Score Normalization in Stratified Biometric Systems

Sergey Tulyakov Nishant Sankaran Srirangaraj Setlur
Venu Govindaraju
Center for Unified Biometrics and Sensors
University at Buffalo, NY, USA
tulyakov,ns6,setlur,govind@buffalo.edu

Abstract

Stratified biometric system can be defined as a system in which the subjects, their templates or matching scores can be separated into two or more categories, or strata, and the matching decisions can be made separately for each stratum. In this paper we investigate the properties of the stratified biometric system and, in particular, possible strata creation strategies, score normalization and acceptance decisions, expected performance improvements due to stratification. We perform our experiments on face recognition matching scores from IARPA Janus CS2 dataset.

1. Introduction

A single biometric system typically contains biometric templates of different origin. For example, enrolled persons possess certain demographic characteristics (gender, age, race), which result in different physical appearances and consequently, invariability of templates. As another example, the biometric template data can be collected by different sensors or in different environments, which could also affect the collected templates. Finally, the template feature vectors or comparison scores could be calculated by different algorithms, e.g. trained to better deal with the specific nature of sensor data.

In many of these situations encountered by biometric systems, the data describing the template origin and characteristics, the *metadata*, may be available to the system engineers, and can be leveraged for system performance improvements. The metadata typically represents additional information separate from the template feature vectors - whereas template feature vectors are customarily derived using only sensor scanned data, the metadata are obtained either by user's input, or as sensor or biometric system parameters. Thus, integration of metadata into a biometric system is a sensible approach to achieve better performance.

In this paper we will define the term *stratum* to designate a particular subset of biometric templates. For example, we can define the male/female strata, as enrolled templates corresponding to males and females. As another example, we will use face image yaw based strata, where the stratum, frontal or profile, is defined by the angle of face yaw in the image. We will also use the term *stratum* to designate the comparison scores obtained from templates of particular stratum. In general, we will assume that there is only a finite number of strata defined for a particular biometric system and such strata include all system templates, or comparison scores.

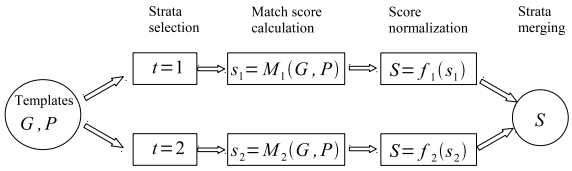

Figure 1. Operating scenario of our system.

The general approach for integrating the strata information used in our paper, is illustrated in Fig. 1. G and P denote gallery and probe biometric templates. In order to calculate a biometric comparison score, we first determine the stratum for these templates. Then, we perform stratum specific comparison score calculations, and/or stratum specific score normalization. Finally, the normalized score is output by the system and effectively, scores from different strata are merged to make a single match decision.

Note, that generally we can define strata for both gallery and probe templates in our scenario. For example, both gallery and probe face images can be frontal or profile, and in combination we can have four strata of comparison scores defined for two strata of gallery and two strata of probe templates. Also, in general, considering more strata has the potential of resulting in bigger improvements, but we would have to make sure that sufficient data is available for each considered stratum to train matching or score normalization algorithms.

In our current work, we look at scenarios with only two comparison score strata, defined by the corresponding two

strata of either gallery or probe templates. In some situations, this is the only possible scenario; for example, we might have information about the gender of the enrolled person (gallery template), but not about the probe template. We will also restrict ourselves to investigating strata specific score normalizations, and will assume that the matching algorithm is the same for both strata.

2. Previous Work

A number of previously published works tried to analyze the biometric system performance on different categories of templates. Klare *et al*. [10] analyzed the performance of six face recognition algorithms with respect to gender, race and age. Even larger number of metadata, or covariates, are considered in [2, 5]. However, demographic information was not used to improve the performance of the whole system in these works, although separate training for different strata of users was suggested. O'Toole *et al*. [13] considered the condition of matching demographic fields for gallery probe templates, but no additional score processing was performed. In contrast, we focus on strata defined by a single, either gallery or probe, metadata attribute.

