defined in other ways (e.g., ranking by illumination or pose). However, they might not always yield ordered densities. To illustrate this, we parsed UHDB11 for relevant counter-examples. When the probe 94001d110 (i.e., ID: 94001, illumination: 1, and pose: 10) is matched with the gallery the similarity scores produced for 94004d11 (i.e., same illumination and different pose) and 94004d25 (i.e., different illumination and different pose) are 0.164 and 0.172, respectively. When the probe 94001d12 is matched with the gallery the similarity scores produced for 94002d32 (i.e., same pose and different illumination) and 94001d412 (i.e., different pose and different illumination) are 0.030 and 0.144, respectively. Hence, ranking the scores based on illumination or pose does not guarantee ordered densities.

2) Step 2—Normalize Each Subset C_r: The set $S^{p,N}$ is initially empty and it 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 $S^{p,N}$. This procedure is repeated until the scores of all the subsets C_r have been normalized and added to $S^{p,N}$.

a) Key remarks: First, we elaborate on when and why RBSN yields increased performance. Under the multimodal scenario a set of scores is obtained from each modality. Since each set is obtained by a different classifier the corresponding score distributions are heterogeneous. Therefore, a score normalization step that maps the scores for each set into a common domain is usually employed. Then, the normalized scores are combined to increase the system's performance [2], [3]. The obtained subsets from step 1 include by construction at most one score per gallery subject and their distributions are also ordered. Using exactly the same arguments made in previous works for the multimodal case it is expected that our approach results in 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. In particular, the main strength of our approach compared to the raw implementation of score normalization techniques is that it utilizes more information on a per probe basis. Consequently, it is expected to primarily increase the separation between match and nonmatch scores for each probe independently. However, as in the case of multimodal systems, an improved overall performance cannot always be guaranteed when our framework is employed. For instance, increased separation between the match and nonmatch scores for a single probe (i.e., better rank-1 accuracy) might come at the cost of a decreased separation between the match and nonmatch scores across probes and vice versa. An important limitation of our rank-based approach 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. Nevertheless, our experiments indicate that RBSN yields increased performance in practice. Moreover, the use of the nonparametric W-score normalization is now feasible as the constructed subsets C_r include at most one score per subject.

b) Implementation details: Score ties can be broken arbitrarily when computing the corresponding ranks since they do not affect the final outcome. Both ranking the scores for each gallery subject and normalizing the scores of each subset C_r can be implemented in parallel. Galleries with different numbers of samples per subject result in subsets with different numbers of elements (see Fig. 1). Consequently, it is possible to obtain a subset which has few elements and thus score normalization should not be applied. This is likely to happen for low rank subsets. In such cases, we may substitute such scores in many ways (e.g., use the corresponding normalized score we would obtain if RBSN was not used). In this paper, we replace such scores with not a number (NaN) and we do not consider them at a decision level. Finally, sometimes it is better to first integrate scores before we normalize them and vice versa. The intuition behind this is that in some cases the integrated scores are of a better quality and thus the score normalization can be more effective. By integrating the scores, though, we might reduce and degrade the quality of the available information. In such cases, it is usually better to normalize the scores before we integrate them. In summary, the final performance depends on a combination of factors such as the quality of the scores

Algorithm 2 GRBSN

1: **procedure** *GRBSN*(G, $S^p = \bigcup_i \{S_i^p\}$, f, h, t)
2: **if** $\max(h(\text{RBSN}(S^p, f))) \geq t$ **then**
Case 1: Probe is part of the Gallery
Step 1.1: Augment G
3: $\{g_{.,n+1}\} = S^p$ ▷ $n \rightarrow n+1$
Step 1.2: Normalize the Augmented G
4: Associate the n^{th} column of G with the gallery identity that corresponds to the Rank-1 score of $h(\text{RBSN}(S^p, f))$
5: **for** $i = 1{:}|g_{.,1}|$ **do**
6: $\{g_{i,.}\} = \text{RBSN}(g_{i,.}, f)$
7: **end for**
Step 1.3: Compute $S^{p,N}$
8: $S^{p,N} = h(\text{RBSN}(S^p, f), g_{.,n})$
9: **else**
Case 2: Probe is not part of the Gallery
Step 2.1: Compute $S^{p,N}$
10: $S^{p,N} = \text{RBSN}(S^p, f)$
11: **end if**
12: **return** $S^{p,N}$
13: **end procedure**

at hand, the score normalization method, the fusion rule, and the order in which they are applied. However, there is not an analytical rule that indicates in which cases it is better to integrate scores before we normalize them or vice versa. Since our framework utilizes information from multiple samples per subject we always normalize the scores first and then integrate them.

C. RBSN Aided by Gallery-Based Information

In this section, we present the Rank-Based Score Normalization aided by Gallery-Based Information (GRBSN) that utilizes the gallery versus gallery matching scores. In particular, we compare the gallery with itself and organize the obtained scores using a symmetric matrix G. Each element $g_{i,j}$ corresponds to the matching 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 matching scores matrix.
- $g_{i,j}$ The matching score obtained by comparing the i^{th} and the j^{th} elements of the gallery, $g_{i,j} \in G$.
- n The number of columns in G.
- $S^{p,N}$ The set of normalized scores S^p.
- h A given integration rule.
- t A given threshold.

By construction there is a correspondence between G and S^p. That is, the ith row/column of G refers to the same gallery sample as the ith score of S^p.

1) Case 1—Probe is Part of the Gallery: The RBSN algorithm is employed to normalize the scores by using a given score normalization technique f. Then, an integration rule h is applied to the obtained normalized scores. If the maximum score is above a given threshold t (i.e., $\max(h(\text{RBSN}(S^p, f))) \geq t$) the probe is believed to be part of the gallery set.

1) *Step 1.1—Augment G:* One more column is added to G that comprises the scores in S^p.
2) *Step 1.2—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. To address this problem, the rank-1 score of $h(\text{RBSN}(S^p, f))$ is used to label the scores of the last column of the augmented matrix G. Normalizing each row of G using RBSN is thus feasible and the row-wise normalized matrix G is obtained.
3) *Step 1.3—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 $\text{RBSN}(S^p, f)$ corresponds to the probe-specific normalized scores for the same probe p. Hence, the two vectors are combined using the relevant rule h.

2) Case 2—Probe is Not Part of the Gallery: If the maximum score obtained from $h(\text{RBSN}(S^p, f))$ is not higher than the given threshold t the probe is not believed to be part of the gallery. Consequently, the gallery-based information cannot be utilized using RBSN.

