IDENTIFICATION MODEL FOR CLASSIFIER COMBINATIONS

Sergey Tulyakov and Venu Govindaraju

Center for Unified Biometrics and Sensors (CUBS)
SUNY at Buffalo, USA

ABSTRACT

This paper considers combinations of biometric matchers in identification system. We assume that the test template is matched not only against the enrolled template of claimed person identity, but also against few enrolled templates of other persons, and all matching scores are available to the combination algorithm. We present a combination method utilizing the dependencies between these scores and showing better performance than comparable traditional combination method using only matching scores related to the claimed identity.

1. INTRODUCTION

Let us consider a person identification system with biometric matcher producing a set of confidence scores $\{s_1, \ldots, s_N\}$ for one identification trial, N is the number of enrolled persons. In order to make an identification decision we have to find an index i_1 so that s_{i_1} is the the highest confidence in this set. One can notice, though, that if there are other scores s_i in this set which are close to s_{i_1}, then it is quite probable that i_1 does not correspond to the true test identity, and the identification attempt should be rejected. [1] investigates this question in detail and concludes that the performance of identification systems can be significantly improved if decision is based not only on the best score s_{i_1} but also on the second best score s_{i_2}.

Thus the second best score can be regarded as the useful statistic about the set of matching scores which is used to transform best matching score into the likelihood of best matching score belonging to genuine user. In this paper we will investigate if similar algorithm could be used in the combination of biometric matchers. The idea behind the algorithm is that not only the best score but all other scores are transformed using second best score statistics to represent likelihoods of corresponding classes being genuine.

1.1. Dependence Between Match Scores

As above, $\{s_1, \ldots, s_N\}$ is the set of match scores output by one biometric matcher during one identification attempt, and we assume that one of these scores, e.g. s_{gen}, is a genuine match score and all other scores are impostor match scores. It turns out that typically these scores are dependent. Let us denote $mean_{imp}$ as the mean of impostor scores produced during one identification trial, $first_{imp}$ and $second_{imp}$ as the best and second best impostor scores for one identification trial. Given experimental data of few identification trials, we can extract the values of s_{gen}, $mean_{imp}$, $first_{imp}$ and $second_{imp}$ for each trial, and find the correlations between them across different trials. Table 1 shows the correlations between s_{gen} and functions of the set of impostor scores extracted from NIST biometric score set [2]. In this table 'li' and 'ri' correspond to left and right index finger match scores produced by the same fingerprint matcher, and 'C' and 'G' correspond to face match scores produced by so named face matchers. The correlations are computed for 6000 identification trials, and each trial has 2999 impostors for faces and 5999 impostors for fingerprints. Positive correlations confirm that there is a dependency between genuine and impostor scores output by biometric matchers.

Matchers	$first_{imp}$	$second_{imp}$	$mean_{imp}$
li	0.3164	0.3400	0.2961
ri	0.3536	0.3714	0.3626
C	0.1419	0.1513	0.1440
G	0.1339	0.1800	0.1593

Table 1. Correlations between s_{gen} and different statistics of the impostor score sets produced during identification trials in NIST BSSR1 data.

Since there exists a dependence between genuine and impostor scores produced during identification trials, it would be desirable to somehow utilize it for score normalization or directly for combination. The intuition behind such need could be illustrated by the following example. Suppose for our matcher bigger score means better match, and suppose we have an output match score of 0.5 for some pair of test and enrolled biometric templates. Suppose we also have a set of matching scores of this test template with other enrolled templates, and we extract some statistic of this set, say, second best score. If, for example, second best score is 0.4, then we have high confidence that the score of 0.5 corresponds to the genuine match. And if the second best score is .6, then we have less confidence that 0.5 is a genuine match. Since

there is a strong correlation between genuine score and best or second best impostor (the second best score is one of them), then we expect that genuine score be on par with second best score, and consequently have less confidence of score .5 being genuine in the second case. Thus the second best score statistics, or any other statistics of the scores produced during one identification trial, can be used to improve the estimate of the likelihood of a score being genuine or impostor.

