Score Fusion by Maximizing the Area Under the ROC Curve *

Mauricio Villegas and Roberto Paredes

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, Edif. 8G Acc. B 46022 Valencia (Spain)
{mvillegas,rparedes}@iti.upv.es

Abstract. Information fusion is currently a very active research topic aimed at improving the performance of biometric systems. This paper proposes a novel method for optimizing the parameters of a score fusion model based on maximizing an index related to the Area Under the ROC Curve. This approach has the convenience that the fusion parameters are learned without having to specify the client and impostor priors or the costs for the different errors. Empirical results on several datasets show the effectiveness of the proposed approach.

1 Introduction

Biometrics is currently a very active area of research due to the numerous applications it offers. By biometrics it is meant the automatic identification of a person by means of an anatomical or behavioral characteristic such as a facial image or a fingerprint. A method for improving the performance of biometric systems is to fuse information from several sources. This information can be fused at different levels, which can be at the sensor, feature, match score or decision levels. This work is focussed at match score fusion.

In the literature numerous score fusion approaches have been proposed. These can be categorized into two groups, which are the non-training based methods and the training based methods [1]. The non-training methods assume that the output of the individual matchers are the posterior probabilities that the pattern belongs to the claimed identity. Because this assumption is not generally true a previous normalization step is required [2]. The training based methods as the name suggest requires a training step. Among these are all of the methods which treat the fusion as a classification problem [3–5].

A significant drawback that the classification approach to score fusion has, is that these methods tend to minimize the classification error. However, the standard way of comparing biometric systems is by using a ROC curve. This way it is not necessary to specify which are the client and impostor priors or what are the costs for each of the possible errors of the system, values which are

* Work supported by the Spanish projects DPI2006-15542-C04 and TIN2008-04571 and the Generalitat Valenciana - Consellería d'Educació under an FPI scholarship.

difficult to estimate and vary depending on the application. From this perspective, minimizing the classification error may not improve the performance in the practice. The Area Under the ROC Curve (AUC) summarizes the ROC curve, and this can be a better measure for assessing biometric systems without having to specify the priors or costs [6]. Motivated by this idea, in this work we propose to learn the parameters of a score fusion model by maximizing the AUC. In the literature there are several works on maximizing the AUC, due to lack of space we limit ourselves to referencing a few [7–9].

The rest of the paper is organized as follows. The next section defines a score fusion model and derives an algorithm which optimizes the model parameters by maximizing the AUC. The experimental results are presented in section 3 and the final section draws the conclusions and directions for future research.

2 Score Fusion by Maximizing the AUC

As was explained in the previous section, the AUC is an adequate measure to assess the quality of a biometric system without having to specify the client and impostor priors or the costs for the different errors. Motivated by this evidence, we propose to derive an algorithm that learns the parameters of a score fusion model by maximizing the AUC. In order to do this we have to address two tasks, the first one is to define a model that fuses scores according to some parameters, and second is to optimize the parameters of the model so that the AUC is maximized.

To choose the model for score fusion we have taken into account the following criteria. The model should be capable of weighting the different scores giving more or less importance to each of them. Also, the model should be able to handle scores with arbitrary input ranges. Finally the model should have few parameters so that they can be well estimated evading the small sample size problem. A simple method that fulfills the previous requirements is to first normalize the scores so that they are all in a common range and afterwards combine linearly the normalized scores.

2.1 Score Normalization

In the literature several methods for score normalization can be found, for a review of the most used ones in biometric fusion refer to [2]. The normalization we have chosen is based on the *tanh-estimators* which is somewhat insensitive to the presence of outliers [2]. This normalization is a nonlinear transformation of the score using a sigmoid function and it depends on two parameters, the sigmoid slope and its displacement. The slope determines how fast is the transition from zero to one, and the displacement indicates at what value the sigmoid is in the midpoint of the transition. The sigmoid normalization is given by

$$\phi_{u,v}(z) = \frac{1}{1 + \exp[u(v - z)]} \ , \tag{1}$$

where u and v are the slope and the displacement of the sigmoid respectively.