1) *Step 2.1—Compute* $S^{p,N}$*:* The gallery-based information is not used and the scores are normalized by employing the RBSN algorithm.

a) Key remarks: We now detail why GRBSN is expected to produce better normalized scores. Each score in S^p is normalized in relation to: 1) scores in S^P and 2) scores contained in each row of G (see Fig. 1). Thus, we obtain both probe- and gallery- specific normalized scores utilizing each time a different source of information. Combining the two scores is reasonable as such rules are suitable when combining evidence for multiple measurements [3]. The combined scores may be integrated using any integration rule as usual.

b) Implementation details: The implementation presented in Algorithm 2 and Fig. 1 increases the readability of this paper and helps the reader understand the proposed approach. In practice an optimized version may be employed to speed up the processing time. For example, each time that RBSN is applied in step 1.2 only the subset C_r that corresponds to the score of the last column of G needs to be normalized. Moreover, the normalization of each row of the augmented matrix G in step 1.2 can be implemented in parallel and thus GRBSN can scale up well. A major limitation of this method, though, is the estimation of the probe's identity (i.e., step 1.2). We can overcome this problem by ignoring the identity estimation (i.e., lines 2, 4, 9, and 10). That is, the gallery-based information is still utilized by normalizing the scores using the normalization technique f but without implementing RBSN in: 1) step 1.2, or 2) steps 1.2 and 1.3. The former approach might produce normalized scores which are not homogeneous and thus induce noise. This heteroskedasticity is attributed to the fact that the gallery-specific scores are normalized in sets of size $n+1$. The probe-specific scores, on the other hand, are partitioned and normalized in sets of size $|I|$ or smaller. Consequently, there

Algorithm 3 ORBSN

procedure *ORBSN*($\mathbb{G}$, G, $S = \bigcup_p \{S^p\}$, f, h, t)
  for all $p = 1{:}|P|$ **do**
    Step 1: Apply GRBSN
    $S^{p,N}$= *GRBSN*(G, S^p, f, h, t)
    Step 2: Augment $\mathbb{G}$ *and Re-compute* G
    if $\max(h(S^{p,N})) \geq t$ **then**
      $\mathbb{G} = \mathbb{G} \bigcup p$ ▷ incorporate probe p
      Recompute the gallery versus gallery similarity scores using the augmented gallery set $\mathbb{G}$
    end if
  end for
  return $[S^N = \bigcup_p \{S^{p,N}\}, \mathbb{G}]$
end procedure

is a mismatch in the amount of information used in the two cases. The latter approach is expected to produce normalized scores of a lower quality since we do not utilize the benefits of RBSN. For these reasons, we adopt the implementation described in Algorithm 2 that both utilizes the benefits of RBSN and normalizes scores organized in subsets of similar sizes. The threshold value of Algorithm 2 can be selected based on prior information. A high threshold will contribute to robust estimates by GRBSN because the identity estimation will be correct in most cases. However, the improvements over RBSN might not be significant due to the small number of cases that the gallery-based information is actually utilized. Using a low threshold oh the other hand might assign incorrect identities to some probes, which would induce noise and degrade the system's performance. An analytical rule that provides a good trade-off cannot be found since the final performance depends on the data at hand along with the score normalization and fusion techniques employed.

D. Online RBSN

In this section, we build upon the GRBSN algorithm and present an online version of the proposed framework. This approach uses information from probes that have already been submitted to the system to dynamically augment the gallery. We provide an overview of the Online RBSN (ORBSN) in Algorithm 3 below. As in Section III-C, we utilize the gallery versus gallery matching scores. The additional notation to be used is as follows.

- $\mathbb{G}$ The set of biometric samples in the gallery.
- 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 A given threshold.

1) Step 1—Apply GRBSN: In this step, GRBSN is applied using the corresponding inputs.

2) Step 2—Augment $\mathbb{G}$ *and Recompute G:* If the probe is believed to be part of the gallery by some rule (e.g., $\max(h(S^{p,N})) \geq t$) p is incorporated to $\mathbb{G}$. Based on the new set of biometric samples in $\mathbb{G}$ the matrix G is updated by recomputing the gallery versus gallery matching scores. When a new probe is submitted to the system it will be compared with the augmented gallery $\mathbb{G}$.

a) Key remarks: The intuition of ORBSN is very similar to the idea presented in Section III-C in the sense that we apply both probe- and gallery- specific normalization. When the system is confident concerning the identity of a submitted probe the corresponding biometric sample is incorporated into the gallery. As a result, the available information is updated in two ways: 1) new probes are compared with an augmented gallery and 2) the gallery-based information utilized by GRBSN is enhanced. Hence, the proposed framework may be applied to single-sample galleries as previously submitted probes may be used to increase the number of samples per gallery subject.

b) Implementation details: An implicit assumption of ORBSN is that some of the submitted probes will be part of the gallery because otherwise $\mathbb{G}$ would never be updated. In addition, each time the gallery is updated, the next submitted probe must be compared with an additional biometric sample. Even worse, each row of G now contains an extra score to be considered when the gallery-specific normalization is employed. This increases the computational complexity of the system. One possible solution to this problem would be to impose an upper bound on the number of samples per gallery subject. Moreover, a rule concerning the most informative subset of biometric samples per gallery subject could be used to determine which samples should be retained in $\mathbb{G}$. However, a general solution to this problem does not currently exist. On the bright side G does not have to be recomputed at each iteration since the matching scores have already been computed. Finally, even though ORBSN and GRBSN share many characteristics, different thresholds could be used in steps 1 and 2. Intuitively, the threshold used to augment $\mathbb{G}$ should have a higher value since any mistakes are going to be propagated.

IV. Experimental Evaluation

In this section, we present the: 1) datasets; 2) evaluation measures; 3) general implementation details; and 4) experimental results.

A. Databases

1) UHDB11: The data in UHDB11 [7] have been acquired from 23 subjects under six illumination conditions. For each illumination condition the subject was asked to face four different points inside the room. This generated rotations on the y-axis. For each rotation on y, three images were also acquired with rotations on the z-axis (assuming that the z-axis goes from the back of the head to the nose and that the y-axis is the vertical axis through the subject's head). Images under six illumination conditions, four y rotations, and three z rotations per subject were acquired. In total, 72 pose/light variations per subject are available for 23 subjects. Hence, as many as 2 742 336 pairwise comparisons can be formed. Each sample consists of both a 2-D image (captured using a Canon DSLR camera) and a 3-D mesh (captured by 3dMD 2-pod optical 3-D system). The scores used in the experiments for UHDB11 were provided by Toderici *et al.* [14].

2) FRGC v2: This database consists of 4007 biometric samples obtained from 466 subjects under different

facial expressions. The method of Ocegueda *et al.* [15] was employed to extract signatures from the 3-D meshes. The Bosphorus database [16] was used as a training set because according to the original paper it yields the best performance when a single training set is used. To compute the Euclidean distance for the 16 056 049 pairwise combinations of the available samples, 27 basis function were used. Finally, the obtained distances were transformed into scores by using the formula score $=$ max(distance) $-$ distance. This way, the scaling of the score distribution was not altered and all the scores lie in the interval $[0, \max(\text{distance})]$. Even though this database provides numerous scores for experimental evaluation it consists of 3-D data. Consequently, distortions such as pose and lighting conditions do not affect the performance of the system and the score distributions are in general homogeneous. This is evident by the high performance reported in the original paper using raw scores (i.e., verification rate (VR) greater than 97% at a FAR of 10^{-3} for the experiments 1–3).

