Towards Predicting Optimal Fusion Candidates: A Case Study on 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. Combining multiple information sources, typically from several data streams is a very promising approach, both in experiments and to some extend in various real-life applications. However, combining too many systems (base-experts) will also increase both hardware and computation costs. One way to selecting a subset of optimal base-experts out of N is to carry out the experiments explicitly. There are $2^N - 1$ possible combinations. In this paper, we propose an analytical solution to this task when weighted sum fusion mechanism is used. The proposed approach is at least valid in the domain of person authentication. It has a complexity that is additive between the number of examples and the number of possible combinations while the conventional approach, using brute-force experimenting, is multiplicative between these two terms. Hence, our approach will scale better with large fusion problems. Experiments on the BANCA multi-modal database verified our approach. While we will consider here fusion in the context of identity verification via biometrics, or simply biometric authentication, it can also have an important impact in meetings because this *a priori* information can assist in retrieving highlights in meeting analysis as in “who said what”. Furthermore, automatic meeting analysis also requires many systems working together and involves possibly many audio-visual media streams. Development in fusion of identity verification will provide insights into how fusion in meetings can be done. The ability to predict fusion performance is another important step towards understanding the fusion problem.

1 Introduction

Combining multiple systems, or base-experts, to boost performance is a very promising approach. In [1], for instance, as many as 14 experts were fused and tested on the XM2VTS database. This study concluded that by adding more experts, the performance of the fused system will not be degraded. While this is true, in practice, fusion leads to added hardware and computational cost. Hence, it is often desirable to select an optimal subset of experts for fusion.

If there are N base-experts, a brute-force experimenting will require $2^N - 1$ fusion experiments to select the smallest optimal subset of base-expert candidates for fusion. Here, in the context of biometric authentication (BA), we attempt to *reduce* the overhead computation cost *without compromising* the effectiveness. The overhead cost is avoided by evaluating the F-ratio criterion $2^N - 1$ times instead of carrying out $2^N - 1$ experiments. F-ratio is a term that is non-linearly a function of Equal Error Rate (EER).

F-ratio arises naturally when assuming that the client and impostor scores are normally distributed. The accuracy of this criterion depends only on how accurately one can estimate the parameters in the VR-EER analysis. The first part of this analysis is Variance Reduction (VR) and the second part is Equal Error Rate (EER) analysis. In short, it has been shown [2] that fusion in BA using multiple experts result in reduced variance, which in turns, results in reduced EER. EER is a commonly used error measure in BA. One specificity about this analysis is that the correlation among experts are explicitly considered and can be described by a class-dependent full covariance matrix.

In this work, we will consider a subtask in meeting analysis: identity verification via biometrics, or simply biometric authentication. Knowing the identity in meeting has an important impact in meetings because this *a priori* information can assist in retrieving highlights in meetings as in “who said what”. Biometric authentication also shares another similarity with meeting analysis; in biometric authentication, users (or participants) are often known and only a few biometric examples (e.g. face with different orientations, speech samples, etc) per user are available to the system. The system is then required to verify the identity. In meetings, the participants are usually known and thus tracking of the person becomes a matter of identity verification. Furthermore, automatic meeting analysis also requires many systems working together and involves possibly many audio-visual media streams, e.g., the speech signal and the facial features. Development in fusion of biometric authentication will provide insights into how fusion in meeting can be done.

In the following, by using weighted sum fusion, we show that the VR-EER analysis can be used to predict optimal fusion candidates if the development set matches the evaluation set. In the presence of slight mismatch between development and evaluation sets, such is the case of the BANCA database, the predicted subset is still acceptable.

Section 2 presents two methods to choosing optimal fusion candidates: the brute-force approach and our proposed analytical approach. Section 3 presents briefly the BANCA experiment setup whereby 70 fusion experiments will be conducted. Sections 4 verifies experimentally that F-ratio calculated from a development set matches its counterpart on an evaluation set. Section 5 further examines the predictability of fusion candidates based on F-ratio calculated from the development set. The complexity of the proposed technique is evaluated in Section 6. This is followed by conclusions in Section 7.

2 Brute-Force Experimenting Versus Analytical Solution To Predicting Optimal Fusion Candidates

This section presents a conventional approach followed by our proposed approach to predicting a subset of candidates (base-experts or systems) that will be optimised in terms of performance when combined. The first approach is termed brute-force experimenting while the second is our proposal using an analytical solution. Note that this analytical solution is only possible when the fusion model is a linear combination of a subset of (the output of) N base-experts from all the available M base-experts.