The main effort of our paper is on the development of score normalization methods. Typically, the score normalization in biometric systems is performed before fusion of scores from multiple matchers [7, 12]. Correspondingly, the effectiveness of score normalization methods is judged by the performance of the final fused biometric system. A single score normalization function is trained for each matcher with the objective of transformed scores falling into some range or having particular statistical distribution parameters, and simple aggregation functions, such as sum, min, max, etc., are used for fusion. Such approaches might not give the objective evaluation of score normalization methods, since one can argue, that a proper fusion function, non-parametric and having universal approximation properties, would account for non-normalized scores. Since in our approach, there is no additional score processing after score normalization, the improvements in system performance serve as a more objective measure of the effectiveness of different score normalization methods.

The benefits of the normalization are limited if all the matching scores are subjected to the same normalization function. For example, min-max normalization will not change the ROC curve or the order of scores during identification trials. But, if we allow the normalization method to change for different sets of comparison scores, then we might expect to see the performance improvements for the system represented by such normalized scores. One way of achieving this is to try to learn user specific parameters of score normalization or fusion algorithms [8, 17]. Since the number of genuine match samples for each user is usually small, only parametric learning methods could be used in such scenarios. A more straightforward approach is to not to use the user specific genuine scores at all, and rely on a set of user specific impostor scores exclusively. T-normalization [1] and more general methods [15, 19] use parameters derived from a set of comparison scores related to a particular gallery or probe template. In all of the above methods the scores are transformed by differently trained functions, and the performance of the system changes without further fusion application. But the usual drawback is the limited number of training scores; only a set of corresponding impostor scores produced by a given gallery or probe template might be used for derivation of score normalization function. In this paper, we consider larger sets of scores available in strata for training, and derive score normalization functions using both impostor and genuine samples.

Some of our strata are defined by template quality based metadata parameters. The existing methods on incorporating template quality into biometric system [18, 9, 6] usually utilize fusion based approach as in section 3.1. In this paper, we explore alternative methods based on explicit stratification of templates and comparison score sets.

Poh and Kittler [14] considered separating matching scores into groups based on clusters of template quality measures. Such cluster based splitting is analogous to our stratification based on quality measures. In [14] generative approaches, i.e. likelihood ratio of section 3.3, decrease the baseline performance, and, as a result, only discriminative approach, i.e. fusion method of section 3.1, is used in experiments. In contrast to this paper, our strata contain more samples available for training and generative approach to score normalization, i.e. likelihood ratio, does improve baseline performance. Due to larger training genuine sample size, we were also able to consider score normalization methods based on strata ROC data, and, in particular, propose a new method of cost based score normalization of section 3.4.

3. Stratification based approaches to score normalization

The score normalization can be defined as a transformation of the original matching score s into some normalized score S: $S = F(s)$. The normalization function F is typically learned from some training data. The usual goal of normalization is to obtain a matching score with predefined distribution parameters. For example, min-max normalization linearly scales the score to the interval $[0, 1]$; the minimum and maximum score values used in the algorithm can be derived from training set. We use traditional z-score normalization method as baseline [7]:

$$F(s) = \frac{s - \mu_i}{\sigma_i} \tag{1}$$

where the normalization parameters μ_i and σ_i are derived from the training set for stratum i. Even though we use all available scores in training sets, these parameters are mostly defined by the impostor score majority.

3.1. Fusion of Strata Information

Let us associate strata with some numeric values t. Then the score normalization can take a form:

$$S = F(s, t) \tag{2}$$

where s is the original matching score, S is the normalized matching score, and F is some function, fusing score and strata identifier. Function F can be trained using specified optimization criterion.

This approach is quite general and easily adaptable to many problems, and have been used in many papers trying to combine some template auxiliary data and matching scores (e.g. [14]). Moreover, it is applicable not only to the problems with discretely defined strata, but also to problems with a continuously varying parameter t. For example, t can just represent the numeric value of a person's age avoiding the need for separate strata based on age ranges. As another example, instead of defining some quality strata, such as "'good"', "'normal"' or "'bad"', we can directly utilize a numeric value representing the quality of templates. However, this approach has disadvantages as well. First, the strata identifiers might not have a natural numeric representation (e.g. person's race), and attempting to introduce such a representation would imply that there is some order in strata (when there is no inherent order with a characteristic such as race). Second, we might introduce too much complexity into the fusion function, which will be difficult to train. Considering separate strata can thus help avoid difficulties in training. Finally, it might be difficult to analyze the constructed fusion function and guarantee its optimality.

