Comparison of Combination Methods Utilizing T-normalization and Second Best Score Model

Sergey Tulyakov, Zhi Zhang, and Venu Govindaraju
Center for Unified Biometrics and Sensors
University at Buffalo, NY, USA
{tulyakov,zhizhang,venu}@cedar.buffalo.edu

Abstract

The combination of biometric matching scores can be enhanced by taking into account the matching scores related to all enrolled persons in addition to traditional combinations utilizing only matching scores related to a single person. Identification models take into account the dependence between matching scores assigned to different persons and can be used for such enhancement. In this paper we compare the use of two such models - T-normalization and second best score model. The comparison is performed using two combination algorithms - likelihood ratio and multilayer perceptron. The results show, that while second best score model delivers better performance improvement than T-normalization, two models are complementary to each other and can be used together for further improvements.

1. Introduction

Biometric matching scores usually represent some variation of distance between two biometric templates - enrolled template and test template. Thus, it is expected that the quality of any of these templates will have an influence on matching score. For example, if test fingerprint image has small area and contains only few minutia, then its matching score with enrolled fingerprint of the same finger will be probably low. If we try to match this fingerprint with some other, impostor, templates enrolled in the database, then we can also expect lower than usual matching scores. Consequently, the quality of test (or enrolled) biometric template can be implicitly estimated using a set of its matching scores with other, genuine or impostor, biometric templates.

This observation can be effectively used to enhance any algorithm relying on biometric matching scores, for example making decisions in person verification applications or combining biometric scores of different modalities. Instead of using a single matching score between test and enrolled biometric templates, we can additionally use a derived quality measure for any of these templates. Such quality measure will provide information on how reliable is original matching score. Alternatively, a matching score can be normalized using derived quality measures.

It is not necessary to explicitly derive a template quality from a set of matching scores as in [17]. Rather, we need to know what information from a set of matching scores is most useful for improving the system and how to utilize it in a best way. In this work we are interested in extracting information with respect to the test template. We call the process of matching a single test template to a set of enrolled templates as an identification trial. The information extracted from identification trial scores and the algorithm utilizing this information is called an identification model. The goal of this paper is to compare two identification models - the identification model obtained by T-normalization algorithm [3] and second best score identification model [15]. We use both models for combination of biometric matchers by means of likelihood ratio and multilayer perceptron.

To be clear, we use a NIST BSSR1 database of biometric scores, which has only a single template for any enrolled person, and a single test template is used for each identification trial. So each set of identification trial scores contains one genuine and $N - 1$ impostor scores, N is the number of enrolled persons. Thus we are not considering cases where multiple templates of the same modality are enrolled for the same person, or where multiple verification attempts are performed for the same person. We will also be dealing with the combinations of two matchers only, though the theory can be readily applied to a bigger number of combined matchers.

2. Terminology

Though in this paper we measure the performance of our systems for the verification mode of operation only, we use the term 'identification trial scores' to designate all the matching scores between a test template and en-

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

rolled templates of all persons in the database. This set of matching scores is normally produced for the biometric systems operating in identification mode. In contrast to traditional combination algorithms using only a single matching score between a test template and an enrolled template of claimed identity, our algorithm requires using this bigger set of scores, and this is one of the reasons we used the term in this paper. The corresponding term to represent a properties of this score set is 'identification model'.

Though there are alternatives to this term, we think they might not exactly convey the meaning of this score set. For example, 'test scores' and 'test model' might have an overlap with already used T(test)-normalization and other usage outside classifier combination field. 'Background' and 'cohort' models have been used in classifier combination field to deal with both identification trial scores and the set of matching scores between a single enrolled (not test) template and other enrolled templates. We provide references of previous usage of these terms in the next section.