Let us introduce the following notations. Let y_i^k be the output of system i indicating how probable a given input stimulus (biometric trait) is a client when the actual class label (target class) is $k \in \{C, I\}$, i.e., either a client or an impostor. Here, the expected value of client is always greater than that of impostor, i.e., $E[y_i^{k=C}] > E[y_i^{k=I}]$, where $E[z]$ is the expectation of z. Furthermore, let us assume that the combined model is of the form:

$$y_{GEN}^k \equiv \sum_{i=1}^{N} y_i^k \alpha_i, \tag{1}$$

where N is the chosen chosen number of experts and $\alpha_i | \forall_i$ weigh the output of each base-expert.

2.1 The Brute-Force Approach

Suppose there are two sets of data containing scores of M base-experts: development and evaluation data sets. The goal is to identify among all the M, which combination of at most N base-experts will give an optimal performance. In the brute-force approach, to choose from at most N out of M base-experts, there are altogether

$$^M C_1 + {}^M C_2 + \ldots + {}^M C_N = \sum_{i=1}^{N} {}^M C_i \tag{2}$$

possibilities, where $^m C_k$ is "m choose k" or $\frac{m!}{k!(m-k)!}$ by definition. To choose from all possible combinations, the total number is $2^M - 1$. The reason for minus one is that we do not consider the solution containing 0 base-expert. The brute-force approach will perform the following:

1. For each of the possible combinations:
   - estimate the best weights in the linear combination (1) from the development set according to a criterion (such as Mean Squared Error)
   - use the weights to evaluate the performance on the development set
2. Choose the best fusion candidate based on the criterion
3. Evaluate the chosen model on the evaluation set

In practice, before the linear combination, the output y_i^k of each expert should be normalised so that none of the base-expert score y_i^k will dominate the combination just because it has large values. Let $y_i^{norm,k}$ be the normalised value of y_i^k. The most common way to normalise the score is as follows:

$$y_i^{norm,k} \equiv \frac{y_i^k - \mu_i^{all}}{\sigma_i^{all}}, i = 1 \ldots, N. \tag{3}$$

The normalizing parameters μ_i^{all} and σ_i^{all} are mean and standard deviation calculated from the development set. These parameters are then applied on both the development and evaluation sets. In this way, the procedure of linear combination actually works on *normalised score space*. The linear combination is thus performed as follows:

$$y_{GEN}^{norm,k} \equiv \sum_{i=1}^{N} y_i^{norm,k} \alpha_i, \tag{4}$$

where $\alpha_{i=1\ldots,N}$ are weights associated to each base-expert i and $y_{GEN}^{norm,k}$ is the fused score. The weights can be found using different methods, such as least-square minimisation or Fisher's linear discriminant [3, Chap. 3]. Two sets of $\left\{y_{GEN}^{norm,k} | k = \{C, I\}\right\}$ are thus obtained, one from the development set and the other from the evaluation set. Note that $y_{GEN}^{norm,k}$ is one-dimensional (after fusion) and $y_i^{norm,k} | \forall_i$, as well as $y_i^k | \forall_i$ are N-dimensional data (of scores). The final decision function $F(\mathbf{x})$ (given a biometric sample $\mathbf{x}$, which is implicit in all variables $y_i^k, \forall_{i=1,\ldots,N}$), accepts or rejects an access claim by comparing $y_{GEN}^{norm,k}$ with a threshold, as follows:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y_{GEN}^{norm} > \Delta \\ reject & \text{otherwise.} \end{cases} \tag{5}$$

This threshold should be calculated from the development set and then applied to the evaluation set.

The above procedure is then repeated for $\sum_{i=1}^{N} {}^{M}C_i$ combinations. The combination or fusion candidate that gives the lowest Equal Error Rate (EER) on the evaluation set is the so-called "optimal" fusion candidate. EER is defined as the point where False Acceptance Rate (FAR) equals False Rejection Rate (FRR). They are defined as:

$$\text{FAR}(\Delta) = \frac{\text{number of FAs}(\Delta)}{\text{number of impostor accesses}}\ , \tag{6}$$

$$\text{FRR}(\Delta) = \frac{\text{number of FRs}(\Delta)}{\text{number of client accesses}}\ . \tag{7}$$

Note that FA and FR are functions of a pre-determined threshold Δ. The empirical procedure to find Δ that satisfies the EER criterion (on the development set) is:

$$\Delta^* = \arg\min_{\Delta} |\text{FAR}(\Delta) - \text{FRR}(\Delta)|. \tag{8}$$

The *empirical* EER value found this way is often reported as a single value called Half Total Error Rate (HTER). It is defined as the average of FAR and FRR:

$$\text{HTER}(\Delta^*) = \frac{\text{FAR}(\Delta^*) + \text{FRR}(\Delta^*)}{2}. \tag{9}$$

The fusion candidate with minimum HTER on the development set is considered the optimal solution. Taking normalisation step into account, for each fusion candidate the brute-force approach thus needs to loop through the data:

- once to obtained normalise scores on the development and evaluation data sets;
- at least once to calculate the weights[1] on the development data set,

[1] An iterative solution may require more passes, e.g. a single-layer Perceptron.