1.2. Score Set Statistics and Identification Model of a Matcher

Suppose that identification system consists of M biometric matchers and each matcher produces a set of scores $\{s_1^j, \ldots, s_N^j\}, 1 \leq j \leq M$. Usually combination algorithms consider score combination functions accepting as parameters only scores related to one class: $S_i = f(s_i^1, \ldots, s_i^M)$. Such combinations effectively discard the relationships between scores output by one classifier. On the other hand, using all MN output scores of M matchers and N classes in order to derive a combined score S_i for class i can be difficult since the number of classes in biometric identification system is frequently large or varying. If combination function accepts all matching scores as parameters, its training becomes problematic. Using score statistics from the identification trial represents a trade-off between these two extreme approaches. Such statistics can deliver important information about the whole set of scores output by one matcher, and the number of parameters to the combination function remains proportionate to the number of matchers - CM, where C is the number of used statistics.

Score statistics, e.g. nth best score, mean, variance, etc., extracted from the set of single identification trial reflect the quality of test biometric template or the test user. Using a training set of match scores containing multiple score sets from many identification trials, we can learn the relationships between these statistics and genuine or impostor scores. The term 'identification model' represents any learned model considering these relationships. The better we are able to learn the identification model of each biometric matcher, the more precise estimates of scores being genuine or impostors we will be able to get.

The general algorithm for utilizing identification models for combination is presented in Figure 1. The rows of the score matrix represent scores produced by each matcher, and columns correspond to scores related to each class or person. The identification model is applied first to rows of the score matrix. Note that each matcher has its own learned identification model. The identification transforms the scores to represent likelihoods of scores being genuine matches, or some other well defined variable. On the second stage normalized scores from different matchers are combined by some predetermined or trained combination function. We performed experiments on two biometric matchers (face and fingerprint),

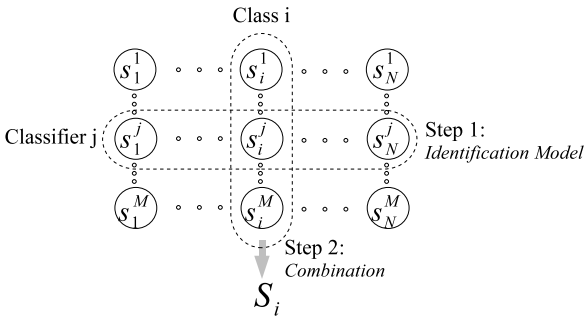


Fig. 1. Combination method utilizing identification model.

and the combination is performed by a predetermined (product of likelihoods) function.

In this work we utilize second best statistic for modeling score relationships in identification models. The previous work of [1] and the correlations of table 1 confirm that using this statistic is reasonable. We use Bayesian classification as a combination algorithm for our experiments. Thus our identification model is a learned joint score densities $p(s_i^j, t_i^j | C_i)$ of score s_i^j and statistic t_i^j, with C_i representing either s_i^j being genuine score or impostor.

2. PREVIOUS WORK

Biometric applications are traditionally separated into verification and identification systems. Verification systems usually have only the confidence scores related to one person available for combination, and their task is to separate genuine and impostor verification attempts. Thus, for verification systems, score matrices in Figure 1 have only one column, and all types collapse into one. Identification systems, on the other hand, can have matching confidence scores produced for all enrolled persons, and score matrices with $N > 2$ columns.

If we have a combination algorithm for verification system, we can use it sequentially for all persons in identification system. Such algorithm will not utilize dependencies between scores output by a single matcher. Most of combination algorithms used in biometric applications are of this type. As an example, some combination algorithms are based on FAR, FRR or ROC curves, and since the construction of these curves assumes the independence of scores in identification trial, such algorithms do not account for score dependence. Below we present approaches which do utilize score dependencies in identification trials: rank based combinations and some score normalization techniques.