B. Performance Measures

To provide an overview of the system's performance for the open-set identification task the osi-ROC is used that compares the detection and identification rate (DIR) against the FAR [17]. For the same task the open-set identification error (OSI-E) is also used. That is, the rate at which the max score for a probe corresponds to an incorrect identity given that the subject depicted in the probe is part of the gallery [1]. In particular, 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. To assess the verification performance and the separation of match and nonmatch scores, the receiver operating characteristic (ROC) that compares VR against the FAR is used. Different methods result in FARs with different ranges. Hence, quantities such as area under the curve (AUC) and max-DIR are not directly comparable. Therefore, a common FAR range is selected by setting the universal lower and upper bound to be equal to the infimum and supremum, respectively. That is, the universal upper bound is defined to be the minimum of the maximum FAR values obtained for a set of curves to be compared. This FAR adjustment is performed per recognition task (i.e., open-set identification and verification) and per integration rule (i.e., sum and max) for each experiment. Finally, ΔAUC denotes the relative improvement of the raw scores performance (e.g., $\Delta\text{AUC} = (\text{AUC}_{\text{RBSN}} - \text{AUC}_{\text{raw}})/\text{AUC}_{\text{raw}}$).

C. General Implementation Details

For the verification task it is assumed that each probe is compared to all the gallery subjects, one at a time. It is further assumed that the rest of the subjects 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 setting. The *W*-score normalization was implemented by selecting five scores each time to fit a Weibull distribution. This number of scores has been empirically found to be sufficient [8]. However, using five scores for fitting sometimes yields discretized normalized scores. Results for which the corresponding FAR range is less than 10^{-1} are omitted. In addition, *W*-score is not directly applicable to multisample galleries and thus the corresponding results are omitted as well. Subsets C_r that included fewer than ten scores and/or had a standard deviation less than 10^{-5} were not normalized. The reason is that in such cases the available information is not rich enough. The corresponding values were replaced by NaN. The notation RBSN:*Z*-score is used to indicate that *Z*-score normalization has been used as an input to Algorithm 1. For GRBSN and ORBSN results are reported based on: 1) optimal identity estimation (i.e., GRBSN:*Z*-score$_o$); 2) pessimal identity estimation (i.e., GRBSN:*Z*-score$_p$); 3) threshold t_i (i.e., GRBSN:*Z*-score$_{t_i}$); and 4) our previous work (i.e., GRBSN:*Z*-score$_b$). To specify the threshold values t_i the matrix G was used. That is, G was normalized using RBSN. The value that results in the minimum nonzero FAR under an open-set identification task was used as the first threshold (i.e., t_1). The second threshold t_2 was defined as the value that yields an FAR value 100 times higher than the minimum nonzero FAR which was used for computing t_1. The user may employ different fusion rules at different steps of GRBSN and ORBSN. In our experience, the max rule yields better rank-1 classification, while the sum rule results in less discretized scores, which is better when the point of interest is the separation of the match and nonmatch scores. Therefore, the max rule was used when needed to estimate the probe's identity and the sum rule to combine scores in the corresponding algorithms. For the integration task results are provided for both rules as indicated in the corresponding tables. To assess the performance of ORBSN a leave-one-out approach was adopted. Each time it is assumed that $|P| - 1$ probes have been submitted to the system and a decision needs 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 samples are incorporated into the gallery $\mathbb{G}$ and ORBSN is implemented as in Algorithm 3. While this approach does not track the improvement over time it provides deterministic results and avoids randomization errors (i.e., ordering of the submitted probes). The matrix G contains only units in the diagonal. These values were replaced by NaN to avoid distortions when estimating the necessary parameters (e.g., mean value in *Z*-score).

D. Experimental Results

1) Experiment 1: The objectives of this experiment are to assess: 1) the performance of RBSN for the open-set identification and verification tasks; 2) the effect of RBSN on the match and nonmatch scores separation on a per probe basis; and 3) the effect of different illumination conditions. For UHDB11 six samples per subject were used to form the gallery, one for each illumination condition. As a result, random effects between the gallery subjects due to lighting conditions were eliminated. Five, six, and ten match scores are reported in [18]–[20], respectively. The gallery for this experiment comprises 138 samples and the remaining 1486 samples

TABLE I
SUMMARY OF THE RESULTS FROM EXPERIMENT 1. THE OSI-E AND MAX-DIR REFER TO ABSOLUTE VALUES, WHERE FOR THE LATTER THE MAXIMUM DIR PERFORMANCE IS REPORTED. THE OSI-ΔAUC AND VR-ΔAUC REFER TO THE RELATIVE IMPROVEMENT OF THE RAW SCORES PERFORMANCE, WHILE THE SUM AND MAX RULES REFER TO THE INTEGRATION METHOD EMPLOYED. THE RELATIVE EXPERIMENTAL PROTOCOL IS DESCRIBED IN SECTION IV

UHDB11	Sum Rule				Max Rule			
	Open-Set Identification			Verification	Open-Set Identification			Verification
Method	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC
BI	28.73	71.20	92.76	93.05	28.73	71.20	95.27	91.21
Z-score	13.53	86.41	267.69	165.69	10.43	86.41	259.77	173.69
RBSN:Z-score	16.35	81.70	297.04	179.19	10.50	81.70	280.00	181.11
RBSN:W-score	24.23	58.75	194.32	158.08	27.99	NaN	NaN	87.00
MAD	13.53	86.41	260.87	151.01	10.43	86.41	222.57	152.44
RBSN:MAD	14.94	84.52	277.15	155.03	13.53	84.52	138.44	133.82
FRGC v2	Sum Rule				Max Rule			
	Open-Set Identification			Verification	Open-Set Identification			Verification
Method	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC
Z-score	0.74	99.26	100.94	93.82	0.30	99.71	83.47	90.53
RBSN:Z-score	0.66	99.34	100.03	95.13	0.22	99.71	82.54	90.49
RBSN:W-score	3.62	96.01	94.32	69.37	2.73	NaN	NaN	52.69
MAD	0.74	99.26	100.58	82.06	0.30	99.70	68.99	84.55
RBSN:MAD	0.66	99.34	101.37	83.90	0.22	99.78	65.33	83.89

are used as probes. Even though this design yields a closed-set problem, reporting results for the open-set identification task remains relevant. The reason is that the overlap of the match and nonmatch score distributions for open-set identification is higher compared to that obtained for verification [1]. To form the gallery for FRGC v2 1893 samples from 350 subjects were randomly selected. The remaining 2114 samples obtained from 302 subjects were used as probes. This experimental protocol yields an open-set problem. The results are summarized in Table I and selected results are depicted in Fig. 3. For the more general AUC indexes, RBSN appears to increase the performance in $9/16$ cases. For the more specific OSI-E and max-DIR measures, RBSN appears to yield increased performance in $7/15$ cases. In summary, no general conclusions can be drawn about whether RBSN increases the overall performance or not. The performance of RBSN appears to yield better results for FRGC v2 compared to UHDB1. One possible explanation is that the number of scores per subset C_r is about 23 for UHDB11 and 350 for FRGC v2. Therefore, partitioning the scores for UHDB11 results in small subsets which are sensitive to noise. Finally, the ΔAUC values for FRGC v2 are always below 100 which means that the score normalization degrades the performance of the raw scores. This is due to the fact that the biometric samples are all 3-D and thus distortion-free. A more detailed analysis is provided in experiment 4.

