Combination of User- and Enrollee-specific Statistical Information in Verification Systems

Xi Cheng, Sergey Tulyakov, and Venu Govindaraju
Center for Unified Biometrics and Sensors
University at Buffalo, NY, USA
xicheng,tulyakov,govind@buffalo.edu

Abstract

Instead of using matching scores between single enrolled and single test (or user) template, the matching scores related to all test templates or all enrolled ones can be considered to enhance the performance of biometric systems. The user-specific methods take into account the dependencies of matching scores assigned to different enrollees being matched to one test template. On the other hand, enrolled template specific methods consider the relationship among matching scores between different user inputs and one enrolled template. In this paper, we consider the combination of user and enrollee specific statistical information by utilizing various statistical models. The experiments show that the combination of user and enrollee specific methods can further improve the performance of both unimodal and multimodal biometric systems compared to solely using either user or enrollee specific models.

1. Introduction

Matching score in biometric systems is the measurement of distance between an enrolled template T_e and a test template T_t. T_e which is enrolled beforehand is also called gallery template and T_t which is tested is also called input or probe template. The matching scores can be used to make accept and reject decisions or to combine different biometric traits or matchers in complex biometric systems.

It is expected that matching scores are influenced by the quality of enrolled or test templates. For instance, if the quality of a test face image is low, the genuine matching score between this input and the enrolled face will be low. If the same test template is matched with some impostor faces in the database, the matching scores will probably be lower than the typical impostor scores. On the other hand, if the quality of test template is good, all the scores generated from both genuine and impostor matchings should be relatively higher than the ones in the previous case. These cases are also true for the situation of one enrolled template with multiple test templates. So the quality of test or enrolled templates should be reflected by using a set of matching scores. The quality is defined as *user quality* to measure the confidence of user biometric sample with its own templates in [7]. It shows significant improvement in the performance for hand-based biometric systems using *user quality* during matching process. Instead of using genuine matching scores to estimate the quality of the bometric template, our methods use all scores associated with the biometric input from both user- and enrollee-specific matchings and large databases to evaluate the performance.

In this paper, we will investigate different models to improve the performance of both unimodal and multimodal verification systems. In unimodal system, scores are used to make decisions about acceptance or rejection and for the latter one, matching scores are used to combine different biometric modalities first. For both cases, instead of using a single score between enrolled and test templates, we use a set of matching scores for one test template with multiple enrolled templates and vice versa. Rather than generating a template quality directly from matching scores in [9], we want to know which information of the scores is most useful.

Fig. 1 shows the sets of possible matching scores produced when the biometric system with m enrolled persons used to identify n test templates. Same sets of scores are available in the biometric score database NIST BSSR [1] which is used in our experiments. Following the notations of this database, the *user* refers to a particular test or probe template. There is only one enrolled template for each person and a single input template for each user so for each column or row in Fig. 1 there is only one genuine score and the rest of them are impostor scores. This database contains two different face matchers and one fingerprint matcher for left and right index fingers.

In this paper, we investigate to use not only all scores related to an enrolled template (one column) but also the scores associated to one test template (one row) to esti-

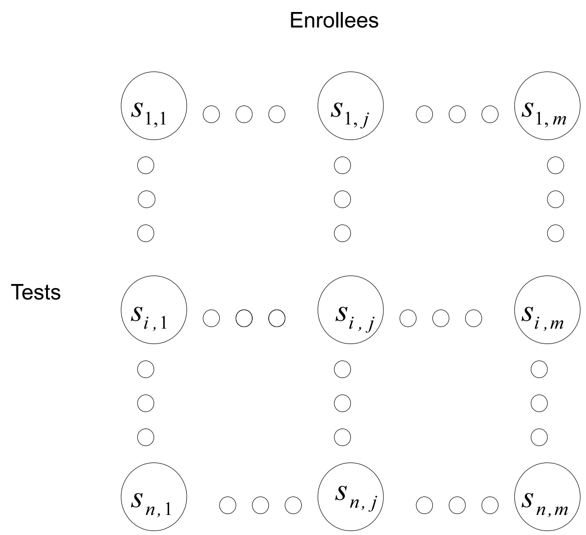


Figure 1. The matrix representation of matching scores produced during biometric system run: each of n test templates is matched against m enrolled templates. The column of scores is obtained using same enrolled template, and can be used to derive enrolled template specific score model. The row of scores is obtained using same test (user) template, and can be used to derive user specific model.