3.2. Error rate based normalization

Suppose we consider a biometric verification system and our decision criteria is based on the comparison of matching score with some threshold $i.e.$ if matching score s belongs to stratum i, then we accept if $s \geq \theta_i$, and reject otherwise. If we perform score normalization in each stratum separately and aggregate the normalized scores, then we would base our decision on comparing the transformed scores with some threshold $S \geq \theta$. Thus, the stratum specific thresholds θ_i are mapped to a single system threshold θ by the normalization functions. Consequently, one of the considerations for constructing stratum specific normalization functions would be to ensure that the thresholds θ_i mapped to the same value θ, have similar properties.

One way of doing this would be to look at the error rates for each stratum associated with these thresholds:

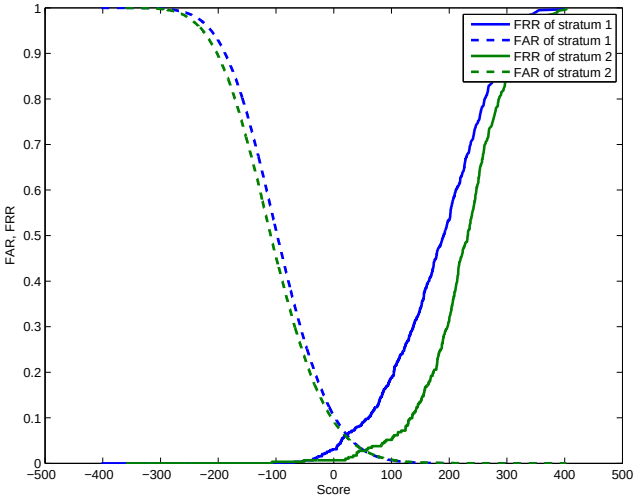

Figure 2. FAR and FRR of the two face yaw quality based strata.

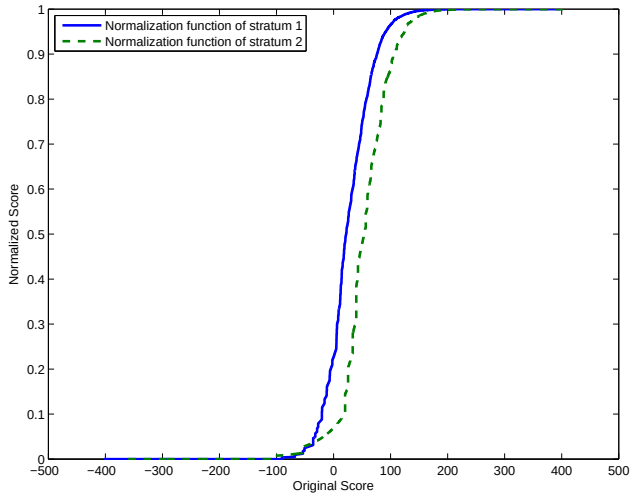

Figure 3. Score normalization functions for the two face yaw quality based strata.

$FRR_i(\theta_i)$ and $FAR_i(\theta_i)$, and have our score normalization function include these error rates in some manner. For example, we can simply define the score normalization to depend only on false accept rate $S = 1 - FAR_i(s)$. Such normalization needs only the samples of impostor scores for construction, and can be a sensible approach in biometric systems, where the number of genuine score samples might be small. But, in order to incorporate both impostor and genuine score distributions, we can propose a ratio based approach to combine error rates:

$$S = \frac{FRR_i(s)}{FAR_i(s) + FRR_i(s)} \tag{3}$$

Such a score normalization function appears to improve performance in our experiments. The advantage of this function is its monotonicity and ease of construction. Note

that there exist other similar published approaches to score normalization(e.g. [15]). However, it might not be the optimal, i.e. leading to the minima of error rates in the final system. We consider theoretically optimal score normalization functions in the next section.

