Improving Fusion with Margin-Derived Confidence In Biometric Authentication Tasks

Norman Poh and Samy Bengio

IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
norman@idiap.ch, bengio@idiap.ch

Abstract. This study investigates a new *confidence criterion* to improve fusion via a linear combination of scores of several biometric authentication systems. This confidence is based on the margin of making a decision, which answers the question, "after observing the score of a given system, what is the confidence (or risk) associated to that given access?". In the context of multimodal and intramodal fusion, such information proves valuable because the margin information can determine which of the systems should be given higher weights. Finally, we propose a *linear discriminative framework* to fuse the margin information with an existing *global* fusion function. The results of 32 fusion experiments carried out on the XM2VTS multimodal database show that fusion using margin (product of margin and expert opinion) is superior over fusion without the margin information (i.e., the original expert opinion). Furthermore, combining both sources of information increases fusion performance further.

1 Introduction

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. Compared to traditional authentication methods such as keys and PIN numbers, biometric authentication has the advantages that it is not susceptible to misplacement or forgetfulness. Unfortunately, its accuracy and reliability still need to be improved to make the system practical in day-to-day applications.

One way to increase its performance accuracy is to combine several biometric systems. In this paper, we show how multimodal or intramodal fusion BA system can be improved by using a new confidence measure based on margin. This quantity can be interpreted as "how confident we are that a given access is correct after observing the score". It is bounded between zero and one; when it is zero, a given access has 50% chance of being correctly classified. The greater the confidence, the higher the chance that the given access is correct. We show that this margin-derived confidence can be used in fusion of multimodal biometric systems. The margin-derived confidence can be used to *modify* the fixed decision boundary. This is done by a linear combination between the confidence-derived function and the fixed discriminative function. The former function is *adaptive*, i.e., it changes *after* observing the access scores. In contrast, the latter function is *fixed* once (hence non-adaptive) and applied to all accesses.

Improving fusion with quality has already been examined by several authors. Toh *et al.* [1] fused fingerprint and speech systems using a modified multivariate polynomial regression function to take the quality information into account. Bigun *et al.* [2] also fused fingerprint and speech systems but using a statistical model (that reconciles expert opinions) modified to take the quality into account. Fierrez-Aguilar [3] fused fingerprint and speech systems, with quality derived from fingerprint, using a modified Support Vector Machine algorithm. Garcia-Romero *et al.* [4] considered quality in speaker authentication task using the first formant. Fusion is done so as to favour speech frames with high quality. Hence, instead of taking the average Log-Likelihood Ratio (LLR) over the entire utterance frames, a weighted LLR (by quality) is used. All these studies provide empirical evidences that *quality information can improve the performance* of single-modal and multimodal biometric systems.

We propose to derive a quality index based on margin. This margin is a function of False Acceptance and False Rejection Rates, which themselves are estimated from a set of expert scores. The main advantage of margin-derived quality is that no additional (and often independent) system is needed to estimate the quality, as compared to the previously mentioned approaches[1].

Section 2 presents the proposed idea of margin and compares it with existing margin definitions in the literature. Section 3 presents how confidence can be integrated with existing fusion functions. Section 4 presents briefly the 32 fusion problems based on the XM2VTS database and Section 5 discusses a pooled EPC curve as a performance visualisation tool. Experiments are reported in Section 6. This is followed by conclusions in Section 7.

2 Margin As Confidence

Given an acquired biometric feature $\mathbf{x}$, an opinion of a BA system $y(\mathbf{x})$ as a function of $\mathbf{x}$ and a preset threshold Δ, a biometric system makes its decision based on the following decision function:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise.} \end{cases} \tag{1}$$

Since $\mathbf{x}$ is present in $y(\mathbf{x})$ and variables derived from it, we simply write y instead of $y(\mathbf{x})$. The system may make two types of mistakes: false acceptance (FA) and false rejection (FR) as a function of threshold Δ. By tracing this function empirically from a development set, and normalising them using the total number of impostor and client accesses, respectively, one obtains the false acceptance rate (FAR) and false rejection rate (FRR) curve as a function of threshold Δ. FAR and FRR are defined as follows:

$$\text{FAR}(\Delta) = \frac{\text{number of FAs}(\Delta)}{\text{number of impostor accesses}} , \tag{2}$$

$$\text{FRR}(\Delta) = \frac{\text{number of FRs}(\Delta)}{\text{number of client accesses}} . \tag{3}$$

[1] The additional measurement system *may* provide additional degree of freedom to describe the biometric classes if the system output is *independent* of the original feature sets. However, in most situations, the additional system derives the quality information from the same feature sets as those used by the verification system, e.g., [1, 2]. Regardless of how the quality information is derived (from the feature sets or from the scores as proposed here), we conjecture that the quality information can provide better information regarding the separation decision.