3. Previous Work - Background and Identification Models

The variation of matching scores produced during identification trials depends on the quality of test biometric templates and makes their straightforward usage unreliable. Similar observation was made by Ho et al. [7] with respect to the problem of handwritten character recognition and the quality of input (test) image. The proposed solution converted matching scores to their respective ranks among all characters in the alphabet, and used ranks instead of original scores in the combination algorithm. In order to convert matching score to its rank we need to consider the whole set of identification trial scores, and thus ranking can be considered as one possible implementation of identification model. As the wide usage of rank based methods for combination suggests (Borda count, decision trees [7] and forests [6], Behavior-Knowledge Spaces [8]) even this simple identification model can be quite beneficial for combination algorithms. On the other hand, it is clear that conversion to ranks discards important information, the matching scores themselves, which is also useful to combination algorithms.

The concept of background model has been previously introduced in the speaker identification applications [13, 4]. The idea of background models is similar to the idea of identification models - the model should reflect the characteristics of a template with respect to other templates. We can define the difference between two models in the way as shown in Figure 1: background models account for the enrolled template, and identification models account for test template. Though earlier developed background models in speaker identification research might include both enrolled and test template models, it is rather convenient to separate them in our research.

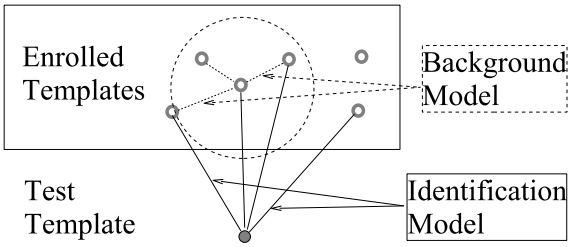


Figure 1. Schematic view of background and identification models.

One example of previous use of background models is the cohort based score normalization for speaker verification [13, 4] and for fingerprint verification [2]. Cohort methods find a cohort - a subset of enrolled templates close to the one under consideration as shown by a circle in Figure 1. During the matching of the enrolled and test templates, the matching score can be modified either by the set of matching scores from background model (matching scores between enrolled template and its cohort templates), or by the set of matching scores from identification model (matching scores between test template and enrolled cohort templates). Thus, cohort methods might include both background models and identification models.

Furthermore, Auckenthaler et al. [3] separated cohort normalization methods into cohorts found during training (constrained) and cohorts dynamically formed during testing (unconstrained cohorts). As in the previous paragraph, constrained cohorts might not include identification model, but only background model. But unconstrained cohorts might only use identification model and no background model. Also, both types of cohorts can utilize both background and identification models at the same time.

As an example of direct construction of background models, we can cite the algorithms learning user-specific biometric parameters of enrolled templates [9, 5, 17]. Such algorithms not only construct background models, but also attempt to make such model different for different users or enrolled templates. Note, that cohort methods usually imply user-specific cohort parameters as well.

T-normalization [3, 10] is the example of simple identification model. Each matching score is normalized using the mean μ and standard deviation σ of the set of all matching scores produced during single identification trial:

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

Another identification model is the second best score model [15, 16], which considers a pair of original score and best score from identification trial besides original score instead of a single original matching score. In terms of cohort

methods, T-normalization is equivalent to considering unconstrained cohort consisting of all enrolled templates, and second best score model is equivalent to considering an unconstrained cohort consisting of only one enrolled template - the one closest to the test template (note, that cohort template is different from the enrolled template under matching consideration). Obviously, some intermediate identification models between these two extremes can be considered similar to [14], where different numbers of enrolled templates closest to the test one are used in a normalization similar to T-norm.

4. Identification Models

The goal of this paper is to compare two identification models, T-normalization and second best score model, in the problem of combining two matchers in biometric person verification application. In this section we present a brief overview of these two models.

The important characteristic of identification model is its ability to correct the variation of scores obtained during same identification trial. As formula 1 for T-normalization suggests, T-normalization can account for score variations involving spreading them by product with some constant factor and addition of some constant to all scores in the identification trial. The constants can change for different identification trials. If our matching scores in different identification trials have only these variations, then T-normalization is an optimal identification model, since it will successfully (with some approximation) account for matching score dependencies. Navratil and Ramaswamy [11] considered this property in more detail and introduced the concept of local gaussianity, so that if matching scores possess this property, their T-normalization will have constant gaussian distribution.