3.3. Likelihood Ratio

Suppose we consider a biometric verification system, whose performance is determined by the trade-off between false accept and false reject rates. Let the prior probability for the samples of stratum i be P_i. Then the total cost of operating the system is

$$Cost = \sum_i P_i Cost_i \tag{4}$$

and the cost of operating system on stratum i is

$$Cost_i = C_{FR} * P_{i,gen} * FRR_i(\theta_i) + C_{FA} * P_{i,imp} * FAR_i(\theta_i) \tag{5}$$

where $P_{i,gen}$ is the prior probability of genuine samples in stratum i, $P_{i,imp}$ is the prior probability of impostor samples in stratum i, C_{FR} is the cost of false rejection of genuine samples and C_{FA} is the cost of false acceptance of impostor samples.

The optimal decisions to accept or reject samples minimizing the cost in the above equation are defined by the likelihood ratios of genuine and impostor scores [16]:

$$lr_i(s) = \frac{p_{i,gen}(s)}{p_{i,imp}(s)} \lessgtr \frac{C_{FA} * P_{i,imp}}{C_{FR} * P_{i,gen}} \tag{6}$$

Here $p_{i,gen}$ is the density of genuine scores in stratum i, $p_{i,imp}$ is the density of impostor scores in stratum i. Effectively, θ_i is such that the Eq. 6 becomes equality if $s = \theta_i$. Note, that this optimal acceptance decision is made separately for each stratum; since the total cost of the biometric system is the summation of costs for each stratum (Eq. 4), this acceptance decision optimizes the cost for each stratum and, consequently, for the whole system.

We can convert the decisions of Eq. 6 into the following optimal stratum score normalization function

$$F(s,i) = \frac{p_{i,gen}(s)}{p_{i,imp}(s)} \times \frac{P_{i,gen}}{P_{i,imp}} \tag{7}$$

After such normalization, the scores from different strata will be compared to the same threshold $\frac{C_{FR}}{C_{FA}}$ to achieve the optimal verification decision.

If we assume that the ratios of prior class probabilities are the same for different strata:

$$\frac{P_{i,gen}}{P_{i,imp}} = \frac{P_{j,gen}}{P_{j,imp}} \tag{8}$$

then our normalization is simply the likelihood ratio:

$$F(s,i) = \frac{p_{i,gen}(s)}{p_{i,imp}(s)} \tag{9}$$

Even though the likelihood ratio score normalization of Eq. 7 is the theoretically optimal normalization method for a stratified biometric system, it has some training related drawbacks. Indeed, the approximation of likelihoods can be a hard task, even in the one dimensional score space considered here. Since our goal is to train a single score normalization function per stratum, this method requires approximating two density functions and might result in poor approximation of the normalization function. Another concern is the possible non-monotonicity of the approximated normalization function. For example, if we use mixtures of Gaussians or Parzen windows method for approximating likelihoods, then their ratio in Eq. 7 will most surely be non-motonic. The non-monotonic score normalization function might be not optimal for many biometric matchers, whose scores represent either distances between templates or similarities, and whose monotonic nature is implied during training.

3.4. Cost based normalization

In order to avoid the difficulties associated with the approximation of likelihoods in Eq. 7, we can try to use error rate functions $FAR_i(s)$ and $FRR_i(s)$ directly for the approximation of the score normalization function. The previous section suggests that normalization function $F(s,i)$ will be optimal if for threshold θ_i optimizing the total stratum cost for particular costs of false rejects and false accepts, C_{FR} and C_{FA}, we will have the mapping

$$F(\theta_i, i) = \frac{C_{FR}}{C_{FA}} \tag{10}$$

Given $FAR_i(s)$ and $FRR_i(s)$ and particular C_{FR} and C_{FA}, we can simply iterate over all possible threshold values θ_i, and find which value optimizes stratum performance. This step can be repeated for the range of values of C_{FR} and C_{FA}; in our experiments we considered iterations with small step over the range: $0 \leq C_{FR} \leq 1$, $0 \leq C_{FA} \leq 1$ and $C_{FR} + C_{FA} = 1$. In addition, instead of map of Eq. 10, we consider an equivalent normalization function of Eq. 11:

$$F(\theta_i, i) = \frac{C_{FA}}{C_{FA} + C_{FR}} \tag{11}$$

This function has range of values between 0 and 1, and maps more confident scores closer to 1 (where false accept costs are higher than false reject costs).