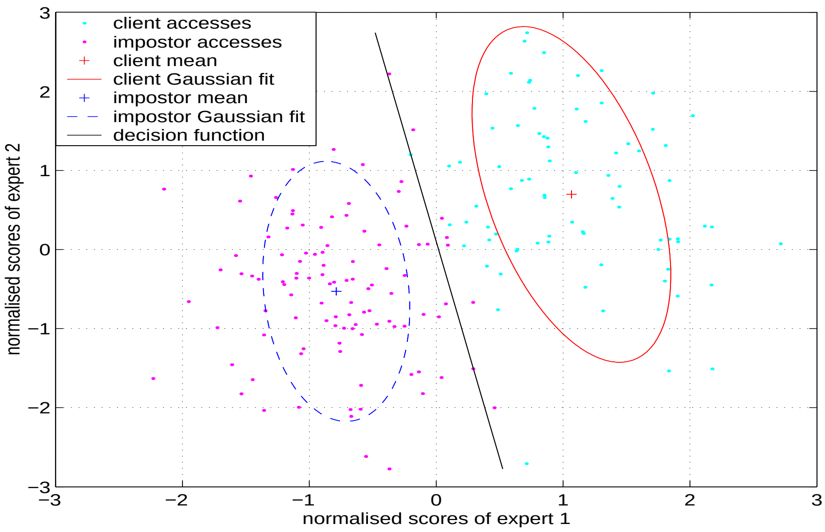

Fig. 1. A geometric interpretation of the proposed approach. Expert 1 is IDIAP's voice system and expert 2 is Surrey's automatic face authentication system, applied on the Ud-g1 BANCA data set.

- once to apply the weights on the development and evaluation data sets, and
- once to evaluate the EER criterion on the development and evaluation data sets

The optimal fusion candidate found on the evaluation set is considered the "ground-truth". When there is no mismatch between development and evaluation data sets, we expect that the optimal fusion candidate found using this procedure to be similar.

2.2 The Analytical Solution

The proposed approach assumes that all y_i^k's are Gaussian distributed. It has the advantage that there is no need to loop through the data set $\sum_{i=1}^{N} {}^M C_i$ times but only once. To quickly give an intuitive picture, in a two-dimensional case (i.e., fusing only two experts), data points of one of the experiments (to be discussed in Section 3) are plotted in Figure 1. Superimposed on the data points are two full-covariance Gaussians, one for the client scores and the other for the impostor scores.

The technical challenge of the proposed method is to estimate the EER of the fused score $y_{GEN}^{norm,k}$. This can be done as follows: first estimate the normalising parameters μ_i^{all} and σ_i^{all} from the development set. Then, estimate the class-dependent Gaussian parameters $\varPhi \equiv \{\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\}$. $\boldsymbol{\mu}^k$ is a vector containing the class-dependent mean score of experts μ_i^k for $i = 1, \ldots, N$ and $\boldsymbol{\Sigma}^k$ is the class-dependent covariance matrix whose elements are $\varSigma_{i,j}^k \equiv E[W_i^k, W_j^k]$, where W_i^k is the noise distribution associated to expert i (see also the appendix). Note that $\varSigma_{i,i}^k$ is simply the variance of scores of expert i.

From these parameters, it is possible to calculate the weights α_i and EER using an intermediate variable called F-ratio. The technical details of deriving F-ratio is beyond the scope of this study. They can be found in [2]. The calculation of weights can be found in classical references such as [3, Chap. 3]. Here, we present how the theoretical EER can be calculated. Let EER_{GEN}^{norm} be the EER of the normalised fused score. Its solution is:

$$\text{EER}_{GEN}^{norm} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}_{GEN}^{norm}}{\sqrt{2}}\right), \tag{10}$$

where

$$\text{F-ratio}_{GEN}^{norm} = \frac{\mu_{GEN}^{norm,C} - \mu_{GEN}^{norm,I}}{\sigma_{GEN}^{norm,C} + \sigma_{GEN}^{norm,I}}, \tag{11}$$

and

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp\left[-t^2\right] dt. \tag{12}$$