To investigate whether RBSN increases the separation of the match and nonmatch scores for a given probe we performed statistical hypothesis testing. In particular, for each probe the ROC curves and the corresponding AUCs were computed. The AUC values were used to perform nonparametric Wilcoxon signed-rank tests [21]. The null hypothesis was set to H_0: the RBSN and raw median AUCs are equal, and the alternative to H_a: the RBSN median AUC is larger than the raw median AUC. The Bonferonni correction was used 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 was set to α/m, where m is the number of tests performed. The obtained p-values for UHDB11 are: 1) Z-score versus raw scores: 1; 2) RBSN:Z-score versus Z-score: $6.8 \cdot 10^{-240}$; 3) RBSN:W-score versus raw scores: $9.5 \cdot 10^{-240}$;

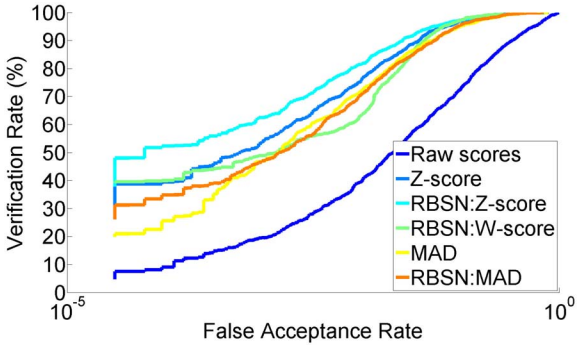

Fig. 3. Depicted are the ROC curves of experiment 1 for UHDB11 when the sum rule is used to integrate the scores for: 1) raw scores; 2) Z-score; 3) RBSN:Z-score; 4) W-score; 5) MAD; and 6) RBSN:MAD.

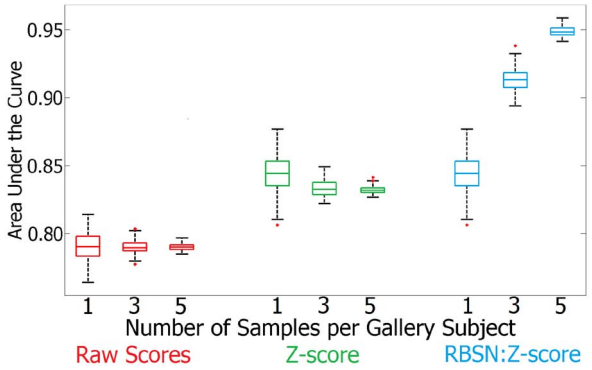

Fig. 4. Depicted are the boxplots for: 1) raw scores; 2) Z-score; and 3) RBSN:Z-score, when one, three, and five samples per gallery subject are randomly selected from UHDB11.

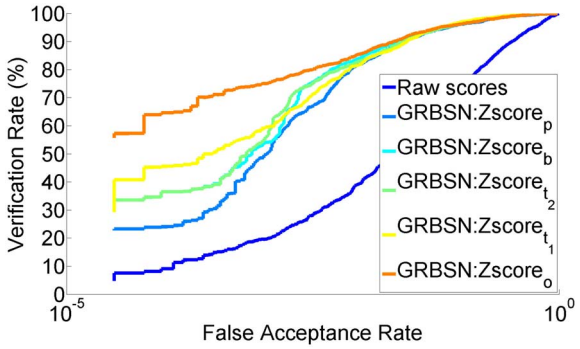

Fig. 5. Depicted are the ROC curves of experiment 3 for UHDB11 when the sum rule is used to integrate the scores for: 1) raw scores; 2) GRBSN:Z-score$_p$; 3) GRBSN:Z-score$_b$; 4) GRBSN:Z-score$_{t_2}$; 5) GRBSN:Z-score$_{t_1}$; and 6) GRBSN:Z-score$_o$.

4) MAD versus raw scores: 1; and 5) RBSN:MAD versus MAD: $4.8 \cdot 10^{-240}$. The corresponding p-values for FRGC v2 are: 1) Z-score versus raw scores: 1; 2) RBSN:Z-score versus Z-score: $2.8 \cdot 10^{-36}$; 3) RBSN:W-score versus raw scores: $2.1 \cdot 10^{-21}$; 4) MAD versus raw scores: 1; and 5) RBSN:MAD versus MAD: $8.1 \cdot 10^{-31}$. The test W-score versus raw scores cannot be performed because W-score is not applicable to multisample galleries. These results indicate that the separation of match and nonmatch scores for each probe significantly increases when RBSN is used (i.e., H_0 is rejected). However, this is not the case for Z-score or MAD, which are linear transformations and yield identical AUC values to the raw scores. As pointed out in Section III, while the separation of the match and nonmatch scores for each probe significantly improves when RBSN is employed it does not always imply a better performance overall.

To systematically assess the effects of different illumination conditions the UHDB11 database was used. In particular, one gallery sample was selected at a time obtained under the same illumination conditions. Then, the performance of the corresponding raw scores was computed. For each of the evaluation measures the illumination condition that resulted in the best performance across probes was retained. The corresponding results are presented in Table I for the method denoted by best illumination (i.e., BI). For the OSI-E and max-DIR measures, BI is compared to Z-score as the latter is equivalent to the raw scores performance. In all cases, using six samples per gallery subject yields better performance than BI. Similarly, both osi- and vr-ΔAUC values for BI are below the ones obtained for the raw scores. We conclude that using multiple samples per gallery subject under various illumination conditions helps the system address the challenge posed by different lighting conditions in the probes. However, the RBSN score normalization utilizes the existing information more effectively and increases the performance even further.

2) Experiment 2: The objective of this experiment is to assess the effect that a different number of samples per gallery subject has on the performance of RBSN. To this end, random subsets of one, three, and five samples per gallery subject were selected for the protocol defined in experiment 1. For each case the ROC curves and the corresponding AUCs were computed. This procedure was repeated 100 times and the obtained values for Z-score are depicted in the corresponding boxplots in Fig. 4 for UHDB11. To statistically evaluate the obtained results nonparametric Mann–Whitney U-tests were performed [21]. Specifically, the null hypothesis is H_0: the median AUCs are equal when three and five samples are used, and the alternative H_a: the median AUC for five samples is larger than the median AUC for three samples. The obtained p-values for UHDB11 are: 1) raw scores: 0.9; 2) Z-score: 0.8; 3) RBSN:Z-score: $1.3 \cdot 10^{-34}$; 4) RBSN:W-score: $1.3 \cdot 10^{-34}$; 5) MAD: 0.14; and 6) RBSN:MAD: $1.3 \cdot 10^{-34}$. The obtained p-values for FRGC v2 are: 1) raw scores: 1; 2) Z-score: 0.5; 3) RBSN:Z-score: 0.3; 4) RBSN:W-score: 0.3; 5) MAD: 0.9; and 6) RBSN:MAD: 0.6. For UHDB11 H_0 is rejected when RBSN is employed and the minimum possible p-value is obtained. For FRGC v2 H_0 is not rejected in any case. However, the smallest p-values are obtained for RBSN. In summary, the statistical analysis suggests that score normalization methods benefit more from an increase in the number of samples per gallery subject if they are used in conjunction with RBSN.