However, there is a problem with the above definition of the score normalization function. It is possible that same values of threshold θ_i will optimize the total stratum cost for different ratios of $\frac{C_{FR}}{C_{FA}}$, and the function of Eq. 11 will not be well defined. In fact, this is a typical situation since $FAR_i(s)$ and $FRR_i(s)$ are approximated as step functions from training data, and the calculated cost is a step function of threshold as well. Moreover, due to variations in

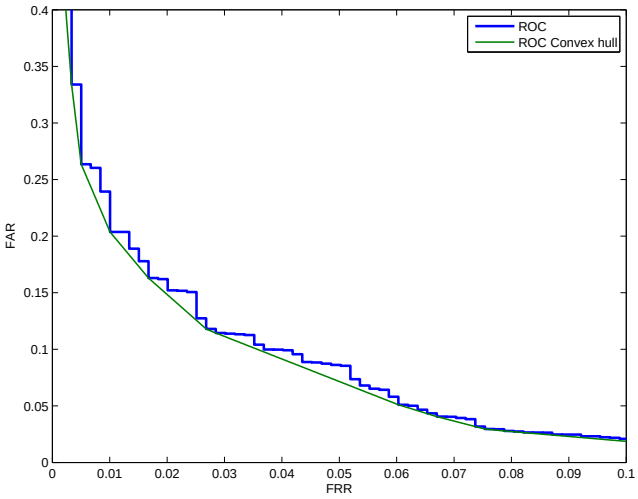

Figure 4. ROC curve and its convex hull; vertices of convex hull determine optimal thresholds for cost minimization.

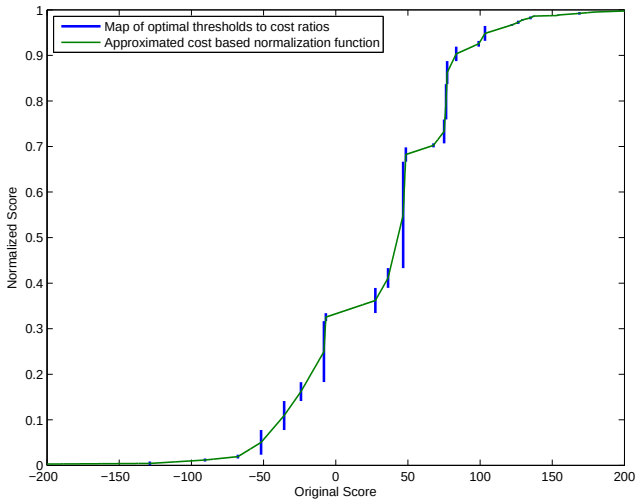

Figure 5. Map of optimal threshold values to cost ratios and corresponding approximation of the score normalization function.

the training score sample set distribution, the ROC curves reconstructed with the help of these samples have inherent concavities, and the optimal thresholds determined for different values of cost ratios do not cover the whole range of scores, but rather belong to the discrete set of convex hull vertices. Figure 4 presents an example from our experiments - we can see that the convex hull of ROC curve has only a small number of vertices, and the cost minimization method searching for tangent lines in ROC curve will find the optimal thresholds located exclusively at these vertices of convex hull.

To further illustrate this situation, Fig. 5 displays a mapping from the found discrete set of thresholds to the values of cost ratio (Eq. 11); a single threshold value will correspond to the interval of cost ratios, i.e. the vertical line interval in the graph. In order to create a valid score normalization function, which should be defined for all possible scores and have single values, we need to apply some kind of smoothing to that map. In our experiments, we use simple linear interpolations to connect the middle points of consecutive vertical intervals; the tails of the function are defined as the tails of the exponents approaching either 0 or 1. An example of such a smoothing is given in Fig. 5.

4. Estimating performance improvements

Even though the template metadata may be readily available, allowing the stratification of matching scores and the stratum specific normalization according to the methods of previous section, the performance improvement is not guaranteed. For example, if scores are already normalized according to strata, then repeated normalization will fail. Thus, it might be helpful to know the expected performance improvement from stratification $apriori$, i.e. given only training data.