mate the quality of the test template which is on the cross of the column and the row. The methods considering the dependencies in scores of a particular enrolled person are called enrollee-specific methods and the methods accounting for the scores related to a particular input user are called user-specific methods [12]. Especially, we will consider the combination of T-normalization [3] and Z-normalization [8] as well as the combination of user- and enrollee-specific second best score models [12]. These models are also explained in the following section.

We will investigate the improvement of unimodal systems and multimodal systems using our methods. Though our theory can apply to more than two classifiers,we only consider combination of two matchers in this paper.

2. Previous Work

Verification systems are usually implied to have the only scores between one test template and one enrolled template. However, as mentioned above, two kinds of score dependencies (column-wise and row-wise scores in Fig. 1) can be used to improve the performance of both identification(1:N) and verification(1:1) biometric systems. Although in this paper we perform the biometric person verification only and measure the performance of the verification system, we collect additional matching scores as well, e.g. the scores between test template and all enrolled templates. This is the same set of scores which is available during identification system trial, and therefore we would call this set of scores as *identification trial score set*.

We first consider user-specific methods which use the scores obtained from the same test template i, $\{s_{i,1}, \ldots, s_{i,m}\}$ as in Fig. 1. There is dependency between these scores since the same test template is used for their derivation. One example of the creation of user-specific score model is the conversion of raw matching scores to ranks; the rank of the score is its order in the set of scores $\{s_{i,1}, \ldots, s_{i,m}\}$ obtained in the current test. Ho *et al*. [11] argue that the use of ranks is more powerful than the use of raw scores in the combination of character recognizers; thus, the user (or test) specific model of ranks carries more useful information than original scores with no user-specific model. Brunelli and Falavigna [5] successfully incorporate ranks along raw matching score in biometric identification problem.

T-normalization [6, 8] is one kind of user-specific model. The mean μ and standard deviation σ is calculated from the set of scores corresponding to a single identification trial and each score in the set is normalized by using the following formula:

$$s \rightarrow \frac{s-\mu}{\sigma} \tag{1}$$

As it shows, the T-normalization successfully accounts for score dependencies if the matching scores in different identification trials have only those variations.

Second best score is another user-specific model [13] which considers the current score and the best score besides the current one from the identification trial. The formula for second best score model is:

$$s \rightarrow \{s, second\_best\_score(s)\} \tag{2}$$

where $second\_best\_score(s)$ is the second best score (best score besides s) in one identification trial. The second best score model can be understood like this: for a currently considered score if second best score (best score besides current) is very high, this high score is likely to be genuine and currently considered score is therefore impostor. Reverse is also true: if second best score is low relative to current, then this score is very likely to be genuine.

Tulyakov *et al*. [13] compare both T-normalization and second best score model on biometric score combination algorithms and conclude that second best score model provides a better performance than T-normalization. They also claim that the performance is further improved if T-normalization and second best score model are used together.

Similar to above user-specific models, the background model also utilize the scores among one template and other templates. The difference between them is that the background model use selective enrolled templates according to one enrolled template instead of all enrolled templates.