3) Experiment 3: The objectives of this experiment are to assess: 1) the improvements obtained over RBSN by utilizing gallery-based information and 2) the performance of the different versions of GRBSN for the open-set identification and verification tasks. The gallery and probe sets were defined in the same way as in experiment 1. Therefore, the results of experiments 1 and 3 are directly comparable. An overview of the GRBSN performance is presented in Tables II and III and selected results are depicted in Fig. 5. For UHDB11, discretized scores were obtained when the MAD score normalization and max rule were used. Similarly, W-score produced discretized scores for FRGC v2 for both rules. Therefore, the relevant results are provided but they are not taken into consideration in our analysis. For the optimal and pessimal conditions the gallery-based information was utilized for all probes. Across databases, the GRBSN under pessimal conditions outperforms RBSN in 16/35 cases, while under optimal conditions it outperforms RBSN in 31/35 cases. In 22/32 cases at least one of the t_1 or t_2 results in increased performance compared to RBSN. This evidence suggest that utilizing the gallery-based information can be beneficial.

TABLE II
SUMMARY OF THE RESULTS FROM EXPERIMENT 3. THE OSI-E AND MAX-DIR REFER TO ABSOLUTE VALUES, WHERE FOR THE LATTER THE MAXIMUM DIR PERFORMANCE IS REPORTED. THE OSI-ΔAUC AND VR-ΔAUC REFER TO THE RELATIVE IMPROVEMENT OF THE RAW SCORES PERFORMANCE, WHILE THE SUM AND MAX RULES REFER TO THE INTEGRATION METHOD EMPLOYED. THE RELATIVE EXPERIMENTAL PROTOCOL IS DESCRIBED IN SECTION IV

UHDB11	Sum Rule				Max Rule			
	Open-Set Identification			Verification	Open-Set Identification			Verification
Method	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC
GRBSN:Z-score$_p$	9.96	87.21	323.51	157.03	11.78	81.43	205.61	156.35
GRBSN:Z-score$_b$	11.37	79.81	343.36	167.56	9.29	87.75	266.35	167.89
GRBSN:Z-score$_{t_2}$	12.79	77.66	348.19	168.04	11.10	86.99	264.98	165.85
GRBSN:Z-score$_{t_1}$	15.61	74.16	412.50	173.60	10.63	86.14	357.41	173.03
GRBSN:Z-score$_o$	7.47	92.46	693.09	193.60	6.93	93.00	296.64	172.04
GRBSN:W-score$_p$	16.35	56.26	15.63	121.40	16.35	56.26	15.63	121.40
GRBSN:W-score$_b$	17.97	44.41	107.79	138.97	17.97	44.41	107.79	138.87
GRBSN:W-score$_{t_2}$	17.50	34.79	101.38	137.94	17.50	34.79	101.38	137.94
GRBSN:W-score$_{t_1}$	19.04	35.26	103.94	137.59	19.04	35.26	103.94	137.59
GRBSN:W-score$_o$	9.96	89.97	326.67	159.82	9.96	89.97	326.67	159.82
GRBSN:MAD$_p$	13.66	81.63	272.29	152.19	48.86	5.79	9.03	79.56
GRBSN:MAD$_b$	12.72	82.57	298.54	155.24	44.21	5.65	6.09	82.07
GRBSN:MAD$_{t_2}$	14.60	78.73	286.46	151.16	42.33	4.37	4.01	79.68
GRBSN:MAD$_{t_1}$	14.94	78.60	351.79	156.30	15.07	76.45	34.13	109.25
GRBSN:MAD$_o$	9.62	90.31	340.71	156.74	44.41	5.25	5.95	81.91

FRGC v2	Sum Rule				Max Rule			
	Open-Set Identification			Verification	Open-Set Identification			Verification
Method	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC
GRBSN:Z-score$_p$	0.44	99.56	94.59	98.88	0.30	99.71	92.10	95.96
GRBSN:Z-score$_b$	0.37	99.63	91.92	98.73	0.22	99.78	87.97	94.38
GRBSN:Z-score$_{t_2}$	0.52	99.48	91.10	98.39	0.22	99.78	87.42	94.09
GRBSN:Z-score$_{t_1}$	0.66	99.34	81.57	95.19	0.22	99.78	82.56	90.52
GRBSN:Z-score$_o$	0.37	99.63	101.35	100.55	0.22	99.78	96.35	95.52
GRBSN:W-score$_p$	0.81	99.19	88.26	85.27	0.22	99.70	97.10	77.89
GRBSN:W-score$_b$	2.80	97.20	91.83	77.28	0.81	99.19	100.25	59.98
GRBSN:W-score$_{t_2}$	2.66	97.34	60.97	77.51	1.11	98.89	36.59	60.03
GRBSN:W-score$_{t_1}$	2.66	97.34	91.61	77.20	1.03	98.97	25.88	59.96
GRBSN:W-score$_o$	2.07	97.93	100.18	100.96	0.15	99.85	101.89	97.90
GRBSN:MAD$_p$	0.37	99.63	74.59	89.07	0.30	99.70	79.85	90.30
GRBSN:MAD$_b$	0.37	99.63	74.25	88.89	0.30	99.70	78.93	89.01
GRBSN:MAD$_{t_2}$	0.66	99.34	74.39	88.69	0.22	99.79	79.10	88.74
GRBSN:MAD$_{t_1}$	0.66	99.34	62.50	84.03	0.22	99.79	65.36	84.50
GRBSN:MAD$_o$	0.30	99.70	81.35	91.79	0.15	99.85	83.92	90.68

For UHDB11, t_2 yields better results compared to t_1 in $9/20$ cases. When it comes to FRGC v2, though, t_2 yields better results compared to t_1 in all cases. The reason becomes clear when referring to Table III. The percentage of correct identity estimates for FRGC v2 is higher for t_1 but for t_2 the gallery-based information is utilized $1266 + 1118 - 64 - 180 = 2140$ more times. For UHDB11 the percentage of correct identity estimates is significantly lower for t_2 and thus the additional information induces too much noise. The version of GRBSN based on our previous work outperforms the proposed approach for both thresholds $11/20$ times for UHDB11 and $10/14$ times for FRGC v2. This is attributed to the fact that FRGC v2 yields better rank-1 performance than UHDB11 (see Table II). Hence, by always using the gallery-based information the additional information benefits the system. Finally, in $8/20$ cases for UHDB11 and in $12/16$ cases for FRGC v2 the performance of the system under pessimal conditions is better compared to when the two threshold values are used. At the same time, in $19/20$ cases for UHDB11 and in $11/12$ cases for FRGC v2 the performance of the system under optimal conditions is significantly better compared to all the other approaches. In summary, we conclude that while a correct

TABLE III
THIS TABLE PROVIDES STATISTICS ABOUT THE NUMBER OF TIMES THAT THE GALLERY-BASED INFORMATION WAS USED FOR GRBSN WHEN THE TWO THRESHOLD VALUES WERE EMPLOYED IN EXPERIMENT 3. INFORMATION CONCERNING THE CORRECT NUMBER OF IDENTITY ESTIMATES IS ALSO PROVIDED

	UHDB11	$\#probes \geq t_i$	# Correct Estimates
t_1	GRBSN:Z-score	806	806
	GRBSN:W-score	1,408	1,009
	GRBSN:MAD	145	145
t_2	GRBSN:Z-score	1,363	1,278
	GRBSN:W-score	1,485	1,057
	GRBSN:MAD	1,257	1,154
	FRGC v2	$\#probes \geq t_i$	# Correct Estimates
t_1	GRBSN:Z-score	64	64
	GRBSN:W-score	1,938	1,304
	GRBSN:MAD	180	180
t_2	GRBSN:Z-score	1,266	1,233
	GRBSN:W-score	2,114	1,318
	GRBSN:MAD	1,118	1,087

identity estimate is important utilizing the gallery-based information is quite robust to incorrect identity estimates. This is supported by the good performance of the previous version of GRBSN and the obtained results under pessimal conditions.