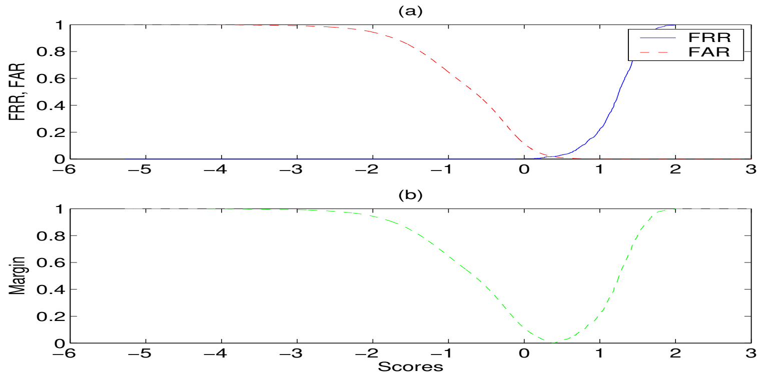


Fig. 1. (a) FAR and FRR as a function of the threshold in the score space. (b) The derived margin based on (a).

A commonly used point to examine the quality of performance is to evaluate the value FAR = FRR. This is the Equal Error Rate (EER) point and it assumes that the costs of FA and FR are equal, and that the class prior probabilities (of client and impostor distributions) are also equal.

The empirical procedure to find Δ that satisfies the EER criterion (on the training set) is:

$$\Delta^* = \arg\min_{\Delta} |\mathrm{FAR}(\Delta) - \mathrm{FRR}(\Delta)| \,. \tag{4}$$

We define the margin as:

$$\mathcal{M}(\Delta) = |\mathrm{FAR}(\Delta) - \mathrm{FRR}(\Delta)|. \tag{5}$$

By replacing Δ by y, we effectively evaluate the margin of the output y. FAR, FRR and margin are shown in Figure 1. The margin derived this way simply tells us how much confident we are given an opinion y. The further it is from the decision boundary Δ^*, the more confident we are. Note that because FAR and FRR are cumulative density functions, they are confined in the range $[0, 1]$. Hence, the margin defined here is also confined in the range $[0, 1]$. The additional scores that are needed to derive the margin function can either be obtained from *additional biometric data* or *cross-validated data* (not used to train the underlying systems) in case the additional data is not available.

Note that the margin defined here is different from the concept of margin in the boosting [5] or Vapnik's *margin slack variable* [6]. Several definitions of margin are defined in [7, Sect. 2]. Suppose that the target output is t_p and the output of a system is y_p for the p-th example. t_p takes on $\{-1, 1\}$, each representing a class (impostor or client here). Using this notation, margin in boosting for a given example p is:

$$margin(y_p) = \underbrace{(y_p - \Delta^*)}\, t_p, \tag{6}$$

whereas, Vapnik's margin slack variable for a given example p is:

$$\xi_p = \max(0, \gamma - \underbrace{(y_p - \Delta^*)}\, t_p), \tag{7}$$

where $\gamma > 0$ is known as *target margin* and is fixed *a priori*. Note that in our notation, the subtraction in the underbraced term $y_p - \Delta^*$ is to make sure that the decision boundary has a value of 0 (normally, the Δ^* has already been absorbed by the output of the system as a bias term; in our context, this bias term corresponds to $-\Delta^*$). Briefly, $margin(y_p)$ measures how far an example is from the decision boundary. The further it is, the better. Negative margin in this case implies wrong classification of example p. In Vapnik's margin, ξ_p measures how much example p fails to have a margin of γ from the hyperplane. If $\xi_p > \gamma$ then example p is misclassified by $y_p - \Delta^*$. The difference between Vapnik's margin slack variable and margin in boosting is that the former takes the target margin into account whereas the latter does not. Both of these margin definitions can only be calculated supposing that the target output (class-label) is known. In fact, they are used to select examples that are difficult to classify. They are only important during the training phrase. Our proposed definition of margin *does not* require the target output (although the margin function is constructed from a labeled training set). Furthermore, it is used exclusively during testing, which differs from the rest of the margin definitions. Perhaps the most remarkable difference is that this margin is based on FAR and FRR, with minimum at EER. The aforementioned margins are also valid but they do not optimise EER directly. Despite their different usages, one similarity among all these margins is that they all have to be derived from labeled (training) data.