$\mu_{GEN}^{norm,k}$ and $\sigma_{GEN}^{norm,k}$ are mean and standard deviation of the fused and normalised score $y_{GEN}^{norm,k}$. Their solutions are:

$$\mu_{GEN}^{norm,k} = \sum_{i=1}^{N} \frac{\alpha_i}{\sigma_i^{all}}\left(\mu_i^k - \mu_i^{all}\right) \tag{13}$$

and

$$(\sigma_{GEN}^{norm,k})^2 = \sum_{m=1}^{N}\sum_{n=1}^{N} \frac{\alpha_m \alpha_n}{\sigma_m^{all}\sigma_n^{all}} E\left[W_m^k W_n^k\right] \tag{14}$$

respectively, for any $k \in \{C, I\}$. The derivation of Eqns. (13 and 14) can be shown in the appendix. As can be seen, all calculations can be solved analytically, including the optimal decision threshold $\varDelta$ in the final decision function $F(\mathbf{x})$, as defined in Eqn. (5). The solution of $\varDelta$ is:

$$\varDelta = \frac{\mu_{GEN}^{norm,I}\sigma_{GEN}^{norm,C} + \mu_{GEN}^{norm,C}\sigma_{GEN}^{norm,I}}{\sigma_{GEN}^{norm,I} + \sigma_{GEN}^{norm,C}}. \tag{15}$$

The decision boundary in Figure 1 was indeed obtained using Eqn. (15). This analytical solution is actually derived from what is called VR-EER analysis due to our preceding work [2]. This analysis links the well-known variance reduction (VR) phenomenon due to committee of classifiers (as discussed in [3, Chap. 9] and elsewhere in the literature) to reduced EER. The parameters are thus called VR-EER parameters.

The next section will present the experiment setup that will be used to test our proposed approach.

3 Experiment Setup

The BANCA database [4] is the principal database used in this paper. It has a collection of face and voice prints of up to 260 persons in 5 different languages. In this paper, we only used the English subset. Hence only 52 people are used here; 26 are males and 26 are females. There are altogether 7 protocols, namely, Mc, Ma, Md, Ua, Ud, P and G, simulating matched control, matched adverse, matched degraded, uncontrolled adverse, uncontrolled degraded, pooled and grant test, respectively. For protocols P and G, there are 312 client accesses and 234 impostor accesses. For all other protocols, there are 78 client accesses and 104 impostor accesses. A set of face and speaker authentication experiments were carried out by University of Surrey (2 face experiments), IDIAP (speaker), UC3M (speaker) and UCL (face)[2]. Hence, there are 5 baseline experiments per protocol, making a total of 35 baseline experiments. Details of these experiments can be found in [5]. For each protocol, we used the following score files:

- `IDIAP_voice_gmm_auto_scale_33_200`
- `SURREY_face_svm_auto`
- `SURREY_face_svm_man`
- `UC3M_voice_gmm_auto_scale_34_500`
- `UCL_face_lda_man`

Moreover, for each protocol, there are two subgroups, called g1 and g2. In this paper, g1 is used as a *development* set while g2 is used as an *evaluation* set. The test set is considered the "ground-truth" data set and is used exclusively for *testing* only. It is particularly useful to determine generalisation performance, i.e., how well a classifier performs on unseen data sets. For each protocol, by combining each time two baseline experts, one can obtain 10 fusion experiments, given by ${}^5C_2$. This results in a total of 70 experiments for all protocols. Similarly by combining each time three baseline experts, one will have a total of $7 \times {}^5C_3 = 70$ experiments.

4 Generalisation Using Weighted-Sum Fusion

In [6], it was shown that given full knowledge about VR-EER parameters, F-ratio of fused score using the *mean* operator can be estimated accurately. Furthermore EER can be predicted fairly accurately, by assuming that the client and impostor scores are drawn from Gaussian distributions. There are two issues to be examined here. The first issue is, given full knowledge about the VR-EER parameters (typically on a development set), would theoretical F-ratios match empirical F-ratios[3]? The second issue is, would it be possible to predict F-ratio on *unseen* data. This is the issue of generalisation. In this case, F-ratio from the development set is compared to F-ratio from the evaluation sets of BANCA protocols (see Section 3) . Note that each of these tests will be repeated $2^5 - 1 = 31$ times for each of the 7 protocols, each time using a different combination of 1–5 base-experts. Hence, there are altogether 217 experiments.

These two issues are detailed below:

[2] Available at "ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores"

[3] *Empirical* F-ratio means that the F-ratio is obtained by actually carrying out a complete experiment, whereas *theoretical* F-ratio means that the F-ratio is estimated analytically. Hence, evaluation of empirical F-ratio requires a pass through the data while its theoretical counterpart requires only direct applications of Eqns. (11), (13) and (14).