4) Experiment 4: The objective of this experiment is to assess the performance of ORBSN for the open-set identification and verification tasks using a single-sample gallery. The gallery in UHDB11 comprises both 3-D and 2-D data, while the probes consist of 2-D data only. Hence, it is not possible to incorporate the probes into the gallery and then apply the same method to compute the corresponding matching scores. Consequently, only the FRGC v2 could be used for the purposes of this experiment. Specifically, a single sample from 350 subjects was randomly selected. The remaining 3657 samples obtained from the 466 subjects were used as probes. For the pessimal and optimal scenarios the same probes were used, but for the former incorrect identity estimates were enforced. Moreover, since the matrix G of this experiment does not produce matching scores from different samples of the same subjects the thresholds t_1 and t_2 obtained in experiment 3 were reused. Since a single-sample gallery was available and a leave-one-out approach was adopted, the probes to be incorporated into the gallery were selected by normalizing the scores without implementing RBSN or GRBSN. The results of this experiment are summarized in Table IV and selected results are depicted in Fig. 6. We observe that t_2 yields better performance compared to t_1 for the previous version of ORBSN in $11/18$ cases, while for the proposed version this is true in $12/18$ cases. In $12/18$ cases a better performance is obtained by the previous version of ORBSN compared to the proposed one for at least one of the two thresholds. By taking into consideration the increased performance obtained for the proposed method under the optimal

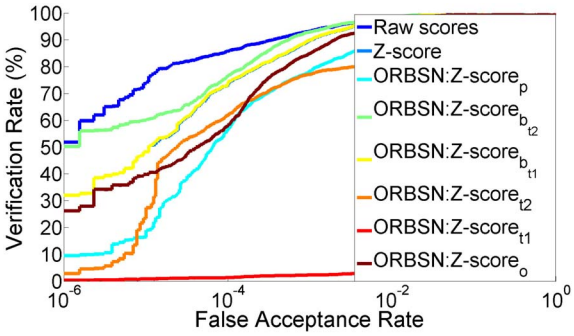

Fig. 6. Depicted are the ROC curves of the experiment 4 when the sum rule is used to integrate the scores for: 1) raw Scores; 2) Z-score; 3) ORBSN:Z-score$_p$; 4) ORBSN:Z-score$_{b_{t_2}}$; 5) ORBSN:Z-score$_{b_{t_1}}$; 6) ORBSN:Z-score$_{t_2}$; 7) ORBSN:Z-score$_{t_2}$; and 8) ORBSN:Z-score$_o$.

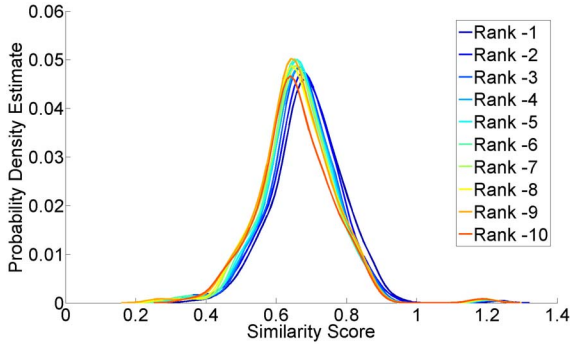

Fig. 7. Each curve depicts the probability density estimate corresponding to a C_r subset. Each subset C_r was constructed by the step 1 of RBSN using the set S^p for a random probe under the optimal scenario defined in experiment 4. That is, the gallery set consists of 3073 biometric samples labeled using the ground truth. Only the first ten ranks are depicted for illustrative reasons.

scenario we conclude that this can be attributed to the fact that the proposed version aggregates the errors. That is, incorrectly labeled samples are incorporated to $\mathbb{G}$ for the proposed version and used when: 1) a new probe is matched against the gallery and 2) the gallery versus gallery scores are recomputed. For the previous version of ORBSN, though, these probes are only used as part of the gallery-based information. In addition, Fig. 7 illustrates that the score distributions of the subsets C_r appear to be homogeneous. Hence, our framework is expected to have problems producing significant improvements for this database since partitioning the scores does not benefit the score normalization. On the contrary, it uses smaller sets to compute the necessary estimates. In relation to Fig. 2, it is evident why the obtained improvements for UHDB11 are more pronounced. Indicatively, for experiment 3 the osi-ΔAUC under the optimal scenario is 693.03 for UHDB11, while for FRGC v2 it is 340.71. Nonetheless, we have demonstrated that our approach has the potential to benefit the performance of the system even in the case of single-sample galleries and approximately distortion-free data.

5) Experiment 5: The goal of this experiment is to systematically evaluate the performance of the proposed method for different poses. To this end, the UHDB11 probes were grouped by pose and individual AUC values were computed for the verification task. To avoid redundancy, only the GRBSN:Z-score$_{t_1}$

TABLE IV
SUMMARY OF THE RESULTS FROM EXPERIMENT 4. THE OSI-E AND MAX-DIR REFER TO ABSOLUTE VALUES, WHERE FOR THE LATTER THE MAXIMUM DIR PERFORMANCE IS REPORTED. THE OSI-ΔAUC AND VR-ΔAUC REFER TO THE RELATIVE IMPROVEMENT OF THE RAW SCORES PERFORMANCE, WHILE THE SUM AND MAX RULES REFER TO THE INTEGRATION METHOD EMPLOYED. THE RELATIVE EXPERIMENTAL PROTOCOL IS DESCRIBED IN SECTION IV