2.2 Score Fusion Model

To be able to represent the model mathematically first we need to state some definitions. Let $\mathbf{z}$ be an M-dimensional vector composed of the M scores we want to fuse $z_1, \ldots, z_M$. Furthermore let $\Phi_{\mathbf{u},\mathbf{v}}(\mathbf{z})$ be a vector composed of the normalized scores $\phi(z_1, u_1, v_1), \ldots, \phi(z_M, u_M, v_M)$ and the vectors $\mathbf{u}$ and $\mathbf{v}$ be a more compact representation of the sigmoid slopes $u_1, \ldots, u_M$ and displacements $v_1, \ldots, v_M$. As mentioned earlier, the model is a linear combination of the normalized scores, then we denote the score weights by $w_1, \ldots, w_M$ which are also represented more compactly by the vector $\mathbf{w}$.

The input scores can be either similarity or distance measures, however the sigmoid normalization can transform all of the scores to be similarity measures by having a positive or negative slope. Given that all the normalized scores are similarity measures, if they contain discriminative information then they should have a positive contribution to the final score, otherwise they should not have any contribution. Therefore without any loss of generality we can restrict the weights to being positive, $w_m \geq 0$ for $m = 1 \ldots M$. On the other hand, scaling the fused score does not have any effect on its discrimination ability, thus we can further restrict the weights so that their sum equals the unity, $\sum_{m=1}^{M} w_m = 1$. Note that given the two restrictions, the fused score has the nice property of being a value between zero and one.

The score fusion model is given by:

$$f_{\mathbf{u},\mathbf{v},\mathbf{w}}(\mathbf{z}) = \mathbf{w}^T \Phi_{\mathbf{u},\mathbf{v}}(\mathbf{z}) \ . \tag{2}$$

The parameters of the model are $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^M$, which means that in total there are $3M$ parameters that need to be estimated.

2.3 AUC Maximization

Although there are few parameters to be estimated in the score fusion model (2), it can be highly computationally expensive to obtain an adequate estimation and clearly brute force is not advisable. Therefore our aim is an index that is directly related to the AUC and use an optimization procedure to maximize it.

Among the different alternatives to compute the AUC the one that lends itself for the simplest optimization process is the one known as the Wilcoxon-Mann-Whitney statistic:

$$\mathcal{A} = \frac{1}{PN} \sum_{p=1}^{P} \sum_{n=1}^{N} \mathcal{H}(x_p - y_n) \ , \tag{3}$$

where P and N are the number of client and impostor samples respectively, and $\mathcal{H}()$ is the Heaviside step function which has a value of zero or one for negative and positive numbers respectively, and a value of $1/2$ at zero.

The expression in equation (3) is not differentiable, therefore inspired on the same ideas as in [10, 11], the Heaviside step function can be approximated using

a sigmoid function. Doing this approximation and using the score fusion model (2), leads to the following optimization index:

$$J(\mathbf{u},\mathbf{v},\mathbf{w}) = \frac{1}{PN}\sum_{p=1}^{P}\sum_{n=1}^{N} S_\beta\Big(\mathbf{w}^T\big(\hat{\mathbf{x}}_p - \hat{\mathbf{y}}_n\big)\Big) \ , \tag{4}$$

where the hat indicates that the sore is normalized, i.e. $\hat{\mathbf{x}}_p = \Phi_{\mathbf{u},\mathbf{v}}(\mathbf{x}_p)$ and $\hat{\mathbf{y}}_n = \Phi_{\mathbf{u},\mathbf{v}}(\mathbf{y}_n)$, and the sigmoid function is defined by

$$S_\beta(z) = \frac{1}{1+\exp(-\beta z)} \ . \tag{5}$$

Care must be taken not to confuse this sigmoid function, which is used for AUC maximization, with the sigmoid used for score normalization from equation (1).