Matcher	mean	variance	Best_imp_score
li	0.2917	0.3302	0.2903
ri	0.3423	0.3639	0.3349
C	0.2301	0.1798	0.2168
G	0.2861	0.2983	0.2453

Table 1. Correlations between genuine score and statistics of enrollee-wise score sets

Matcher	mean	variance	Best_imp_score
li	0.2857	0.3227	0.2901
ri	0.3541	0.3691	0.3269
C	0.1439	0.1263	0.1485
G	0.1587	-0.0352	0.1569

Table 2. Correlations between genuine score and statistics of user-wise score sets

The cohort based method which is one implementation of background model finds a cohort - a subset of enrolled templates close to a particular enrolled one under some considerations. The cohort of enrolled templates are used during the matching process. The cohort method has been used for fingerprint verification [2] and speaker verification [3].

On the other hand, the enrollee-specific statistical information of matching scores can also be used in biometric systems. The analogue of T-normalization here is the Z-normalization which has the same formula 1 as T-normalization but uses only enrollee-specific mean and variance. Particularly, the mean and variance are calculated according to each column in Fig. 1. The name of second best score model is still used in enrollee-specific methods and share the same formula 2 as the user-specific one. Tulyakov *et al*. [12] claim both Z-normalization and second best score model can enhance not only the verification decisions of single matcher but also the performance of combination of matchers in verification systems.

3. Combination Rules

We conducted experiments on NIST BSSR1 dataset in which 'li','ri','C' and 'G' are different matchers. Tables 1 and 2 show correlations between genuine score and statistics of enrollee- or user-wise score sets for each matcher. We can notice that generally all matchers exhibit relatively high dependencies so that accounting for these dependencies might benefit any system utilizing these matchers. In addition, the correlations for 'C' and 'G' are higher when the statistics are calculated from enrollee-specific scores than the statistics from user-specific sets. This implies that the matching scores for 'C' and 'G' is non-symmetric with regards to enrolled and test templates. That is why we want to use both user- and enrolle-specific statistical information compared to previous research which states that user- and enrollee-specific models can improve the performance of biometric systems individually. For unimodal verification systems which contain a single matcher, ROC curve is the best way to evaluate the performance of the system. But for multimodal biometric system, combination of matchers should be done first.

The combination algorithm we use here is the likelihood ratio method which is theoretically optimal combination method for verification systems [10]:

$$S = \frac{p_{gen}(s_1, s_2)}{p_{imp}(s_1, s_2)} \tag{3}$$

where s_i is the score from matcher i. Instead of using either T- or Z-normalization in [12, 13], we use both methods together with different order, T-normalization first then Z-normalization and vice versa. As we present in the next section, such combination with sequential normalizations by both methods in either order can indeed improve the performance results.

where $sbs(s_i)$ is calculated differently for user- and enrollee-specific methods as indicated before. In order to use both user- and enrollee-specific second best scores, six dimensional probability densities should be constructed for genuine and impostor scores:

$$S = \frac{p_{gen}(s_1, sbs_u(s_1), sbs_e(s_1), s_2, sbs_u(s_2), sbs_e(s_2))}{p_{imp}(s_1, sbs_u(s_1), sbs_e(s_1), s_2, sbs_u(s_2), sbs_e(s_2))} \tag{4}$$

where $sbs_u(s_i)$ is the user-specific second best score for matcher i and $sbs_e(s_i)$ is the enrollee-specific one for matcher i. We use Parzen window with Gaussian kernels to estimate the densities by using bootstrap technique.

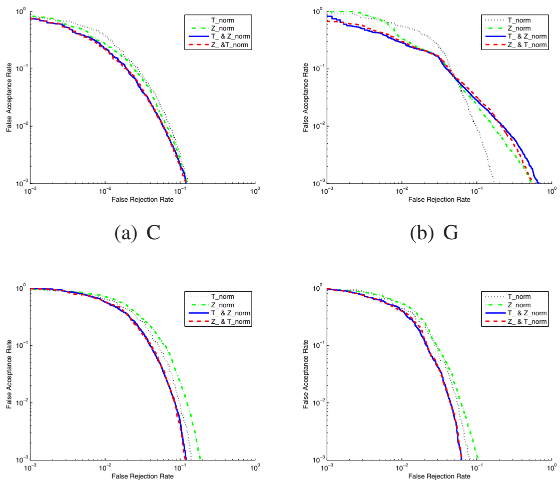
4. Experiments

The formula of second best score model for either user- or enrollee-specific situation is the same:

$$S = \frac{p_{gen}(s_1, sbs(s_1), s_2, sbs(s_2))}{p_{imp}(s_1, sbs(s_1), s_2, sbs(s_2))} \tag{5}$$