FRGC v2	Sum Rule				Max Rule			
	Open-Set Identification			Verification	Open-Set Identification			Verification
Method	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC	OSI-E (%)	max-DIR (%)	osi-ΔAUC	vr-ΔAUC
Z-score$_p$	2.5	97.5	93.59	89.41	2.5	97.5	93.62	89.41
ORBSN:Z-score$_p$	1.43	98.57	106.70	103.52	1.10	98.90	91.00	101.44
ORBSN:Z-score$_{b_{t_2}}$	2.72	97.28	89.30	95.28	2.72	97.28	89.30	95.28
ORBSN:Z-score$_{b_{t_1}}$	2.50	97.50	93.73	89.52	2.50	97.50	93.70	89.52
ORBSN:Z-score$_{t_2}$	2.31	97.69	59.57	87.81	2.94	97.03	29.92	75.90
ORBSN:Z-score$_{t_1}$	2.64	97.36	83.62	87.66	3.08	96.92	93.92	89.40
ORBSN:Z-score$_o$	0.95	99.05	106.31	100.38	0.77	99.23	95.41	99.58
W-score$_p$	2.50	NaN	NaN	65.85	2.50	NaN	NaN	65.85
ORBSN:W-score$_p$	8.81	NaN	NaN	66.69	12.41	NaN	NaN	67.35
ORBSN:W-score$_{b_{t_2}}$	2.24	NaN	NaN	71.49	2.24	NaN	NaN	71.49
ORBSN:W-score$_{b_{t_1}}$	2.46	NaN	NaN	72.77	2.46	NaN	NaN	72.77
ORBSN:W-score$_{t_2}$	11.75	NaN	NaN	80.83	28.06	NaN	NaN	54.94
ORBSN:W-score$_{t_1}$	2.61	NaN	NaN	77.99	2.60	NaN	NaN	77.99
ORBSN:W-score$_o$	0.99	NaN	NaN	109.00	0.51	NaN	NaN	113.66
MAD$_p$	2.50	97.47	70.43	78.33	2.50	NaN	NaN	78.33
ORBSN:MAD$_p$	1.65	98.35	111.97	104.92	1.40	NaN	NaN	100.85
ORBSN:MAD$_{b_{t_2}}$	4.22	95.70	47.83	78.52	4.22	NaN	NaN	78.52
ORBSN:MAD$_{b_{t_1}}$	2.64	97.32	54.31	72.73	2.64	NaN	NaN	72.73
ORBSN:MAD$_{t_2}$	4.59	95.30	54.94	81.61	10.65	NaN	NaN	64.67
ORBSN:MAD$_{t_1}$	7.30	92.43	18.47	65.30	19.87	NaN	NaN	58.73
ORBSN:MAD$_o$	2.42	97.58	82.99	89.13	1.62	NaN	NaN	87.11

performance is reported for the GRBSN algorithm. In particular, the biometric samples in UHDB11 contain four rotations around the y-axis and three around z. The notation P1–P12 is used to denote a different combination of y and z rotation. The three groups P1–P4, P5–P8, and P9–P12 correspond to the three z rotations. Within each group the same y rotations are used. Specifically, P1, P5, and P9 correspond to the same y rotation and so forth. Finally, P1 corresponds to a subject looking at mark on the bottom left while P12 on the top right. A summary of the results is provided in Table V and selected results are depicted in Fig. 8. As illustrated, one of the RBSN or GRBSN yields increased performance compared to the baselines Z-score and MAD in 19/24 cases. Even though the differences in terms of AUC appear to be small, examining Fig. 8 reveals that for low FAR values the VR improvements obtained are more pronounced. Specifically, the VR values for the lowest FAR possible are 23.88, 53.73, 65.67, and 67.91 for the raw scores, Z-score, RBSN:Z-score, and GRBSN:Z-score$_{t_1}$, respectively.

6) Time Complexity: We used a random probe from FRGC v2 under the protocol defined in experiment 1 and computed the time in seconds for Z-score, RBSN:Z-score and GRBSN:Z-score$_b$. The obtained values are 0.01 s, 0.13 s, and 145.25 s, respectively. By utilizing two, four, and eight cores for GRBSN:Z-score$_b$ we obtained 71.21 s, 40.70 s, and 31.90 s, respectively. In a similar manner, we computed the time in seconds for Z-score by using a random probe from FRGC v2 under the protocol defined in experiment 4. The time needed to normalize the scores was 0.01 s. We provide the corresponding results for ORBSN:Z-score$_o$ when utilizing two, four, and eight cores. The obtained values are 142.44 s, 80.56 s, and 62.21 s, respectively.

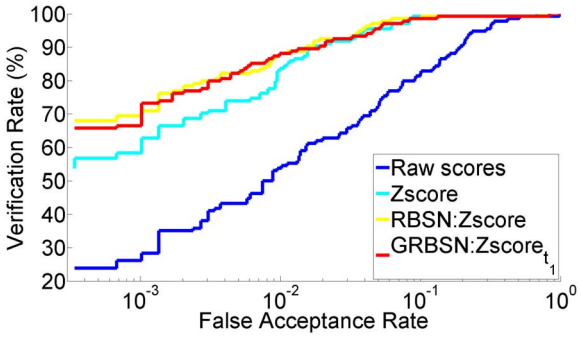

Fig. 8. ROC curves of experiment 5 when the sum rule is used to integrate the scores for: 1) raw scores; 2) Z-score; 3) RBSN:Z-score; and 4) GRBSN:Z-score$_{b_{t_1}}$ for pose P6.

V. FUTURE RESEARCH

In this section, we provide a brief discussion concerning possible directions for future research. First, it is not clear in which cases an improved separation of the match

TABLE V
SUMMARY OF THE RESULTS FROM EXPERIMENT 5. THE RELATIVE AUC IMPROVEMENT OF THE RAW SCORES PERFORMANCE FOR THE VERIFICATION TASK IS REPORTED WHEN THE SUM RULE IS USED TO INTEGRATE THE SCORES. THE NOTATION P1–P12 IS USED TO DENOTE THE 12 DIFFERENT POSES OF PROBES AVAILABLE FOR UHDB11. THE RELATIVE EXPERIMENTAL PROTOCOL IS DESCRIBED IN SECTION IV

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Z-score	118.34	109.12	106.07	108.48	106.60	104.95	105.22	104.89	111.15	106.38	110.55	115.20
RBSN:Z-score	118.79	109.17	106.20	108.59	106.77	105.12	105.35	105.24	111.67	106.73	111.09	115.98
GRBSN:Z-score$_{t_1}$	118.31	109.18	106.11	108.66	106.64	104.91	105.41	105.14	111.39	106.31	110.80	115.56
RBSN:W-score	118.61	108.65	105.46	108.48	106.05	104.61	105.18	104.61	111.09	106.34	110.39	115.45
GRBSN:W-score$_{t_1}$	117.48	108.76	105.79	108.31	105.70	104.22	105.14	104.60	110.39	106.12	110.19	114.20
MAD	117.55	108.79	105.77	108.16	106.48	104.88	104.80	104.64	110.85	105.53	109.75	114.12
RBSN:MAD	117.87	108.41	105.48	107.90	106.11	104.55	104.54	104.65	111.02	105.50	110.10	114.77
GRBSN:MAD$_{t_1}$	117.79	108.41	105.46	107.79	106.06	104.53	104.51	104.68	110.89	105.44	110.03	114.69

and nonmatch scores per probe increases the overall performance of the system. This phenomenon was evident in the results of experiment 1. Future research could provide sophisticated solutions to address this problem. Second, it remains unclear under which conditions it is better to fuse scores before normalizing them and vice versa. Third, investigation of the performance of the proposed method in multimodal systems is being planned. Fourth, assessing the impact of other fusion methods (see [22]) is of great interest. Fifth, methods for automatically selecting the threshold values for GRBSN and ORBSN could increase the usability and performance of our framework greatly. Finally, as illustrated by the time complexity analysis, the composition of the gallery is very important for the scalability of our approach. Developing appropriate gallery selection techniques could simultaneously reduce the time cost and increase the recognition performance.