In the next section, we will propose a method to incorporate the margin-derived confidence measure into an existing fusion function.

3 Combining *a priori* Weights with Confidence

3.1 General Fusion Function

The most used form of fusion function in biometric authentication is perhaps a linear combination of several expert opinions passed through an activation function. Suppose y'_j is the j-th opinion and α_j is the weight associated to y'_j, respecting the constraint that $\sum_j \alpha_j = 1$. The combined opinion of M base experts, y_{COM} can be written as:

$$y_{COM} = f\left(\sum_{j=1}^{M} \alpha_j y'_j\right) \tag{8}$$

where f is an activation function. Suppose that there are N biometric systems but there are $M \geq N$ opinions. The number of opinions can be more than the number of systems because we assume here that each system can give more than one opinion, derived in one way or another. For instance, for the case of fusing two systems with output y_1 and y_2, we could have:

$$y'_j \in \{y_1, y_2, y_1^2, y_2^2, y_1 y_2, 1\}, \tag{9}$$

where 1 is a bias term, and

$$f(z) = \frac{1}{1 + \exp[-a(z-b)]}, \tag{10}$$

which yields a *polynomial logistic regression* function (with $a = 1, b = 0$). The full expansion of polynomial is exponential with respect to its degree. In [8], a reduced polynomial expansion is used to reduce the complexity (the degree of freedom of the classifier) and to make it practical enough for fusion problems. When y'_j is defined as:

$$y'_j \in \{y_i | i = 1, \ldots, N\} \tag{11}$$

and using Eqn. (10) with $a = 1, b = 0$, one obtains a *logistic regression* function [9] In this study, we concentrate on the linear function f, i.e., $f(z) = z$ (a linear function) and establish a means to combine margin-derived confidence with a fixed discriminative function. We will show how the form of fusion in Eqn. (8) occurs naturally.

3.2 Fusion Function With Quality

In the literature, to the best of our knowledge, there are two forms to integrate the quality information with an *a priori* weight that modifies α_i in Eqn. (8). Suppose that w_j is the *a priori* weight (found by optimising Equal Error Rate, for instance) and q_j is the quality associated to y'_j. The two forms that incorporate the quality information are as follow:

$$\alpha_j \propto w_j + q_j \tag{12}$$

and

$$\alpha_j \propto w_j \times q_j \tag{13}$$

Note that in the absence of the quality information, we have $\alpha_j \propto w_j$. The usage of Eqn. (12) can be found in [1] using a reduced polynomial expansion of logistic regression function, i.e., using Eqn. (9) for the case of polynomial degree 2 and Eqn. (10). In the mentioned work, only polynomial up to degree 3 was examined. Experiments were conducted on fusion of fingerprint and speech biometrics with quality information obtained only from the fingerprint.

The usage of Eqn. (13) was found in [10, 11]. In [10], a speech expert ($j = 1$) and a lip expert ($j = 2$) were fused. Suppose that y_j^k is the j-th opinion given that the access is $k = \{C, I\}$, i.e., client or impostor. Suppose that y_j^k is generated from a normal distribution with mean μ_j^k and variance $(\sigma_j^k)^2$, i.e., $y_j^k \sim \mathcal{N}\left(\mu_j^k, (\sigma_j^k)^2\right)$. In [10], w_1 is defined as:

$$w_1 = \frac{\zeta_2}{\zeta_1 + \zeta_2} \tag{14}$$

where,

$$\zeta_j = \sqrt{\frac{(\sigma_j^C)^2}{NC} + \frac{(\sigma_j^I)^2}{NI}} \tag{15}$$

and NC is the total number of client accesses and NI is the total number of impostor accesses. By the summation constraint, $w_2 = 1 - w_1$. ζ_j is called the standard error. In [10], it was assumed that this error gives relative discrimination of an expert. High ζ_j indicates that expert j has high class dependent variance and hence, lower performance. As a result, its weight is lowered and the other expert's weight is increased[2]. q_j is defined as:

$$q_j \propto |\mathcal{M}_j^C(y_j) - \mathcal{M}_j^I(y_j)|, \tag{16}$$

[2] Although this criterion is valid, examining class-dependent variance is not sufficient; the mean difference is an important factor [12].

where

$$\mathcal{M}_j^k(y_j) = \frac{(y_j - \mu_j^k)^2}{(\sigma_j^k)^2} \tag{17}$$

for $k = \{C, I\}$ and $\sum_j q_j = 1$. Note that in this context, only the speech expert ($j = 1$) can be corrupted by noise whereas the lip expert ($j = 2$) stays intact. It was demonstrated experimentally [10] that under clean conditions, q_1 is relatively large (as compared to q_2) whereas under noisy conditions, q_1 is relatively small.