2.1. Rank Based Combinations

The frequent approach to combination in identification systems is to use rank information of the scores. This approach

transforms combination problems with measurement level output classifiers to combination problems with ranking level output classifiers ([3]). T.K. Ho has described classifier combinations on the ranks of the scores instead of scores themselves by arguing that ranks provide more reliable information about class being genuine [4, 5]. Thus, if the input image has low quality, then the genuine score, as well as the impostor scores will be low. Combining low score for genuine class with other scores could confuse a combination algorithm, but the rank of the genuine class remains to be a good statistic, and combining this rank with other ranks of this genuine class should result in true classification. Brunelli and Falavigna [6] considered a hybrid approach where traditional combination of matching scores is fused with rank information in order to achieve identification decision. Hong and Jain [7] consider ranks, not for combination, but for modeling or normalizing classifier output score. Saranli and Demirekler [8] provide additional references for rank based combination and a theoretical approach to such combinations.

Rank-based methods do utilize the score dependencies in identification trials, and, as many authors suggest, these methods provide a better performance in identification systems. The problem with rank based methods, however, is that the score information is somewhat lost. Indeed, genuine score can be much better than second best score, or it could be only slightly better, but score ranks do not reflect this difference. It would be desirable to have a combination method which retains the score information as well as the rank information.

2.2. Score normalization approaches

Usually score normalization [9] means transformation of scores based on the classifier's score model learned during training, and each score is transformed individually using such a model. Thus the other scores output by a matcher during the same identification trial are not taken into consideration. If these normalized scores are later combined class-wise, then score dependence will not be accounted for by the combination algorithm.

Some score normalization techniques can use a set of identification trial scores output by classifier. For example, Kittler et al. [10] normalize each score by the sum of all other scores before combination. Similar normalization techniques are Z(zero)- and T(test)- normalizations [11, 12]. Z- normalization uses impostor matching scores to produce a class specific normalization. Z-normalization does not include the set of identification trial scores, and thus does not utilize score dependency. On the other hand, T-normalization does use a set scores produced during single identification trial, and can be considered as a simple form of identification model. T-normalization uses statistics of mean and variance of identification score set. Note that identification model implies some learning algorithm, but T-normalization is a predetermined routine with no training. Still, using this simple kind of score modeling turns out to be quite useful; for example, [13] argued for applying T-normalizations in face verification competition.

In [14] we showed theoretically on an artificial example that if combination algorithm does not perform score normalization based on the set of scores produced during identification trial, then such algorithm might perform suboptimally. Even if the combination is based on optimal Bayesian algorithm, the performance of combination might be worse than the performance of a single classifier. In this paper we are presenting specific score normalization methods, which do use a set of identification trial scores, and deliver a better performance of the combination algorithm. Note that we try to compare similar Bayesian combination methods with and without such normalization, and are not using heuristic methods such as sum and product rule. In preliminary testing sum rule performed worse than Bayesian combination, and comparison with heuristic rules is not a purpose of this paper.

3. COMBINATIONS USING IDENTIFICATION MODEL

The goal of this section is to derive specific algorithms for using identification models in combinations. We use a Bayesian approach of modeling match score densities, and make a simplifying assumption that our matchers are statistically independent, that is any score from one matcher is independent from any score of another matcher. Note that the scores output by the same matcher are dependent, and the purpose of identification model is to account for this dependence.

3.1. Combinations by Modeling Score Densities

Suppose that we combine M independent classifiers, and each classifier outputs N dependent scores corresponding to N classes. The Bayesian classifier used as combination algorithm chooses the class which maximizes the posterior class probability. An input is a whole set of NM scores output by all the M combined classifiers. Thus the goal of classification is to find

$$\arg\max_k P(C_k|\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M})$$

Term C_k refers to the fact that the class k is the genuine class. By the Bayes theorem

$$P(C_k|\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}) = \frac{p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}|C_k)P(C_k)}{p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M})} \quad (1)$$

and since the denominator is the same for all classes, our goal is to find

$$\arg\max_k p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}|C_k)P(C_k)$$

or, assuming all classes have the same prior probability,

$$\arg\max_k p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}|C_k)$$

By our current assumption, classifiers are independent, which means that any subset of scores produced by one classifier is statistically independent from any other subset of scores produced by another classifier. Hence, our problem is to find