To maximize the index (4) we propose to use a batch gradient descent procedure. To this end, we take the partial derivatives of the index with respect to the parameters obtaining:

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{u}} &= \mathbf{w} \bullet \frac{1}{PN}\sum_{p=1}^{P}\sum_{n=1}^{N} S'_\beta\Big(\mathbf{w}^T\big(\hat{\mathbf{x}}_p - \hat{\mathbf{y}}_n\big)\Big)\bullet \\
&\qquad \Big((\mathbf{x}_p - \mathbf{v}) \bullet \Phi'(\mathbf{x}_p) - (\mathbf{y}_n - \mathbf{v}) \bullet \Phi'(\mathbf{y}_n)\Big) \ ; \\
\frac{\partial J}{\partial \mathbf{v}} &= \mathbf{w} \bullet \mathbf{u} \bullet \frac{1}{PN}\sum_{p=1}^{P}\sum_{n=1}^{N} S'_\beta\Big(\mathbf{w}^T\big(\hat{\mathbf{x}}_p - \hat{\mathbf{y}}_n\big)\Big)\bullet \\
&\qquad \Big(\Phi'(\mathbf{y}_n) - \Phi'(\mathbf{x}_p)\Big) \ ; \\
\frac{\partial J}{\partial \mathbf{w}} &= \frac{1}{PN}\sum_{p=1}^{P}\sum_{n=1}^{N} S'_\beta\Big(\mathbf{w}^T\big(\hat{\mathbf{x}}_p - \hat{\mathbf{y}}_n\big)\Big)\Big(\hat{\mathbf{x}}_p - \hat{\mathbf{y}}_n\Big) \ .
\end{aligned} \tag{6}$$

The big dot $\bullet$ indicates a Hadamard or entry-wise product, $S'_\beta()$ is the derivative of the sigmoid function (5) and the elements of the vectors $\Phi'(\mathbf{x}_p)$ and $\Phi'(\mathbf{y}_n)$ are given by

$$\phi'(z) = \frac{\exp[u(v-z)]}{(1+\exp[u(v-z)])^2} \ . \tag{7}$$

Finally the corresponding gradient ascend update equations are

$$\mathbf{s}^{(t+1)} = \mathbf{s}^{(t)} + \gamma \frac{\partial J}{\partial \mathbf{s}}^{(t)} \ , \tag{8}$$

where $\mathbf{s} = \{\mathbf{u},\mathbf{v},\mathbf{w}\}$ and γ is the learning rate. After each iteration the weights are renormalized so that the restrictions of being positive and sum to unity are met.

This approach to maximization of the AUC has been previously mentioned in the work of Yan et al. [7], however they report having significant numerical problems for values of $\beta > 2$, in which case the sigmoid function is a poor estimate of the step function. Our experience differs completely from this notion, being the optimization quite stable for higher values of β on several data sets.

2.4 Notes on the Implementation of the Algorithm

The algorithm has a very high computational cost which makes it unpractical for large datasets. However there are several approaches that can be used to speed up the computation without sacrificing performance. For most of the client and impostor score pairs the derivative of the sigmoid function is practically zero. Therefore in each iteration a large amount of pairs can be discarded depending on their relative difference. Another approach could be to use the stochastic gradient ascend instead of the batch. This reduces significantly the amount of iterations that the algorithms needs to converge.

2.5 Extensions of the Algorithm

An initial clarification must be made. Although in this paper a score fusion model is defined and optimized, the maximization by AUC is a general approach which can be applied to other models and other problems different from score fusion. Furthermore the proposed score fusion model is very simple and linear and therefore it is unable to handle complex distributions. Depending on the problem, improvements to the model must be made.

Up to this point, the proposed model has very few parameters, and it is a simple linear combination of normalized scores. The algorithm can be extended to be nonlinear, it is as straight forward as adding new virtual scores which are a nonlinear combination of the original scores. This nonlinear extension can also be useful to increase the number of parameters of the score fusion model and thus it increases its representation capability.

Along with the research on biometric score fusion there is another related topic. This topic is the use of quality measures to determine how confident a biometric score is. This information can greatly improve the recognition accuracy of the systems if they are taken into account during the fusion. An approach to integrate the quality measures into the proposed model could be to include these values as if they were other scores like it is done in [12]. However the quality values can mean different things under different circumstances [5], making this approach unsatisfactory. A simple and better approach would be to include the quality measures as scores but removing the restriction of the weight being positive. This way the quality can reward or penalize the final score depending on the circumstance.