In [11], face and speech experts are fused and the speech expert is susceptible to noise whereas the face expert remains intact. The quality of the speech signal is estimated by using a statistical model (Gaussian Mixture Model) from the unvoiced part of speech frames. The unvoiced part of speech was obtained from the speech features right before an utterance begins. The output of the model (Log-Likelihood Ratio, LLR) is normalised into the range $[0, 1]$ by using a sigmoid function, as shown in Eqn. (10). a and b were tuned by heuristics, such that q_j is close to one for good quality speech and close to 0 for bad quality speech. According to the authors, the likelihood normalisation step is necessary because the normalised LLR is used directly to influence the *a priori* weight. $w_j|\forall_j$ are estimated using standard methods to minimise Equal Error Rate (EER), to be discussed in the later section.

We will use the method in Eqn. (12) because, as will be shown, it can be used to fuse different information sources. Furthermore, the multiplicative effect in Eqn. (13) can adversely influence α_j drastically as compared to Eqn. (12). To begin with, we consider a linear function of f, i.e., $f(z) = z$. We wish to fuse existing weight w_i with quality q_i for all $i = 1, \ldots, N$. Hence, α_i can be written as:

$$\alpha_i = \beta_{1,i} w_i + \beta_{2,i} q_i \tag{18}$$

where β_i control the contribution between the *a priori* weight w_i and the quality information q_i. Using $f(z) = z$, Eqns. (8) and (18), we obtain:

$$\begin{aligned} y_{COM} &= \sum_i (\beta_{1,i} w_j + \beta_{2,i} q_i) y_i \\ &= \sum_{m=1}^{N} \Big(\underbrace{\beta_{1,m} w_m}\, \underbrace{y_m} \Big) + \sum_{n=1}^{N} \Big(\underbrace{\beta_{2,n}}\, \underbrace{q_n y_n} \Big) \end{aligned} \tag{19}$$

where the four under-braces in Eqn. (19) can be written in the form of Eqn. (8). with y'_j defined by:

$$y'_j \in \{y_i, q_i y_i | i = 1, \ldots, N\}$$

Hence, fusion of *a priori* weight with the quality information can be performed by a linear combination of y_i and $q_i y_i$, for all i. The corresponding weights α_j can be found using standard methods such as Fisher-ratio or linear regression. The use of non-linear solutions is direct. For instance, one can use a Multi-Layer Perceptron with $y'_j|\forall_j$ as an input vector. Standard Support Vector Machine (SVM) algorithm with a polynomial kernel can also be used to classify the secondary features, thus, eliminating the need to create a dedicated classifier to fuse the quality information, as in [1] or to apply heuristics, as in [10, 11].

4 Database

The XM2VTS database [13] contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence. The database is divided into three sets: a training set, an evaluation set and a test set. The training set was used to build client models, while the evaluation set was used to compute the decision thresholds as well as other hyper-parameters used by classifiers and normalisation. Finally, the test set was used to estimate the performance. The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II (LP1 and LP2). The most important thing to note here is that there are only 3 samples in LP1 and 2 samples in LP2 for client-dependent adaptation and fusion training. Instead of reimplementing base experts and applying them on this database, we used scores from [14]. The score files are made publicly available and are documented in [15][3]. There are altogether 7 face experts and 6 speech experts for LP1 and LP2, respectively. By combining 2 baseline experts at a time according multimodal or intramodal fusion problems, 32 fusion experiments are further identified. The 13 baseline experiments have 400 $\times$ 13 = 5,200 client accesses and 111,800 $\times$ 13 = 1,453,400 impostor accesses. The 32 fusion experiments have 400 $\times$ 32 = 12,800 client accesses and 111,800 $\times$ 32 = 3,577,600 impostor accesses.