The second best score identification model can be represented by the following formula:

$$s \rightarrow \{s, sbs(s)\} \tag{2}$$

where $sbs(s)$ is the best score besides s obtained in the same identification trial. In contrast to T-normalization, second best score model produces two numbers instead of one, which might allow bigger flexibility in training combination algorithm. Whereas the score variations are rigidly modeled by T-normalization model, the subsequent training of algorithm using second best score model can automatically account for different score variations. For example, if we assume that the scores in identification trials are subjected to the same addition by the constant as in T-normalization, both s and $sbs(s)$ will be shifted by the same constant. The combination algorithm can be trained to use the difference $s - sbs(s)$, which will be the same for all such shifts.

Another factor which we might consider when choosing used identification model is information contained in derived score set statistics, such as mean and standard deviation for T-normalization and $sbs(s)$ for second best score identification model, with respect to predicting a performance of considered matching score. One way to measure such information is to verify that such statistics can indeed predict that the score is genuine or impostor. In our experiments we use biometric score set distributed by NIST[1] consisting from matching scores for two fingerprints by one matcher, and two sets of face scores produced by two different face matchers. As it was noted before [16], the best impostor score has similar or higher correlation with genuine score than the mean of the impostor scores. This implies that second best score model has the same or more reliability in evaluating the genuine score than T-normalization. Note, that T-normalization also utilizes the standard deviation of identification trial matching scores, but it is plausible it has little effect on genuine score evaluation.

5. Combination methods

The first algorithm which we used to compare two models is the likelihood ratio combination method. This is theoretically optimal combination method for verification system [12] and consists in assigning a combined score a value of the ratio between genuine and impostor score densities:

$$S = \frac{p_{gen}(s^1, s^2)}{p_{imp}(s^1, s^2)} \tag{3}$$

where s^i is the verification matching score assigned by the matcher i. The likelihood ratio with T-normalization will operate by the same formula, only using T-normalized scores s^i. The likelihood ratio method using second best score model will consider the joint densities of scores and second best score statistics:

$$S = \frac{p_{gen}(s^1, sbs(s^1), s^2, sbs(s^2))}{p_{imp}(s^1, sbs(s^1), s^2, sbs(s^2))} \tag{4}$$

The use of T-normalization and second best score model at the same time implies first T-normalization of combined scores, and then using second best score model likelihood ratio combination using above formulas. Note that for methods utilizing T-normalization, the training (approximation of score densities) is performed on T-normalized scores. In order to approximate score densities we use the Parzen window method with gaussian kernels.

We also repeated the comparison of the same identification models using multilayer perceptron instead of likelihood ratio. Direct approximation of score densities might be problematic in a higher dimensional space corresponding to the case of bigger number of combined matchers, and using alternative classification methods can be beneficial. We fixed the structure of perceptron to have 3 layers with 8 nodes in first hidden layer and 9 nodes in second hidden layer. The output layer had one node with

goal values of 0 and 1 corresponding to impostor and genuine verification attempts. The input layer traditional and for T-normalization had two nodes for two original or T-normalized scores $\{s^1, s^2\}$ from two matchers. The input layer for second best score identification model had 4 nodes for two pairs of scores and second best score statistics $\{s^1, sbs(s^1), s^2, sbs(s^2)\}$ from two matchers. The logistic function is used as a threshold function for all layers.

6. Experiments

We performed experiments on a set of biometric scores by NIST[1]. All six possible two-matcher combinations has been investigated using fingerprint matching scores for left index 'li' and right index 'ri' fingers produced by the same matcher, and face matching scores produced by two different matchers, 'C' and 'G'. We used the bigger subsets of the database involving 6000 users (identification trials). Since the scores in these subsets originate from different persons, we assumed the independence of fingerprint and face matching scores, and considered randomly paired set of scores corresponding to 6000 identification trials of 3000 enrolled persons. Note that correspondence of scores to the same physical person was retained when combining scores of the same modality. Also note, that some enrollee and user scores had to be discarded due to apparent template acquisition errors, resulting in 5982 identification trials and 2991 enrollees.