3 Experiments

The proposed approach was evaluated using three publicly available datasets. The first dataset was the LP1 set of scores obtained from the XM2VTS face and voice multimodal database [13]. This dataset includes eight biometric scores per claim, five for face images and the remaining three for speech. The experimentation protocol for this dataset is clearly defined, first there is an evaluation set, which is used to optimize the fusion parameters, and then there is a test set which is used to assess the performance. In total the evaluation set has 600 client and 40k impostor claims, and the test set has 400 client and around 112k impostor claims.

The other two datasets used in the experiments were the Multimodal and the Face datasets from the NIST Biometric Scores Set - Release 1 (BSSR1) [14]. The Multimodal dataset is composed of four scores per claim, two correspond to face matchers and the other two to the right and left index fingerprints for the same matcher. This dataset has 517 client and around 267k impostor claims. Finally the Face dataset is composed of two scores per claim, each one for a different face matcher. In this case there are 6k client and 1.8M impostor claims. For these datasets there is no experimentation protocol defined. In our experiments we did a repeated hold-out using half of the data for training and the other half for test, and repeated 20 times.

The results of the experiments for the test sets are summarized in the table 1. For each dataset three results are presented. The first one is for the single matcher which obtained the best result without doing score fusion. The second result is the best one obtained by trying among several baseline techniques. The baseline techniques tried were the sum, product, min and max rules, each one either with z-score or maxmin normalization. The final result is for our technique (SFMA). For each dataset and method the table presents three performance measures, the AUC given as a percentage of the total area, the Equal Error Rate (EER) and the Total Error Rate at a False Acceptance Rate of 0.01% (TER@FAR=0.01%). The 95% confidence intervals are included for the BSSR1 datasets.

On biometric research papers it is common to plot either a ROC or a DET curve to compare various systems. Nonetheless these curves do not take into account how the thresholds are selected, making the comparison of systems somewhat unreliable. In this paper we have opted to use the Expected Performance Curves (EPC) [15], which plots the HTER using a threshold ($\hat{\theta}_\alpha$) obtained on a development set by $\arg\min(\theta_\alpha) = \alpha\text{FAR} + (1-\alpha)\text{FRR}$. The parameter α is a value between zero and one which weights the importance of the errors. The EPC curves for two of the datasets are presented in figure 1.

The results for the proposed technique are very promising. In all of the datasets SFMA improves the AUC even though the maximization was done on the training set, this suggests that the technique has good generalization capability. Only the TER@FAR=0.01% for the BSSR1 Face dataset is slightly worse than Sum Rule with z-score, however this is an extreme operating point. For this dataset the improvement is significant for a wide range of operating thresholds as can be observed on its EPC curve.

Dataset	Method	AUC (%)	EER (%)	TER@ FAR=0.01% (%)
XM2VTS (LP1)	Best Matcher	99.917	1.14	15.0
	Sum Rule/z-score	99.973	0.56	3.0
	SFMA	**99.997**	**0.28**	**1.0**
BSSR1 (Multimodal)	Best Matcher	98.84 ±0.10	4.67 ±0.23	26.8 ±1.07
	Sum Rule/z-score	**99.99** ±0.00	**0.50** ±0.07	3.2 ±0.41
	SFMA	**99.99** ±0.00	**0.50** ±0.18	**1.5** ±0.25
BSSR1 (Face)	Best Matcher	65.39 ±1.07	5.26 ±0.05	28.9 ±0.27
	Sum Rule/z-score	98.62 ±0.03	5.09 ±0.03	**24.2** ±0.31
	SFMA	**99.07** ±0.03	**4.25** ±0.05	25.3 ±0.33

Table 1. Summary of score fusion results on different datasets.