5 Evaluation Using Pooled EPC Curves

Perhaps the most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [16]. It has been pointed out [17] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [17] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [17] was proposed. We will adopt this evaluation method, which is also in coherence with the original Lausanne Protocols defined for the XM2VTS database. The criterion to choose an optimal threshold is called weighted error rate (WER), defined as follows:

$$\mathrm{WER}(\alpha, \Delta) = \alpha \mathrm{FAR}(\Delta^*) + (1-\alpha)\, \mathrm{FRR}(\Delta^*), \tag{20}$$

where FAR and FRR are False Acceptance Rate and False Rejection Rate, respectively. Note that WER is optimised for a given $\alpha \in [0,1]$. Let Δ^*_α be the threshold that *minimises* WER on a *development set*. The performance measure tested on an *evaluation set* at a given Δ^*_α is called Half Total Error Rate (HTER), which is defined as:

$$\mathrm{HTER}(\alpha) = \frac{\mathrm{FAR}(\Delta^*_\alpha) + \mathrm{FRR}(\Delta^*_\alpha)}{2}. \tag{21}$$

The EPC curve simply plots HTER versus α, since different values of α give rise to different values of HTERs. The EPC curve can be interpreted in the same manner as the DET

[3] Accessible at http://www.idiap.ch/$\sim$norman/fusion

curve, i.e., the lower the curve is, the better the performance but for the EPC curve, the comparison is done at a given cost (controlled by α). Furthermore, one can plot a pooled EPC curve from several experiments. For instance, in order to compare two methods over M experiments, only one pooled curve is necessary. This is done by calculating HTER at a given α point by taking into account all the false acceptance and false rejection accesses over all M experiments. The pooled FAR and FRR across $j = 1, \ldots, M$ experiments for a given $\alpha \in [0, 1]$ is defined as follow:

$$\text{FAR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \text{FA}(\Delta_\alpha^*(j))}{NI \times M}, \tag{22}$$

and

$$\text{FRR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \text{FR}(\Delta_\alpha^*(j))}{NC \times M}, \tag{23}$$

where $\Delta_\alpha^*(j)$ is the optimised threshold at a given α, NI is the number of impostor accesses and NC is the number of client accesses. FA and FR count the number of false acceptance and the number of false rejection at a given threshold $\Delta_\alpha^*(j)$. The pooled HTER is defined similarly as in Eqn. (21).

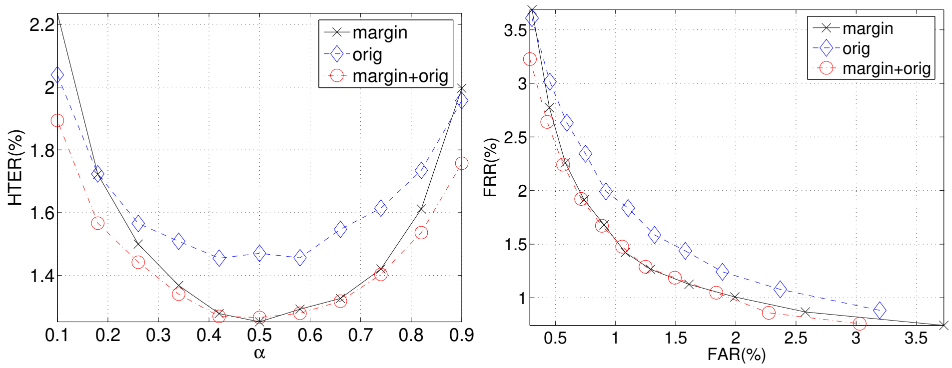
6 Experimental Results

Figure 2 shows both pooled EPC and ROC curves calculated from all 32×3 fusion experiments using original expert opinion ($y'_j \in \{y_i|\forall_i\}$), margin ($y'_j \in \{\mathcal{M}(y_i)y_i|\forall_i\}$) and both ($y'_j \in \{y_i, \mathcal{M}(y_i)y_i|\forall_i\}$). The ROC curves were plotted using FAR and FRR defined in Eqns. (22 and 23), whose *common* threshold was adjusted on a development (training) set. Note that for all these experiments, $\alpha_j|\forall_j$ were set to be equal. This reduces the fusion into the mean operator[4]. As can be seen, fusion with margin is better than the one using only the original expert opinions. Combining the two actually improves the performance even further. In fact, this improvement is significantly better than fusion using the original expert opinions across different α values according to the HTER significant test [18] with 95% of confidence. As a control experiment, we also performed fusion with $y'_j \in \{y_i, \mathcal{M}(y_i)|\forall_i\}$ using weighted sum. As expected, this approach does not improve the performance because $\mathcal{M}(y_i)$ does not contain any discriminative information. As a result, this control experiment is worse than using $y'_j \in \{y_i|\forall_i\}$ with EPC ranging between 1.5% and 3% of HTER (not shown here).