We can modify the strata score normalization method of section 3.4 to derive a method of estimating the performance improvements due to stratified score normalization. Suppose, for a particular cost ratio $r = \frac{C_{FR}}{C_{FA}}$, we find thresholds $\theta_i(r)$ minimizing the total cost $Cost_i$ for each of two strata i (Eq. 5). Let us denote $Cost'_i$ as the system's cost on stratum i, if threshold for a different stratum, θ_{1-i}, is used instead of the optimal threshold θ_i. The difference $Cost'_i - Cost_i$ will be the additional cost if a non-optimal threshold is used for strata i. We integrate this additional cost over the range of possible cost ratios r to get the estimate of the possible performance improvement due to the stratification method:

$$D(i) = \int_r (Cost'_i(r) - Cost_i(r))p(\theta_i(r))dr \qquad (12)$$

Here, the integration is weighted by the average of genuine and impostor score sample distributions: $p(\theta) = p_{i,gen}(\theta) + p_{i,imp}(\theta)$. Equation 12 estimates the performance improvement for stratum i; we repeat the calculations for other strata and average the results to get the final average estimate of performance improvement.

5. Experiments

We perform our experiments on the set of comparison scores of face recognition algorithm by Chen *et al*. [4]. The algorithm extracts face features from images using a deep convolutional neural network. The joint Bayesian metric learning method is used to derive the face recognition scores between template feature vectors. We conduct our experiments on face recognition scores obtained on IARPA Janus Challenge Set 2 (CS2) dataset, which is a superset of IJB-A dataset [11]; the comparison between CS2 and IJB-A sets is given in [4].

Strata method	Stratum 1	Stratum 2	Estimated D	Actual Improvement
1	88.01 ± 3.60	87.41 ± 2.00	0.11 ± 0.05	-0.03
2	89.00 ± 2.38	86.70 ± 2.27	0.10 ± 0.04	0.06
3	87.39 ± 2.58	87.53 ± 2.62	0.11 ± 0.05	0.03
4	83.97 ± 2.25	94.00 ± 1.54	0.29 ± 0.14	0.41
5	82.87 ± 2.26	93.76 ± 1.52	0.37 ± 0.19	0.72

Table 1. Characteristics of the considered stratification methods: performance of individual strata (% TAR at FAR=1%), estimated performance improvement using only training set (section 4) and actual improvement (% TAR at FAR=1%) of the best performing method measured on test set.

Although CS2 dataset has 10 splits and there is a separate image subset designated for training in each of those splits, the training set does not have a separation into gallery and probe subsets, and, as a result, we do not have proper comparison score training sets for our experiments. Thus, in order to have proper training and testing comparison score subsets, we use bootstrap testing technique [3]. In each step of the bootstrap, the gallery and probe templates are randomly divided into training and testing parts (the division is performed with respect to identities, and since each person can be used to derive multiple templates, the numbers of templates in each part is variable). The training gallery and probe sets are used for creating training comparison score set, and the same is done for the testing set. In total, each testing and training set has either 83 or 84 gallery templates, and around 900 probe templates; correspondingly we have around 900 genuine and 73,000 impostor scores. We perform 10 bootstrap experiments for every split, and, thus, the total number of experiments is 100. The results for these 100 experiments are averaged to obtain a mean performance value (EER) and its standard deviation.

We explore five stratification methods (three based on demographics and two on template quality)in our experiments based on the available metadata in the CS2 dataset:

1. Female/male strata (0=female, 1=male).
2. Age strata (0=young, 1=old).
3. Skin color strata (0=light, 1=dark).
4. Face yaw strata (0=side view, 1=frontal).
5. Number of images based strata (0=small, 1=large).

The templates in this dataset consist of variable numbers of images, and the metadata information is provided for separate images. Thus, we average the image metadata to obtain the stratum number. All metadata except gender have numeric values with particular ranges; after averaging these numeric values, we choose a threshold separating two strata so that the number of templates in both strata would be similar. In the first three methods, the stratification is done with respect to gallery templates assuming that the demographic metadata is more likely to be available for the gallery templates. In the last two methods, we perform stratification with respect to probe templates; the gallery templates in CS2 dataset have more images resulting in more uniform quality, and subsequently, a small difference in strata. Note, that we do not perform a match between demographic data of gallery and probe templates, but only split the data based on the value of demographic data of a single gallery template. Matching demographic data would most probably further improve the performance, but this is not a focus of this paper.

Table 1 provides the performance characteristics of individual strata, the estimated performance improvement of the system due to stratification and observed actual performance improvements from the best score normalization approach (relative to baseline 87.79% TAR). Although the difference between strata performance is present in all cases, the estimate of possible performance improvement reveals that the benefits will be rather modest. This is confirmed by the reported performance of presented score normalization methods on test set in Table 2.