$$\arg\max_k \prod_j p(\{s_i^j\}_{i=1,\ldots,N}|C_k) \quad (2)$$

The problem now is to reliably estimate the densities $p(\{s_i^j\}_{i=1,\ldots,N}|C_k)$, which is a rather hard task given that the number N of classes is large and we do not have many samples of each class for training. The last problem is solved by noticing that we do not construct class specific combination, and thus class indexes can be permuted. Thus all training samples pertaining to different genuine classes can be used to train only one density, $p(s_k, \{s_i^j\}_{i=1,\ldots,N,i\neq k}|C_k)$. Now s_k^j is a score belonging to genuine match, and all other scores $\{s_i^j\}_{i=2,\ldots,N}$ are from impostor matches. Since there are many impostor scores participating in this density, we might somehow try to eliminate them. This is where we apply our identification model.

Instead of $p(s_k, \{s_i^j\}_{i=1,\ldots,N,i\neq k}|C_k)$ we can consider $p(s_k^j, t_k^j|C_k)$, where t_k^j is a statistic of identification trial score set, e.g. second best score for classifier j, given that the genuine class is k. Note that if s_k is the best matching score, then t_k^j is the second best score, and if s_k is not the best score, then t_k^j is the best score. Thus the combination rule is the following:

$$\arg\max_k \prod_j p(s_k^j, t_k^j|C_k) \quad (3)$$

3.2. Combinations by Modeling Posterior Class Probabilities

As above we consider posterior class probability, apply Bayes formula, but now use independence of classifiers to decompose the denominator:

$$\begin{aligned} P(C_k|\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}) = \\ \frac{p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}|C_k)P(C_k)}{p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M})} = \\ \frac{\prod_j p(\{s_i^j\}_{i=1,\ldots,N}|C_k)P(C_k)}{\prod_j p(\{s_i^j\}_{i=1,\ldots,N})} = \\ P(C_k)\prod_j \frac{p(\{s_i^j\}_{i=1,\ldots,N}|C_k)}{p(\{s_i^j\}_{i=1,\ldots,N})} \end{aligned} \quad (4)$$

The next step is similar to the step in deriving the algorithm for background speaker model [12]. We consider class $\overline{C_k}$ meaning some other class is genuine, and decompose

$$\begin{aligned} p(\{s_i^j\}_{i=1,\ldots,N}) = P(C_k)p(\{s_i^j\}_{i=1,\ldots,N}|C_k)+ \\ P(\overline{C_k})p(\{s_i^j\}_{i=1,\ldots,N}|\overline{C_k}) \end{aligned} \quad (5)$$

By assuming that N is large and $P(\overline{C_k}) \gg P(C_k)$, we can discard the first term and get the following classifier decision:

$$\arg\max_k \prod_j \frac{p(\{s_i^j\}_{i=1,\ldots,N}|C_k)}{p(\{s_i^j\}_{i=1,\ldots,N}|\overline{C_k})} \quad (6)$$

In comparison with decision 2 of the previous section we have have additional density $p(\{s_i^j\}_{i=1,\ldots,N}|\overline{C_k})$. Such density can be viewed as a background of impostors for the genuine class C_k. As research in speaker identification suggests, utilizing such background model is helpful.

One way to model these ratios could be a direct reconstruction of the posterior class probabilities (ratios in equation 4 are exactly these probabilities without priors). The other way is by additional modeling of $p(\{s_i^j\}_{i=1,\ldots,N}|\overline{C_k})$. We used an approach similar to the previous section to estimate this density as $p(s_k, t_k^j|\overline{C_k})$, but t_k^j now is not the best impostor (we do not know what score is genuine, and thus can not know the best impostor), but simply the second best score.

The technique described in this section can be characterized as a composition of identification model and background model. The identification model considers $p(s_k, t_k^j|C_k)$ and $p(s_k, t_k^j|\overline{C_k})$ instead of $p(s_k|C_k)$ and $p(s_k|\overline{C_k})$, and background model considers $p(s_k, t_k^j|\overline{C_k})$ or $p(s_k|\overline{C_k})$ in addition to $p(s_k, t_k^j|C_k)$ or $p(s_k|C_k)$. The background model makes score normalization under the assumption of the independence of scores assigned to different classes, and identification model accounts for dependencies of scores.