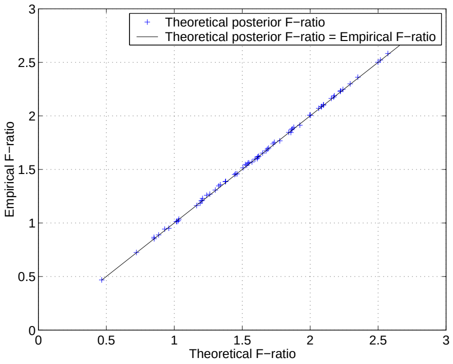

Fig. 2. Comparison of a theoretical prior F-ratio and empirical F-ratio, based on BANCA development set, over all possible combinations and all protocols, i.e., 217 data points.

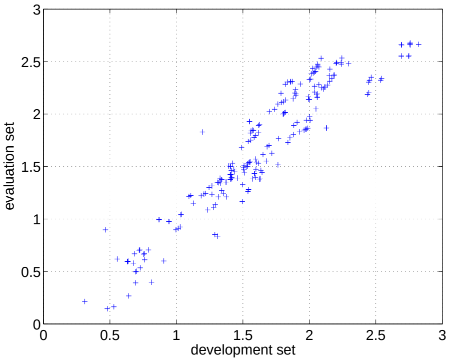

Fig. 3. F-ratios of combined scores of the development set versus those of the evaluation set, over all possible combinations and all protocols, i.e., 217 data points.

1. **Posterior test**. One knows all the VR-EER parameters. This is typically the case for the development set. The empirical F-ratio of each fusion candidate is compared with its theoretical counterpart, with the weights estimated from the same set. This test is called posterior because one has all the information about the data set.
2. **Evaluation test**. Having all the information about the development set, the goal here is to test if one can extrapolate this information on an (unseen) evaluation set. This tests how closely the development set corresponds to the evaluation set. Empirical F-ratio of evaluation set is plotted against its development set counterpart.

The results are shown in Figure 2 for the first test and Figure 3 for the second test. As can be seen, the posterior test shows that, given full knowledge about the VR-EER parameters, each data point (which corresponds to each fusion experiment) can be predicted accurately. The evaluation test shows that F-ratios of the development set is correlated to that of the evaluation set. Hence, prediction is possible. The inaccuracy is due to the inherent mismatch between the development and the evaluation sets.

5 Predicting Optimal Subsets of Base-Expert Candidates

In the previous evaluation test, it was shown that the F-ratios between the development and the evaluation sets due to fusion are correlated. This implies that good candidates for fusion in a development set would also be good candidates in the corresponding evaluation set. The next experiment is to examine how accurate the prediction of the best fusion candidate can be if we were to choose *from all* M base-experts. Note that we could have also conducted a series of experiments to find out the accuracy of predicting *at most* the N best fusion candidates from all M base-experts for different values of $N \leq M$. Since choosing from all is a more difficult task than choosing at most N, we will only illustrate the former problem. Before doing so, let us label some of the

Table 1. Labels of corresponding fusion experiments using 1, 2, 3 and 4 base-experts. The numbers 1–5 in the right columns of each table correspond to the five base-experts discussed in Section 3. The letter "E" is assigned to fusion of all experts.

(a) base-expert	
labels	experts
a	1
b	2
c	3
d	4
e	5

(b) 2-expert fusion	
labels	experts
f	1 2
g	1 3
h	1 4
i	1 5
j	2 3
k	2 4
l	2 5
m	3 4
n	3 5
o	4 5

(c) 3-expert fusion	
labels	experts
p	1 2 3
q	1 2 4
r	1 2 5
s	1 3 4
t	1 3 5
u	1 4 5
v	2 3 4
w	2 3 5
x	2 4 5
y	3 4 5

(d) 4-expert fusion	
labels	experts
z	1 2 3 4
A	1 2 3 5
B	1 2 4 5
C	1 3 4 5
D	2 3 4 5