VI. CONCLUSION

In this paper, we presented a RBSN framework that consists of three different algorithms. Unlike other approaches (see [23], [24]), our first algorithm uses the rank of the scores for each gallery subject to partition the set of scores for a probe and then normalizes the scores of each subset individually. The second algorithm utilizes gallery-based information, while the third algorithm dynamically augments the gallery in an online fashion and can be employed to single-sample galleries. The experimental evaluation indicates that the use of our framework for unimodal systems may yield increased performance. Utilizing the gallery-based information by estimating the probe's identity appears to be robust. Furthermore, the online version of our framework appears to result in significantly better performance for single-sample galleries when the identity estimation is successful. According to the relevant statistical tests, the performance of our framework appears to improve as we increase the number of samples per subject. Also, it appears to yield increased match and nonmatch scores separation for each probe. As illustrated, a gallery which comprises multiple samples per subject yields increased performance. However, our framework utilizes the additional information more effectively than raw score normalization methods. Finally, the experiments demonstrated that even for probes of fixed pose the proposed scheme has the potential to increase the recognition performance.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Evangelopoulos, O. Ocequeda, Prof. M. Papadakis, Prof. P. Tsiamyrtzis, and Dr. X. Zhao for fruitful conversations, and the anonymous reviewers for their comments that helped to improve this paper. 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] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," in *Proc. Speaker Lang. Recognit. Workshop*, Toledo, Spain, May 2004, pp. 369–376.

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

[3] J. F. G. Shakhnarovich and T. Darrell, "Face recognition from long-term observations," in *Proc. Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, May 2002, pp. 851–865.

[4] P. Moutafis and I. A. Kakadiaris, "Can we do better in unimodal biometric systems? A novel rank-based score normalization framework for multi-sample galleries," in *Proc. 6th IARP Int. Conf. Biometr.*, Madrid, Spain, Jun. 2013, pp. 1–8.

[5] P. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 947–954.

[6] P. J. Phillips *et al.*, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 831–846, 2010.

[7] G. Toderici, G. Evangelopoulos, T. Fang, T. Theoharis, and I. Kakadiaris, "UHDB11 database for 3D-2D face recognition," in *Proc. Pac.-Rim Symp. Image Video Technol.*, Guanajuato, Mexico, Oct. 2013, pp. 1–14.

[8] W. Scheirer, A. Rocha, R. Micheals, and T. Boult, "Robust fusion: Extreme value theory for recognition score normalization," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 481–495.

[9] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[10] E. Wolfstetter, U. Dulleck, R. Inderst, P. Kuhbier, and M. Lands-Berger, *Stochastic Dominance: Theory and Applications*. Berlin, Germany: Humboldt Univ. Berlin, School Bus. Econ., 1993.
[11] S. Durlauf *et al.*, *The New Palgrave Dictionary of Economics*. New York, NY, USA: Palgrave Macmillan, 2008.
[12] M. Birnbaum, J. Patton, and M. Lott, "Evidence against rank-dependent utility theories: Tests of cumulative independence, interval independence, stochastic dominance, and transitivity," *Organ. Behav. Human Decis. Process.*, vol. 77, no. 1, pp. 44–83, 1999.
[13] J. Hadar and W. Russell, "Rules for ordering uncertain prospects," *Amer. Econ. Rev.*, vol. 59, no. 1, pp. 25–34, 1969.
[14] G. Toderici *et al.*, "Bidirectional relighting for 3D-aided 2D face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2721–2728.
[15] O. Ocegueda, G. Passalis, T. Theoharis, S. Shah, and I. Kakadiaris, "UR3D-C: Linear dimensionality reduction for efficient 3D face recognition," in *Proc. Int. Joint Conf. Biometr.*, Washington, DC, USA, Oct. 2011, pp. 1–6.
[16] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Proc. 1st COST 2101 Workshop Biometr. Identity Manage.*, Roskilde, Denmark, May 2008, pp. 47–56.
[17] (Sep. 14, 2006). *National Science and Technology Council Subcommittee on Biometrics, Biometrics Glossary*. [Online]. Available: http://biometrics.gov/Documents/Glossary.pdf
[18] D. Garcia-Romero, J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "U-NORM likelihood normalization in PIN-based speaker verification systems," in *Proc. Audio Video-Based Biometr. Person Authentication*, Guildford, U.K., Jun. 2003, pp. 208–213.
[19] K. Toh, X. Jiang, and W. Yau, "Exploiting global and local decisions for multimodal biometrics verification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3059–3072, Oct. 2004.
[20] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Exploiting general knowledge in user-dependent fusion strategies for multimodal biometric verification," in *Proc. Int. Conf. Acoust. Speech. Signal Process.*, vol. 5. Montreal, QC, Canada, May 2004, pp. 617–620.
[21] M. Hollander, D. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*. New York, NY, USA: Wiley, 2013.
[22] M. Vatsa, R. Singh, and A. Noore, "Improving iris recognition performance using segmentation, quality enhancement, match score fusion, and indexing," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 1021–1035, Aug. 2008.
[23] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995.
[24] K. Nandakumar, A. Jain, and A. Ross, "Fusion in multibiometric identification systems: What about the missing data?" in *Proc. 3rd Int. Conf. Biometr.*, Alghero, Italy, Jun. 2009, pp. 743–752.

Panagiotis Moutafis (S'13) received the B.Sc. degree in statistics from the Athens University of Economics and Business, Athens, Greece, in 2010. He is currently pursuing the Ph.D. degree in computer science from the University of Houston (UH), Houston, TX, USA.

Since 2011, he has been a Research Assistant in biometric applications at the Computational Biomedicine Laboratory, UH. His current research interests include statistical learning, score normalization, cross-domain matching, and semi-supervised learning.

Mr. Moutafis was the recipient of the 2014 Hellenic Professional Society of Texas Scholarship Award.

Ioannis A. Kakadiaris (SM'91) received the B.Sc. degree in physics from the University of Athens, Athens, Greece, the M.Sc. degree in computer science from Northeastern University, Boston, MA, USA, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, PA, USA.

He was a Post-Doctoral Fellow at the University of Pennsylvania. In 1997, he joined the University of Houston (UH), Houston, TX, USA, where he is currently a Hugh Roy and Lillie Cranz Cullen University Professor of Computer Science, Electrical and Computer Engineering, and Biomedical Engineering. He is the Founder of the Computational Biomedicine Laboratory, UH. In 2008, he was the Director of the Methodist-University of Houston-Weill Cornell Medical College, Institute for Biomedical Imaging Sciences, Houston, Texas, USA. His current research interests include biomedical image computing, computer vision, pattern recognition with application on cardiovascular informatics, cancer informatics, neuro-informatics, face recognition, nonverbal human behavior understanding, and energy informatics. His research has been featured on The Discovery Channel, National Public Radio, KPRC NBC News, KTRH ABC News, and KHOU CBS News.

Dr. Kakadiaris was the recipient of a number of awards, including the National Science Foundation Early Career Development Award, the Schlumberger Technical Foundation Award, the UH Computer Science Research Excellence Award, the UH Enron Teaching Excellence Award, and the James Muller Vulnerable Plaque Young Investigator Prize. He was the General Co-Chair of the 2013 Biometrics: Theory, Applications and Systems Conference, and the 2014 SPIE Biometric and Surveillance Technology for Human and Activity Identification, the Program Co-Chair of the 2015 International Conference on Automatic Face and Gesture Recognition Conference, and the Vice-President for Technical Activities for the IEEE Biometrics Council.