We use the NIST BSSR1 dataset which contains matching scores for one fingerprint matcher applied on right index 'ri' and left index 'li' fingers and two face matchers, 'C' and 'G'. Each dataset is tested as a unimodal system. Due to data collection errors(for example, all match scores for a particular user or enrollee being 0), some scores have to be discarded to get 2991 enrollees and 5982 inputs in each. For the purpose of testing multimodal systems, two of the datasets are combined as each combination is treated as a virtual person: 'C G', 'C li', 'C ri', 'G li', 'G ri'and 'li ri'.

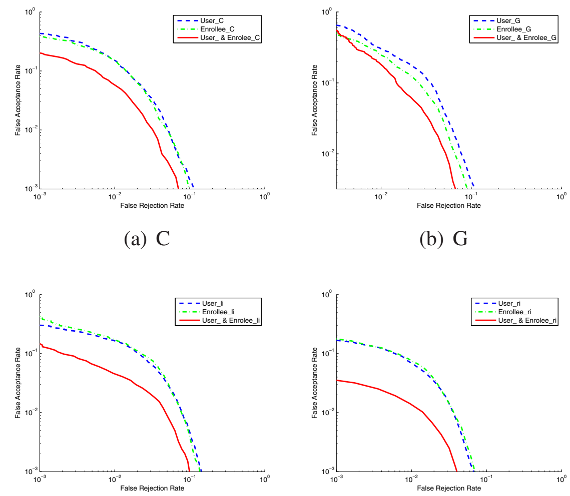
For both unimodal and multimodal systems, we use Parze



(a) C (b) G

(c) li (d) ri

Figure 2. ROC curves for T-normalization, Z-normalization, T- & Z-normalization and Z- & T-normalization of unimodal systems



(a) C (b) G

(c) li (d) ri

Figure 3. ROC curves for user-specific, enrollee-specific and user- & enrollee-specific second best score models of unimodal systems

mated by the maximum likelihood method. Bootstrap sample testing technique is used in our experiments [4]. We perform 100 bootstrap tests in each experiment. For each bootstrap test,out of total 5982 identification trials, 2991 of them are selected randomly as testing set, 1400 trials are chosen randomly as training set and remaining 1591 trials are left for validation which is used to estimate the kernel width.

4.1. Verification Systems with Single Matcher

In these experiments we use normalization and score statistics for enhancing the performance of unimodal verification systems. To accept or reject a verification trial, as in the traditional approach, we compare matching scores to some thresholds. The T-normalization and Z-normalization also use thresholds but the scores are T-normalized or Z-normalized first. The scores are T-normalized first then Z-normalized in T- & Z-normalization method and in Z- & T-normalization method scores are Z-normalized and then T-normalized. The results are shown in Fig. 2. The performance has been improved in both Z- & T-normalization and T- & Z-normalization compared to either Z-normalization or T-normalization in all cases except G. Tulyakov *et al*. [12] explain that Z-normalization fails for G because the linear normalization model does not reflect the score dependencies for the enrollee-specific score sets in G.