combinations as listed in Table 1. The numbers 1–5 correspond to the five base-experts discussed in Section 3. Figure 4 shows the top five fusion candidates due to choosing from all 31 fusion candidates. There are 7 sub-figures, each corresponding to a BANCA protocol. The EERs of the candidates are sorted from the smallest to the biggest in the x-axis. Hence the first item in the x-axis is the best candidate fusion candidate. For example, the best candidate according to protocol G, as shown in Figure 4, is z (1-2-3-4), according to the development set but is A (1-2-3-5) according to the evaluation set. Since the evaluation set is taken as the "ground-truth", i.e, A is the correct answer, we need to consider the top 3 candidates in order to "remedy" this error. For the protocol Mc as well as Md and Ua , it takes the top two candidates to remedy this error. Ideally, it is desirable that the top candidate as proposed by the development and evaluation set to be the same. Such is the case for protocols P, Ma and Ud. By varying the top-k candidates where $k = 1, 2, 3$ and applying the analysis for other protocols, we obtain Table 2. As can be observed, most of the errors committed by choosing the top fusion candidate can be rectified when choosing the top-2 fusion candidates. Note that even though some proposed optimal fusion candidates are not coherent with the evaluation set, the proposed optimal fusion candidates are not very far off from their follow-up candidates in terms of EER, across different protocols, i.e., top few optimal candidates have very similar EERs. As a result, even if the proposed candidate according to the de-

Table 2. Mistakes committed by choosing the top-k fusion candidate(s) by choosing from all 31 fusion candidates, over all 7 protocols, for $k = 1, 2, 3$. As k increases, errors will decrease.

Top k	Errors committed over 7 protocols
1	4
2	1
3	0

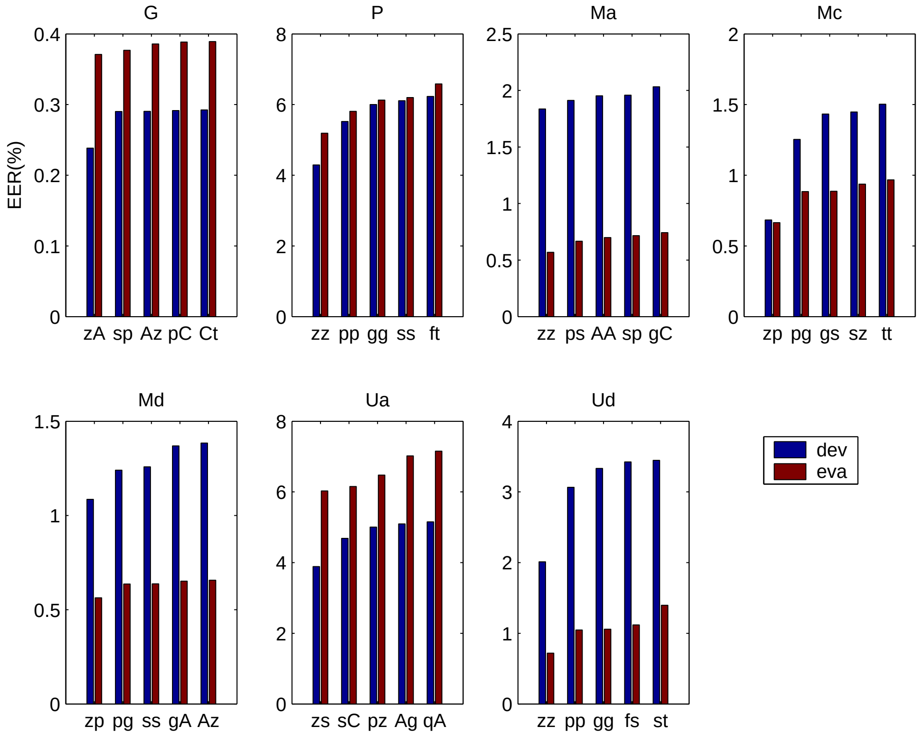

Fig. 4. Top five fusion candidates according to the EER criterion (sorted in the x-axis) by choosing from all 31 fusion candidates for each of the 7 protocols. The alphabets "XY" represent the fusion candidate X proposed by the development set and Y by the evaluation set. Tables 1(a)–(d) show the corresponding base-experts with E denoting fusing all the experts.

velopment set might be incorrect, the increase of EER error due to this wrong selection is not big.

6 Analysis of Complexity

We now analyse the complexity of our proposed approach and compare it with the brute-force approach. Let the number of development examples and evaluation examples be l_{dev} and l_{eva} In the brute-force approach, to choose one best fusion candidate from all possible M base-experts, one would have to carry out the experiment $2^M - 1$ times. Furthermore, in each experiment, one has to loop through $l_{dev} + l_{eva}$ examples. The complexity is thus:

$$O\left((l_{dev} + l_{eva}) \times (2^M - 1)\right). \tag{16}$$

In the proposed approach, one only has to loop through the both development and evaluation sets once to derive all the VR-EER parameters (i.e., class-dependent mean and covariance matrix plus the global mean and standard deviations) and then to evaluate the F-ratio criterion $2^N - 1$ times on the evaluation set. Hence, the complexity is thus:

$$O\left(l_{dev} + l_{eva}) + 2^M - 1\right). \qquad (17)$$

In the brute-force approach, the $2^M - 1$ is multiplicative with the number of development and evaluation examples whereas in the proposed approach, these two terms are additive. Therefore, our approach is scalable to larger fusion problems with significant reduction of computation cost. It should be noted that the computation involved even in the brute-force approach in this case is simple ($M = 5$). However, for large problems, this benefit will be more obvious.