3.3. Extension to Combinations of Dependent Classifiers

The combination algorithms presented in the previous two sections deal with independent classifiers. How should we address dependent classifiers?

By looking at the combination formulas 2 and 6 we can see that each classifier contributes terms $p(\{s_i^j\}_{i=1,\ldots,N}|C_k)$ and $\frac{p(\{s_i^j\}_{i=1,\ldots,N}|C_k)}{p(\{s_i^j\}_{i=1,\ldots,N}|\overline{C_k})}$ correspondingly to the combination algorithm. Thus one can conclude that it is possible to model the same terms for each classifier, and then combine them by some other trainable function.

Note that any relationships between scores $s_{i_1}^{j_1}$ and $s_{i_2}^{j_2}$ where $i_1 \neq i_2$ and $j_1 \neq j_2$ will be essentially discarded.

4. EXPERIMENTS

We conducted experiments using NIST BSSR1 biometric score database [2]. We used two subsets of fingerprints: li (left index) and ri (right index), and two subsets of face scores from

two face matchers C and G. Since we wanted to consider the case of independent matchers we performed four sets of experiments on combining fingerprint and face scores : li combined with C, li combined with G, ri combined with C, and ri combined with G.

Results are presented in Table 2. The columns represent the combination method. 'Traditional' is the method of reconstructing densities of genuine and impostor score pairs, and performing Bayesian classification using this densities. This approach discards score dependencies. 'Density' is the method outlined in section 3.1. 'PP'(Posterior Probability) is the method from section 3.2. All the densities are reconstructed using original scores linearly normalized to interval $[0, 1]$, and the kernel sizes are derived using the maximum likelihood method.

Matchers	# of tests	Traditional	Density	PP
li & C	516	5	7	4
li & G	517	9	11	6
ri & C	516	3	3	2
ri & G	517	3	2	2

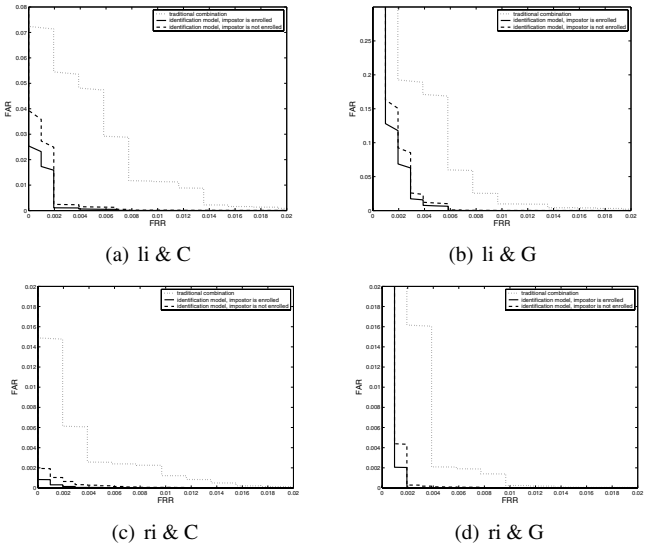
Table 2. Experiments on combinations in identification systems. Entries are the numbers of failed test identification trials.

All experiments were performed in leave-one-out framework. The numbers in the tables are the numbers of failed tests, and total number of tests is also given. Failed test means that the impostor got the best combination score in this particular identification attempt.

The algorithm for traditional combinations models the optimal Bayesian decision by approximating score densities. For each pair of scores the combined score is derived as a ratio of genuine and impostor density function approximations at this score pair. Thus, this combination method automatically deals with the background model - the density of impostors participates in the combined score. This might explain why traditional combinations got better results than combinations based on genuine score density approximation as in section 3.1 ('Density' method in table 2). But if identification model is combined with background model as in section 3.2 ('PP' method), then we are able to obtain better combination than the traditional method.

5. IDENTIFICATION MODEL FOR VERIFICATION SYSTEMS

Although there are examples where score normalization techniques with background models are used for speaker identification tasks[6], even more applications use such techniques for speaker verification systems [15, 16, 12]. We also applied the combinations utilizing identification models for biometric person verification tasks. The drawback of using either the background models or the identification models in verification tasks is that we have to produce not only one match per person and per matcher, but also some set of matching scores for other persons enrolled in the system, or some artificially modeled persons.