7 Conclusion

In this study, we proposed to use margin as a measure of confidence. When fusing two system opinions, their derived margins provide a relative information to which system is more important. This margin definition has the property that it is confined in the range $[0, 1]$, because it is derived from the distance between two cumulative density functions. Hence, margin can be used as a quality index. To the best of our knowledge, using margin to boost fusion has not been found in the literature yet. The second contribution of this work is the analysis of fusion function and how the quality information can be integrated with *a priori*

[4] In this database, weighted sum fusion with weights optimised using Fisher-ratio did not provide better performance than the mean operator.



(a) EPC (b) ROC

Fig. 2. Pooled (a) EPC and (b) ROC curves of fusion experiments using original expert opinion (labeled as "orig"), product of expert opinion with margin (labeled as "margin"), and combination of both information (labeled as "margin+orig"), all using the mean operator. According to the HTER significant test, the "margin+orig" curve is always better than the "orig" curve, at different α, at 95% of confidence. These experiments were carried out on the XM2VTS database using 32 intramodal and multimodal fusion datasets, and each dataset contains the scores of two experts. Note that both (a) EPC and (b) ROC curves are *consistent* in that "margin+orig" is the lowest curve (for EPC) or closest to the origin (for ROC), implying the best generalisation performance among the three curves.

weights of an existing fusion function. Suppose that y_i is the i-th opinion of an expert system and q_i is the associated quality. The fusion problem now can be treated as a fusion of $\{y_i, q_i y_i | \forall_i\}$. This has the same effect as modifying the *a priori* weight by adding q_i directly. 32×3 intramodal and multimodal fusion experiments were carried out on the XM2VTS multimodal database. Using pooled EPC curves (which summarise over each of the 32 experiments), we show that fusion using the confidence enhanced opinion $y_i q_i$ is better than using the original opinion y_i. Furthermore, combining the two, i.e., $\{y_i, y_i q_i\}$ improves the performance even further, and significantly, over different operating costs.

8 Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

1. K-A. Toh, W-Y. Yau, E. Lim, L. Chen, and C-H. Ng., "Fusion of Auxiliary Information for Multimodal Biometric Authentication," in *Springer LNCS-3072, Int'l Conf. on Biometric Au-*

thentication (ICBA), Hong Kong, 2004, pp. 678–685.
2. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communnications," in *12th Int'l Conf. on Image Analysis and Processing*, Mantova, 2003, pp. 2–11.
3. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-Based Multimodal Biometric Verification Using Quality Signals," in *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, 2004, vol. 5404, pp. 544–554.
4. D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the Use of Quality Measures for Text Independent Speaker Recognition," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 105–110.
5. Y. Freund and R. Schapire, "A Short Introduction to Boosting," *J. Japan. Soc. for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.
6. V. N. Vapnik, *Statistical Learning Theory*, Springer, 1998.
7. N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
8. K.-A. Toh, W.-Y. Yau, and X. Jiang, "A Reduced Multivariate Polynomial Model For Multimodal Biometrics And Classifiers Fusion," *IEEE Trans. Circuits and Systems for Video Technology (Special Issue on Image- and Video-Based Biometrics)*, vol. 14, no. 2, pp. 224–233, 2004.
9. Patrick Verlinde, Gerard Chollet, and Marc Acheroy, "Multimodal Identity Verification Using Expert Fusion," *Information Fusion*, vol. 1, no. 1, pp. 17–33, 2000.
10. T. Wark, S. Sridharan, and V. Chandran, "Robust Speaker Verification via Asynchronous Fusion of Speech and Lip Information," in *2nd Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington, D.C., 1999, pp. 37–42.
11. C. Sanderson and K. K. Paliwal, "Noise Compensation in a Person Verification System Using Face and Multiple Speech Features," *Pattern Recognition*, vol. 36, no. 2, 2003.
12. N. Poh and S. Bengio, "How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks?," Research Report 04-18, IDIAP, Martigny, Switzerland, 2004, accepted for publication in *IEEE Trans. Signal Processing*, 2005.
13. J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.
14. N. Poh and S. Bengio, "Non-Linear Variance Reduction Techniques in Biometric Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 123–130.
15. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Research Report 04-44, IDIAP, Martigny, Switzerland, 2004, Accepted for publication in *AVBPA 2005*.
16. A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.
17. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
18. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.