To use the user-specific and enrollee-specific second best score models, we perform the following transformation on the matching scores:

$$S = \frac{p_{gen}(s_1, sbs(s_1))}{p_{imp}(s_1, sbs(s_1))} \tag{6}$$

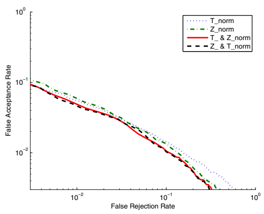
For the user-specific second best score model, the $sbs(s_1)$ is the second best score in one identification trial. On the other hand, the $sbs(s_1)$ is the second best score out of all matching scores corresponding to one enrollee in the enrollee-specific model. Eq. 6 is the optimal Bayesian classification decision by treating $(s_1, sbs(s_1))$ as inputs to the classification process. The user- and enrollee-specific second best score model which uses both user-specific and enrollee-specific second best scores has this formula:

$$S = \frac{p_{gen}(s_1, sbs_{user}(s_1), sbs_{enrollee}(s_1))}{p_{imp}(s_1, sbs_{user}(s_1), sbs_{enrollee}(s_1))} \tag{7}$$

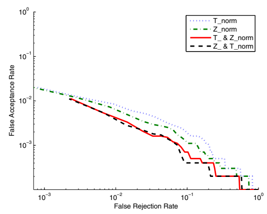
The three dimensional probability densities are estimated for genuine and impostor scores. Therefore, if sbs_{user} and $sbs_{enrollee}$ have any useful information complementary to the original score and the densities p_{gen} and p_{imp} are estimated sufficiently, the performance of single matchers can be improved compared to the methods using only user-specific or enrollee-specific statistical information of scores. The experiment results are shown in Fig. 3. As it shows, utilizing both user- and enrollee-specific second best scores can further improve the performance compared to using only either user- or enrollee-specific second best scores in all single matcher systems.

4.2. Verification Systems with Combination of Matchers

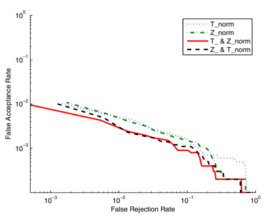
In this section we use normalizations and score statistics in the combination of biometric matchers. Though more than two matchers can be used, in this paper we only combine two matchers out of 'C' 'G' 'li' and 'ri' so six combinations are tested. The likelihood ratio of Eq. 3 is used in the



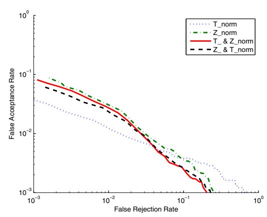
(a) C and G



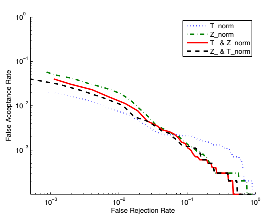
(b) C and li



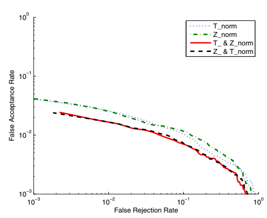
(c) C and ri



(d) G and li

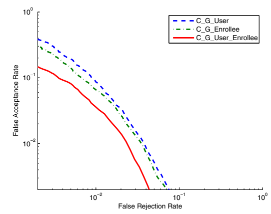


(e) G and ri

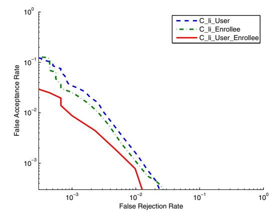


(f) li and ri

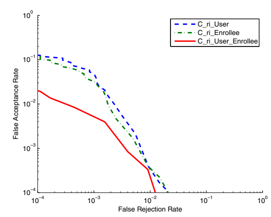
Figure 4. ROC curves for T-normalization, Z-normalization, T- & Z-normalization and Z- & T-normalization of multimodal systems



