Fixed FAR correction factor of score level fusion

Susyanto, N.; Veldhuis, R.N.J.; Spreeuwers, L.J.; Klaassen, C.A.J.

Published in:
The IEEE Eighth International Conference on Biometrics: Theory, Applications and Systems

DOI:
10.1109/BTAS.2016.7791173

Fixed FAR Correction Factor of Score Level Fusion

N. Susyanto[1] R.N.J. Veldhuis[2] L.J. Spreeuwers[2] C.A.J. Klaassen[1]
[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]Faculty of EEMCS, University of Twente
[1]{n.susyanto,c.a.j.klaasen}@uva.nl, [2]{r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

Abstract

In biometric score level fusion, the scores are often assumed to be independent to simplify the fusion algorithm. In some cases, the "average" performance under this independence assumption is surprisingly successful, even competing with a fusion that incorporates dependence. We present two main contributions in score level fusion: (i) proposing a new method of measuring the performance of a fusion strategy at fixed FAR via Jeffreys credible interval analysis and (ii) subsequently providing a method to improve the fusion strategy under the independence assumption by taking the dependence into account via parametric copulas, which we call fixed FAR fusion. Using synthetic data, we will show that one should take the dependence into account even for scores with a low dependence level. Finally, we test our method on some public databases (FVC2002, NIST-face, and Face3D), compare it to Gaussian mixture model and linear logistic methods, which are also designed to handle dependence, and notice its significance improvement with respect to our evaluation method.

1. Introduction

In a score based biometric person verification system, a *threshold* has to be set to decide whether a matching score between two biometric samples (query and template) is a *genuine* or an *impostor* score. A genuine score leads to the conclusion that the query and template originate from the same person while an impostor score means that the query and template stem from different people. We will assume that the matching score is a similarity score. Note that once the threshold is set, the system can make two different errors: accept an impostor score as genuine score and reject a genuine score. The probability of accepting an impostor score is called the *False Acceptance Rate (FAR)*, while the probability of rejecting a genuine score is called the *False Rejection Rate (FRR)*. The complement of the FRR is called the *True Positive Rate (TPR)*, which is defined as the probability of accepting a genuine score as genuine score. Since every genuine score will be either accepted or rejected by the system, we have $\mathrm{TPR} = 1 - \mathrm{FRR}$. The most common method to evaluate a biometric person verification system is by plotting the relation between FAR and TPR, which is known as *Receiver Operating Characteristics* (ROC).

When there are two or more matchers, one has to transform these multiple scores to a new score (a scalar) as a fused score, which is called score level fusion. There are three categories in score level fusion. The most commonly used one is the transformation-based one which is done by mapping all components of the vector of matching scores to a comparable domain and applying some simple rules such as sum, mean, max, med, etc. [10]. However, this approach relies heavily on the niceness of the training set used for the transformation. For example if one wants to normalize each component of the vector of matching scores to the unit interval [0,1] (which is called minmax normalization), then the maximum and the minimum of all scores have to be determined. Unfortunately, when the maximum and minimum scores have to be estimated from the training set that has outlier(s), the estimation will be very bad. The second approach is classifier-based fusion which is done by stacking all components of the vector of matching scores and applying a classifier to separate the genuine and impostor scores [11]. The last approach is based on estimation of the densities of the genuine and impostor scores [14]. According to [17] this approach, which is also known as likelihood ratio based, would be optimal if the underlying densities were known. However, in practice, such densities have to be estimated from data so that the performance relies on how well the two densities are estimated.