We applied bootstrap sample testing technique in our experiments. For each bootstrap test, 2991 identification trials were randomly chosen as test set, 1000 trials were chosen as training set and remaining 1991 trials were chosen as validation set. 100 bootstrap tests were performed for each experiment.

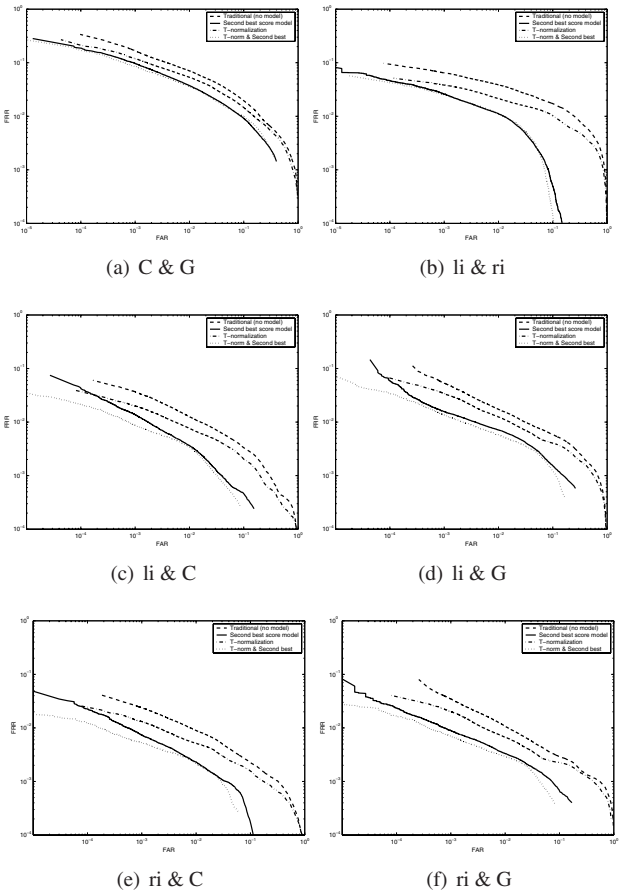
Since each identification trial had 1 genuine and 2990 impostor scores we chose to use only a single random impostor score for each genuine score from the same identification trial for training. The validation sets were used to estimate the kernel sizes for likelihood ratio methods, and to stop the training of multilayer perceptrons when the minimum MSE is achieved for validation set.

The results of testing are presented in figures 2 and 3. In almost all cases second best score identification model was able to show better improvement than T-normalization. This results seem to confirm the analysis of used statistics (mean and $sbs(s)$) in predicting the genuine property of considered score presented in [16]. We also note that using both models at the same time had even better performance, which indicates the complementary nature of both models.

7. Conclusions

The identification models can provide a significant improvement to biometric systems utilizing biometric matching scores. In this paper we showed how biometric score combination algorithm can be improved using two such models - T-normalization and second best score model. We also tried to analyze the strengths of these methods, and it seems that for considered biometric matchers second best score model can provide a better performance than T-normalization.

As we also described in this paper, it is possible to consider a whole range of identification models in the future research. It might turn out that some identification models will be suited for one type of biometric matchers, and other models for other types. There has been only little research in this area so far, and more theoretical and practical results are needed.