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

(c) ri & C (d) ri & G

Fig. 2. ROC curves for traditional combinations and combinations utilizing identification models in verification tasks.

In our experiments for each test person we performed match of input biometric with biometric templates of all enrolled persons. All these scores were used to derive a score normalized by identification and background models as in section 3.2 ('PP' method). The ROC curves were obtained by means of thresholding these normalized scores for both genuine and impostor verification attempts. These curves are drawn in Figure 2 together with ROC curves corresponding to traditional combinations.

We distinguish two possible cases with respect to impostors in such verification systems: impostor is enrolled in the database, and impostor is not enrolled in the database. If the impostor is in the database, and impostor attempts to be verified as another person, we expect that match score to the true impostor's identity will be higher than impostor's match score to the claimed identity. Thus a verification system utilizing the identification model (and hence all matching scores) is more likely to reject this impostor's matching attempt. Experimental results in Figure 2 show that the performance of verification system is better if impostor is enrolled in the database than when the impostor is not in the database. But this difference in performance is small, and both cases have better performance than traditional combination. The small difference in performance can be explained by the fact that our identification model algorithm uses second-best impostor statistics instead of best impostor statistics (section 3.2).

6. CONCLUSION

In order to account for the relationships between scores assigned by one classifier to different classes, we introduced the concept of the identification model. The identification model application is a score normalization algorithm where normalization depends on all scores output by a classifier in one identification trial, and the algorithm is the same for all classes. Thus our identification model is simpler than similar attempts to normalization [17, 18]. In these previous attempts normalizations were class specific and required huge amount of training data. Biometric identification problems can have large number of enrolled persons, and such combinations are not feasible due to the lack of training data. By restricting ourselves to non-class-specific normalizations of the identification model we avoid the problem of combination algorithm training.

At the same time, our approach is more complex than traditional combination algorithms which disregard the dependence between scores in identification trials. The experimental results presented in Table 2 and in Figure 2 show that we were able to achieve significant improvement in the performance of identification and verification systems by utilizing the dependence of matching scores in identification model.

7. REFERENCES

[1] S. Tulyakov and V. Govindaraju, "Combining matching scores in identification model," in *8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, Seoul, Korea, 2005.

[2] "Nist biometric scores set. http://www.nist.gov/biometricscores/," .

[3] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods for combining multiple classifiers and their applications to handwriting recognition," *IEEE transactions on System, Man, and Cybernetics*, vol. 23, no. 3, pp. 418–435, 1992.

[4] Tin Kam Ho, *A Theory of Multiple Classifier Systems And Its Application to Visual Word Recognition*, Ph.d thesis, SUNY Buffalo, 1992.

[5] Tin Kam Ho, J.J. Hull, and S.N. Srihari, "Decision combination in multiple classifier systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 1, pp. 66–75, 1994.

[6] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 10, pp. 955–966, 1995.

[7] Lin Hong and Anil Jain, "Integrating faces and fingerprints for personal identification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 12, pp. 1295–1307, 1998.

[8] Afsar Saranli and Mubeccel Demirekler, "A statistical unified framework for rank-based multiple classifier decision combination," *Pattern Recognition*, vol. 34, no. 4, pp. 865–884, 2001.

[9] Anil Jain, Karthik Nandakumar, and Arun Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[10] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998.

[11] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

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

[13] Patrick Grother, "Face recognition vendor test 2002 supplemental report," Tech. Rep. NISTIR 7083, NIST, 2004.

[14] 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.

[15] A.E. 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*, 1996, vol. 1, pp. 81–84 vol. 1.

[16] 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.

[17] D. Bouchaffra, V. Govindaraju, and S. Srihari, "A methodology for mapping scores to probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, September 1999.

[18] Krassimir Ianakiev, *Organizing Multiple Experts for Efficient Pattern Recognition*, Ph.D Thesis, SUNY at Buffalo, 2000.