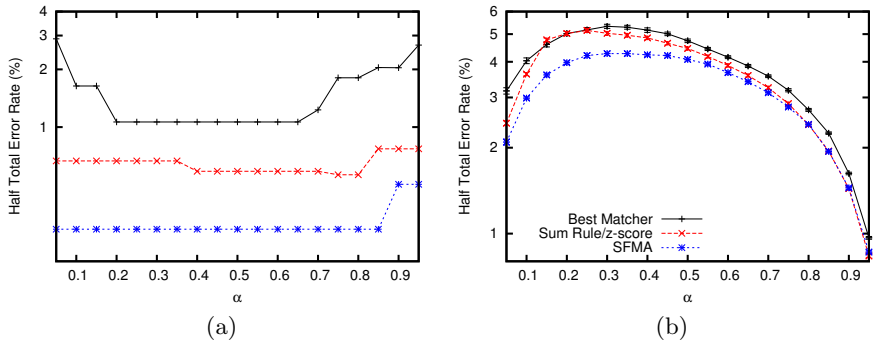

Fig. 1. Expected Performance Curves of the fusion algorithms for (a) XM2VTS LP1 and (b) BSSR1 Face.

4 Conclusions

This paper presented a novel method for optimizing the parameters of a score fusion model based on maximizing an index related to the Area Under the ROC Curve (AUC). A score fusion model based on a linear combination of normalized scores was chosen and the AUC optimization procedure was derived for it.

The proposed algorithm was empirically evaluated using three publicly available datasets, the XM2VTS LP1, the BSSR1 Multimodal and the BSSR1 Face. The results show that the technique works as expected. The AUC is iteratively improved by the algorithm and the result generalizes well to new data. Also, by maximizing the AUC, specific operating points on the ROC curve also improve without having to choose which one will be used in the final system.

Several research topics are left for future work. One topic is to analyze the computational cost of the algorithm. To speedup the algorithm some approximations can be made and a stochastic gradient ascend procedure can be employed.

Therefore the question remains about how much time is required by the algorithm and how much do the approximations affect the results. For future work also is how to integrate quality measures into the score fusion model and how does the algorithm perform using this type of information. Another topic to work on is using the AUC maximization for other problems, such as biometric verification or biometric sample quality estimation.

References

1. Toh, K.A., Kim, J., Lee, S.: Biometric scores fusion based on total error rate minimization. Pattern Recognition **41**(3) (2008) 1066–1082
2. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition **38**(12) (December 2005) 2270–2285
3. Gutschoven, B., Verlinde, P.: Multi-modal identity verification using support vector machines (svm). Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on **2** (July 2000) THB3/3–THB3/8 vol.2
4. Ma, Y., Cukic, B., Singh, H.: A classification approach to multi-biometric score fusion. In: AVBPA. (2005) 484–493
5. Maurer, D.E., Baker, J.P.: Fusing multimodal biometrics with quality estimates via a bayesian belief network. Pattern Recogn. **41**(3) (2008) 821–832
6. Ling, C.X., Huang, J., Zhang, H.: Auc: a statistically consistent and more discriminating measure than accuracy. In: Proc. of IJCAI 2003. (2003) 519–524
7. Yan, L., Dodier, R.H., Mozer, M., Wolniewicz, R.H.: Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), Washington, DC, USA, AAAI Press (August 2003) 848–855
8. Marrocco, C., Molinara, M., Tortorella, F.: Exploiting auc for optimal linear combinations of dichotomizers. Pattern Recogn. Lett. **27**(8) (2006) 900–907
9. Marrocco, C., Duin, R.P.W., Tortorella, F.: Maximizing the area under the roc curve by pairwise feature combination. Pattern Recogn. **41**(6) (2008) 1961–1974
10. Paredes, R., Vidal, E.: Learning prototypes and distances: a prototype reduction technique based on nearest neighbor error minimization. Pattern Recognition **39**(2) (2006) 180–188
11. Villegas, M., Paredes, R.: Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (2008) 1–8
12. Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.: Likelihood ratio-based biometric score fusion. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(2) (Feb. 2008) 342–347
13. Poh, N., Bengio, S.: A score-level fusion benchmark database for biometric authentication. In: AVBPA. (2005) 1059–1070
14. National Institute of Standards and Technology: NIST Biometric Scores Set - Release 1 (BSSR1). http://www.itl.nist.gov/iad/894.03/biometricscores/ (2004)
15. Bengio, S., Mariéthoz, J., Keller, M.: The expected performance curve. In: Proceedings of the Second Workshop on ROC Analysis in ML. (2005) 9–16