(a) C & G (b) li & ri

(c) li & C (d) li & G

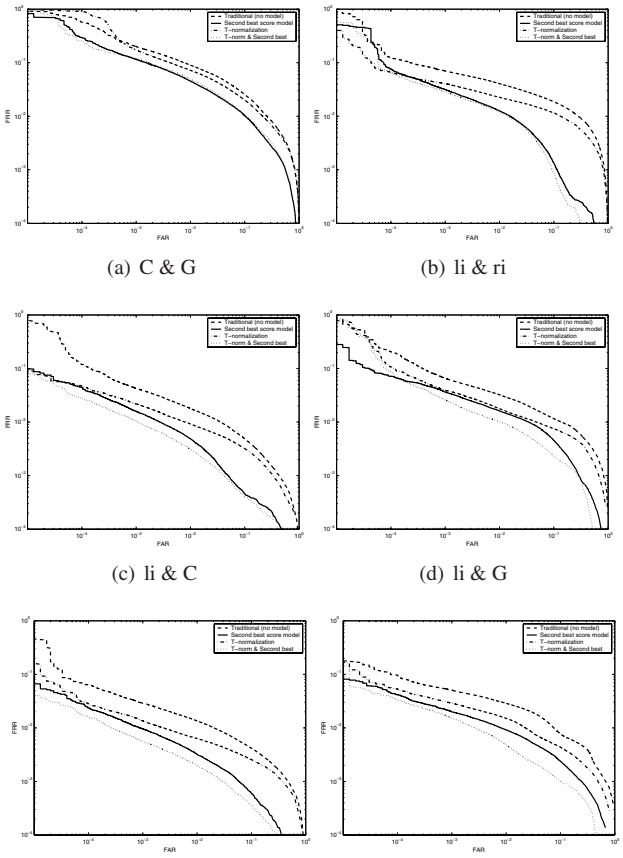
(e) ri & C (f) ri & G

Figure 2. ROC curves for likelihood ratio combinations utilizing and not utilizing identification models.

References

[1] Nist biometric scores set. http://www.nist.gov/biometricscores/. 3, 4

[2] G. Aggarwal, N. Ratha, and R. Bolle. Biometric verification: Looking beyond raw similarity scores. In



(a) C & G (b) li & ri

(c) li & C (d) li & G

(e) ri & C (f) ri & G

Figure 3. ROC curves for neural network combinations utilizing and not utilizing identification models.

Computer Vision and Pattern Recognition Workshop, 2006 Conference on, page 31, 2006. 2

[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. 1, 2

[4] J. Colombi, J. Reider, and J. Campbell. Allowing good impostors to test. In *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, volume 1, pages 296–300 vol.1, 1997. 2

[5] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Bayesian adaptation for user-dependent multimodal biometric authentication. *Pattern Recognition*, 38(8):1317–1319, 2005. 2

[6] T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998. 2

[7] T. K. Ho, J. Hull, and S. Srihari. Decision combination in multiple classifier systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(1):66–75, 1994. 2

[8] Y. Huang and C. Suen. A Method of Combining Multiple Experts for Recognition of Unconstrained Handwritten Numerals. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 17(1):90–94, 1995. 2

[9] A. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–57–I–60 vol.1, 2002. 2

[10] J. Mariethoz and S. Bengio. A unified framework for score normalization techniques applied to text independent speaker verification. *IEEE Signal Processing Letters*, 12, 2005. 2

[11] J. Navratil and G. N. Ramaswamy. The awe and mystery of t-norm. In *8th European Conference on Speech Communication and Technology (EUROSPEECH-2003)*, pages 2009–2012, Geneva, Switzerland, 2003. 3

[12] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35(4):861–874, 2002. 3

[13] A. Rosenberg and S. Parthasarathy. Speaker background models for connected digit password speaker verification. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 81–84 vol. 1, 1996. 2

[14] A. Schlapbach and H. Bunke. Using hmm based recognizers for writer identification and verification. In *9th Intl Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)*, 2004. 3

[15] S. Tulyakov and V. Govindaraju. Classifier combination types for biometric applications. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), Workshop on Biometrics*, New York, USA, 2006. 1, 2

[16] S. Tulyakov and V. Govindaraju. Identification model for classifier combinations. In *Biometrics Consortium Conference*, Baltimore, MD, 2006. 2, 3, 4

[17] P. Wang, Q. Ji, 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. 1, 2