7 Conclusion

In this paper, using a Gaussian model with full covariance matrix to model the client and impostor distributions, on the zero-mean unit-variance normalised score space, we showed how to predict theoretically the performance of an authentication system based on Equal Error Rate (EER) using weighted sum fusion. This approach is based on VR-EER analysis due to [2]. The advantage of the proposed approach is that one does not have to make the assumption that the base-experts are independent and that their scores are not correlated, as frequently done in the literature. In fact, the dependency is already captured by the covariance matrix. Although a single full covariance matrix seems to be overly simple (as compared to mixture of Gaussians with diagonal covariance matrix), we have shown that it is adequate to model EER [6] as a function of *F-ratio*, a quantity that measures how separable the client distribution is from the its impostor counterpart.

The central idea of this work is to use F-ratio as a criterion to search for an optimal subset of base-experts for fusion in an efficient way. Although F-ratio was previously established in [2], this study demonstrates a way to predict the performance of fusion analytically, without compromising the effectiveness when one actually carries out the fusion experiments. Hence the proposed technique allows us to select an optimal subset of base-experts for fusion in an efficient way. To choose one optimal fusion candidate from M base-experts, the brute-force approach needs to carry out $2^M - 1$ experiments and for each experiment, this approach will need to cycle through the data set several times. The proposed approach needs only to loop through the development and evaluation set once and to evaluate the F-ratio criterion $2^M - 1$ times. Hence, our approach has only a fixed computation cost with respect to the size of the available data set and will scale well with large fusion problems.

We tested our approach on the BANCA database and showed that F-ratio can be predicted accurately if one has the full knowledge about the data distribution (e.g. development data set). The prediction degrades when one knows less and less about the data (e.g. the test data set). In fact, by actually carrying out 217 fusion experiments on the BANCA database, we showed that the F-ratio on the development set (g1) is correlated to that of the test set (g2), despite their mismatch. Exploiting this ability, we were able to predict an optimal subset of fusion (base-expert) candidates fairly accurately on the 7 BANCA protocols. The accuracy cannot be 100% since there exists an intrinsic mismatch between the development and evaluation sets.

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. The authors also thank the anonymous reviewers for detailed comments and suggestions. This publication only reflects the authors' views.

A Derivation of Solutions

Suppose $y_{i,t}^k$ is the t-th observed sample of the i-th response of class k, recalling that $i = 1, \ldots, N$ and $k = \{C, I\}$. We assume that this observed variable has a deterministic component and a noise component and that their relation is additive. The deterministic component is due to the fact that the class is discrete in nature, i.e., during authentication, we know that a user in *either* a client or an impostor. The noise component is due to some random processes during biometric acquisition (e.g. degraded situation due to light change, miss-alignment, etc) which in turns affect the quality of extracted features. Indeed, it has a distribution governed by the extracted feature set $\mathbf{x}$ often in a non-linear way. By ignoring the source of distortion in extracted biometric features, we actually assume the noise component to be random (while in fact they may be not if we were able to systematically incorporate all possible variations into the base-expert model).

Let μ_i^k be the deterministic component. Note that its value is *only dependent on* the class $k = \{C, I\}$ and independent of sample t. We can now model $y_{i,t}^k$ as a sum of this deterministic value plus the noise term $w_{i,t}^k$, as follows:

$$y_{i,t}^k = \mu_i^k + w_{i,t}^k, \tag{18}$$

for $k \in \{C, I\}$ where $w_{i,t}^k$ is an instance of variable W_i^k which follows an unknown distribution $\mathcal{W}_i^k$ with zero mean and $(\sigma_i^k)^2$ variance, i.e., $w_{i,t}^k \sim W_i^k$ and $p(W_i^k) \propto \mathcal{W}\left(0, (\sigma_i^k)^2\right)$. By adopting such a simple model, from the fusion point of view, we effectively encode the i-th expert score as the sum of a deterministic value and another random variable, in a class-dependent way. Following Eqn. (18), we can deduce that the distribution of Y_i^k is proportional to the that of the noise, i.e., $p(Y_i^k) \propto \mathcal{W}\left(\mu_i^k, (\sigma_i^k)^2\right)$. The expectation of Y_i^k (over different sample t) is:

$$E[Y_i^k] = E[\mu_i^k] + E[W_i^k] = \mu_i^k. \tag{19}$$

Let $\Sigma_{i,j}^k$ be the i-th and j-th element of the covariance matrix of $Y_i^k|\forall_i$, i.e., $\boldsymbol{\Sigma}^k$. It can be calculated as:

$$\begin{aligned} \Sigma_{i,j}^k &\equiv E\left[(Y_i^k - \mu_i^k)(Y_j^k - \mu_j^k)\right] \\ &= E[W_i^k W_j^k] \end{aligned} \tag{20}$$

where $\Sigma_{i,j}^k$ is the covariance between two distributions Y_i^k and Y_j^k. When $i = j$, we have the definition of variance of Y_i^k, i.e.,

$$\Sigma_{i,i}^k \equiv (\sigma_i^k)^2 = E[W_i^k W_i^k]. \tag{21}$$

Let $\mu_{GEN}^{norm,k}$ and $\sigma_{GEN}^{norm,k}$ be the mean and standard deviation of combined scores derived from y_i^k for $i = 1, \ldots, N$ (see Eqn. (4)). Let $Y_{GEN}^{norm,k}$ be the distribution from which $y_{GEN}^{norm,k}$ is drawn. Note that μ_{GEN}^k and σ_{GEN}^k can be defined by Eqn. (19) and Eqn. (21) by replacing the index i by GEN and similarly for $\mu_{GEN}^{norm,k}$ and $\sigma_{GEN}^{norm,k}$. The expected value of $Y_{GEN}^{norm,k}$, for $k = \{C, I\}$, is:

$$\begin{aligned}
\mu_{GEN}^{norm,k} &\equiv E[Y_{GEN}^{norm,k}] \\
&= \sum_{i=1}^{N} \alpha_i E[Y_i^{norm,k}] \\
&= \sum_{i=1}^{N} \frac{\alpha_i}{\sigma_i^{all}} \left(E[Y_i^k] - \mu_i^{all} \right) \\
&= \sum_{i=1}^{N} \frac{\alpha_i}{\sigma_i^{all}} \left(\mu_i^k - \mu_i^{all} \right) \qquad (22)
\end{aligned}$$

The variance of $Y_{GEN}^{norm,k}$ is:

$$\begin{aligned}
(\sigma_{GEN}^{norm,k})^2 &= Cov(Y_{GEN}^{norm,k}, Y_{GEN}^{norm,k}) \\
&= E\left[\left(Y_{GEN}^{norm,k} - E[Y_{GEN}^{norm,k}] \right)^2 \right] \\
&= E\left[\left(\sum_{i=1}^{N} \frac{\alpha_i (Y_i^k - \mu_i^{all})}{\sigma_i^{all}} - \sum_{i=1}^{N} \frac{\alpha_i (\mu_i^k - \mu_i^{all})}{\sigma_i^{all}} \right)^2 \right] \\
&= E\left[\left(\sum_{i=1}^{N} \frac{\alpha_i (Y_i^k - \mu_i^k)}{\sigma_i^{all}} \right)^2 \right] \\
&= E\left[\left(\sum_{i=1}^{N} \frac{\alpha_i W_i^k}{\sigma_i^{all}} \right)^2 \right] \qquad (23)
\end{aligned}$$

To expand Eqn. (23), one should take care of possible correlation between different W_m^k and W_n^k, as follows:

$$\begin{aligned}
(\sigma_{GEN}^{norm,k})^2 &= E\left[\left(\sum_{m=1}^{N} \sum_{n=1}^{N} \frac{\alpha_m W_m^k \alpha_n W_n^k}{\sigma_m^{all} \sigma_n^{all}} \right) \right] \\
&= \sum_{m=1}^{N} \sum_{n=1}^{N} \frac{\alpha_m \alpha_n}{\sigma_m^{all} \sigma_n^{all}} E\left[W_m^k W_n^k \right] \qquad (24)
\end{aligned}$$

for any $k \in \{C, I\}$.

References

1. J. Kittler, K. Messer, and J. Czyz, "Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems," in *Proc. Cost 275 Workshop*, Rome, 2002, pp. 17–24.

2. N. Poh and S. Bengio, "Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks?," in *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, 2004, pp. vol. V, 893–896.
3. C. Bishop, *Neural Networks for Pattern Recognition,* Oxford University Press, 1999.
4. E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Springer LNCS-2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. 2003, Springer-Verlag.
5. Christine Marcel, "Multimodal Identity Verification at IDIAP," Communication Report 03-04, IDIAP, Martigny, Switzerland, 2003.
6. 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.