In general, the two optimal methods viz. likelihood ratio and cost based normalization, perform slightly worse than fusion based normalization, but as we noted, they might be superior for strata defined by non-numeric attributes. The cost based normalization seems to perform consistently better than likelihood ratio, which is explained by the monotonicity of normalization function. Finally, all presented methods seem to perform better than traditional Z-score normalization.

6. Conclusions

In this paper we have achieved the following goals:

- We have presented a general framework for creating a stratified biometric system.
- We have demonstrated how score normalization methods can be evaluated in such a system.
- We have derived four score normalization methods and discussed their optimality. The last method, cost based

Strata method	Z-score	Fusion	FAR/FRR	LR	Cost
1	$-.21 \pm .39$	$-.03 \pm .50$	$-.03 \pm .41$	$-.16 \pm .60$	$-.08 \pm .49$
2	$-.04 \pm .32$	$.02 \pm .43$	$.04 \pm .40$	$-.02 \pm .47$	$.06 \pm .48$
3	$.03 \pm .27$	$-.18 \pm .57$	$-.05 \pm .44$	$-.16 \pm .55$	$-.07 \pm .42$
4	$.21 \pm .26$	$.41 \pm .45$	$.41 \pm .54$	$.28 \pm .51$	$.40 \pm .54$
5	$.41 \pm .34$	$.72 \pm .53$	$.68 \pm .57$	$.50 \pm .56$	$.68 \pm .57$

Table 2. Changes in TAR(%) & FAR=1% for different strata selection and score normalization methods.

score normalization, appears to be a new, well performing, method of score normalization.

- We have derived a method to estimate the magnitude of possible performance improvements due to stratification.

Even though the observed performance improvements were not significant for all considered stratification methods, the derived measure of estimated performance improvement explains this result. Note that this does not measure the difference in performance of different strata, but rather the difference in optimal thresholds, and thus directly indicates the magnitude of possible performance improvement.

The proposed score normalization method of section 3.4 could be applied not only to stratified, but to arbitrary biometric systems. Since the meaning of normalized scores in terms of error costs is well defined, it would be easy for system users and administrators to set the threshold parameters, or integrate them with other security devices.

Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[2] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762, 2009.

[3] R. M. Bolle, N. K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems–the subsets bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33, 2004.

[4] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep CNN features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[5] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. Draper, Y. M. Lui, and D. Bolme. Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics & Data Analysis*, 67:236–247, 2013.

[6] F. Hua, P. Johnson, and S. Schuckers. Utilizing automatic quality selection scheme for multi-modal biometric fusion. In *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*, pages 664–670, 2013.

[7] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recogn.*, 38(12):2270–2285, 2005.

[8] A. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *International Conference on Image Processing. 2002*, volume 1, pages I–57–I–60 vol.1, 2002. TY - CONF.

[9] N. D. Kalka, J. Zuo, N. A. Schmid, and B. Cukic. Estimating and fusing quality factors for iris biometric images. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(3):509–524, 2010.

[10] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[11] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[12] C.-L. Liu. Classifier combination based on confidence transformation. *Pattern Recognition*, 38(1):11–28, 2005.

[13] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–176, 2012.

[14] N. Poh and J. Kittler. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):3–18, 2011.

[15] V. Štruc, J. Ž. Gros, and N. Pavešić. Non-parametric score normalization for biometric verification systems. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2395–2399, 2012.

[16] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[17] K. A. Toh, J. Xudong, and Y. Wei-Yun. Exploiting global and local decisions for multimodal biometrics verification. *IEEE Transactions on Signal Processing*, 52(10):3059–3072, 2004.

[18] K. A. Toh, W. Y. Yau, E. Lim, L. Chen, and C. H. Ng. Fusion of auxiliary information for multi-modal biometrics authentication. In D. Zhang and A. K. Jain, editors, *Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004. Proceedings*, pages 678–685. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[19] S. Tulyakov, J. Li, and V. Govindraju. Enrolled template specific decisions and combinations in verification systems. In *IEEE Second International Conference on Biometrics: Theory, Applications and Systems (BTAS 08)*, 2008.