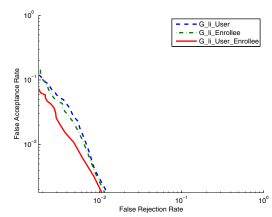
(a) C and G



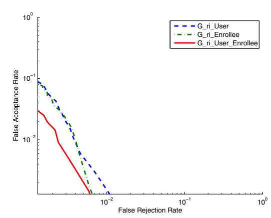
(b) C and li



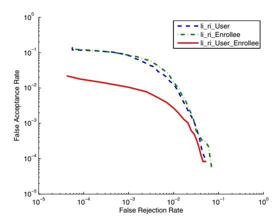
(c) C and ri



(d) G and li



(e) G and ri



(f) li and ri

Figure 5. ROC curves for user-specific, enrollee-specific and user- & enrollee-specific second best score models of multimodal systems

T-normalization , Z-normalization , T- & Z-normalization and Z- & T-normalization to make the acceptance or rejection decisions. Similar to single matcher cases, the difference among these methods is that scores are normalized in different ways or in different orders. The results are shown in Fig. 4. As it shows, both T- & Z-normalization and Z- & T-normalization perform better than either T-normalization or Z-normalization with the exception of combination involving matcher G in Fig. 4(d) and Fig. 4(e). The explanation would be the same as in previous section that the linear normalization model does not represent the score dependencies of the enrollee score sets in dataset G.

Eq. 5 is used for either user-specific or enrollee-specific second best score models. The difference between them is how the second best score is retrieved. To use both user- and enrollee-specific second best scores, the Eq. 4 with six dimensional probability densities of genuine and impostor scores is used to generate the likelihood ratio. The results of combination experiments are shown in Fig. 5. It says that the utilization of both user-specific and enrollee-specific second best scores can further improve the performance of biometric systems compared to using either user-specific or enrollee-specific second best scores individually in all cases.

5. Conclusions

Both user-specific and enrollee-specific statistical information can provide a significant improvement to biometric systems. In this paper we investigated the simultaneous use of user- and enrollee-specific normalization methods as well as user- and enrollee-specific second best score statistics in both unimodal and multimodal verification systems. The experimental results show that using both user- and enrollee-specific statistical information at the same time can further improve the performance of biometric systems compared to using either user- or enrollee-specific information individually. In future, it is possible to consider the whole range of models so that some models are suited for one type of biometric matchers and other models for other types.

References

[1] Nist biometric scores set. http://www.nist.gov/biometricscores/. 161

[2] G. Aggarwal, N. K. Ratha, and R. M. Bolle. Biometric verification: Looking beyond raw similarity scores. *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pages 31–31, 2006. 163

[3] 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. doi: DOI: 10.1006/dspr.1999.0360. 162, 163

[4] 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. doi: DOI: 10.1016/j.cviu.2003.08.002. 164

[5] R. Brunelli and D. Falavigna. Person identification using multiple cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(10):955–966, 1995. 162

[6] G. N. R. Jiri Navratil. The awe and mystery of t-norm. *In 8th European Conference on Speech Communication and Technology (EUROSPEECH-2003)*, pages 2009–2012, 2003. 162

[7] A. Kumar and D. Zhang. Improving biometric authentication performance from the user quality. *Instrumentation and Measurement, IEEE Transactions on*, 59(3):730–735, 2010. doi: 10.1109/TIM.2009.2028773. 161

[8] J. Mariethoz and S. Bengio. A unified framework for score normalization techniques applied to text-independent speaker verification. *Signal Processing Letters, IEEE*, 12(7):532–535, 2005. 162

[9] W. Peng, J. Qiang, and J. L. Wayman. Modeling and predicting face recognition system performance based on analysis of similarity scores. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):665–670, 2007. 161

[10] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35(4):861–874, 2002. doi: DOI: 10.1016/S0031-3203(01)00103-0. 163

[11] H. Tin Kam, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(1):66–75, 1994. 162

[12] S. Tulyakov, L. Jiang, and V. Govindaraju. Enrolled template specific decisions and combinations in verification systems. *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–7, 2008. 162, 163, 164

[13] S. Tulyakov, Z. Zhi, and V. Govindaraju. Comparison of combination methods utilizing t-normalization and second best score model. *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–5, 2008. 162, 163