In this paper, we will focus on score level fusion for dependent matchers. The likelihood ratio

based fusion automatically incorporates the dependence between matchers. However, this approach needs to estimate two density functions, which is a challenging task. While the choice of an appropriate parametric model is sometimes difficult, nonparametric estimators suffer from the difficulty that they are sensitive to the choice of the bandwidth or of other smoothing parameters. To simplify, many researchers assume that all genuine and impostor scores are independent so that the likelihood ratio is only the product of the individual likelihood ratios of the matchers (henceforth called PLR fusion); see [13, 22, 23]. However, the independence assumption is not realistic since the scores are obtained from the same sample. A study of incorporating dependence instead of using PLR fusion is presented in [15] where the authors investigate the effect of considering correlation and compare their method to PLR fusion by computing the difference between the areas their respective ROCs. However, in practice the FAR has to be set in advance. For example, in a security application, the FAR is set to be very small and usually less than 0.1% or even 0.01%. Since area under ROC does not always reflect the performance at small FAR, we will compare the performance between dependent and PLR fusion at specific FAR.

This paper has two main contributions: proposing an evaluation of biometric fusion at fixed FAR and proposing a method to improve PLR fusion. In Section 2, we present our method to evaluate biometric fusion at fixed FAR. Instead of using parametric or nonparametric models, we propose a semiparametric approach, which will be called *fixed FAR fusion*, by modeling the marginal densities nonparametrically and the dependence between them by parametric copulas as explained in Section 3. We will see the gain of considering dependence using synthetic data and subsequently compare our method to GMM [14] and Logit [13] fusions, which are also intended to deal with matcher dependence, on some real databases (FVC2002, NIST-face, Face3D) in Section 4. Although also vector machine (SVM) fusion can handle dependence, we do not include it because it is a classifier tool so that we cannot set the FAR value beforehand (the FAR value of SVM fusion is automatically determined by the classifier). Finally, our conclusions are presented in Section 5.

2. Performance of biometric fusion at fixed FAR

Suppose we have d matchers. In biometric fusion, one has to find a function $\psi : \mathbb{R}^d \to \mathbb{R}$, which will be called a *fusion.* Let

$$\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}} \tag{2.1}$$

and

$$\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}} \tag{2.2}$$

be i.i.d copies of the d-dimensional random variable of genuine scores $\mathbf{S}_{\text{gen}}$ and impostor scores $\mathbf{S}_{\text{imp}}$, respectively. In this section, we will present how to measure the performance of a fusion at fixed FAR.

Let α be a fixed FAR. The exact TPR is

$$\text{TPR} = P(\psi(\mathbf{S}_{\text{gen}}) \geq \tau) \tag{2.3}$$

where the threshold τ is explicitly determined via relation

$$P(\psi(\mathbf{S}_{\text{imp}}) \geq \tau) = \alpha. \tag{2.4}$$

This means that all fused scores greater than or equal to τ will be recognized as genuine scores. In practice, we do not know the distribution functions of $\mathbf{S}_{\text{gen}}$ and $\mathbf{S}_{\text{imp}}$. However, we can compute the empirical value of TPR based on (2.1) and (2.2) by

$$\widehat{\text{TPR}} = \widehat{F}^{\psi}_{\text{gen}}(\hat{\tau}). \tag{2.5}$$

where

$$\hat{\tau} = \inf\{x \,:\, \widehat{F}^{\psi}_{\text{imp}}(x) \geq 1 - \alpha\}. \tag{2.6}$$

Here, $\widehat{F}^{\psi}_{\text{gen}}$ and $\widehat{F}^{\psi}_{\text{imp}}$ are *modified* empirical distribution functions based on the two samples

$$\psi(\mathbf{W}_1), \ldots, \psi(\mathbf{W}_{n_{\text{gen}}})$$

and

$$\psi(\mathbf{B}_1), \ldots, \psi(\mathbf{B}_{n_{\text{imp}}}),$$

respectively. Our modified empirical distribution function based on a sample $X_1, \ldots, X_n$ is defined by

$$\hat{F}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}. \tag{2.7}$$

The $\widehat{\text{TPR}}$ is only an estimated rate, which may be viewed as the probability of a Bernoulli experiment [20]. With n_{gen} genuine scores $\widehat{\text{TPR}}$ has a binomial distribution with success probability TPR, which may be approximated by $\text{Bin}(n_{\text{gen}}, \widehat{\text{TPR}})$. We employ Jeffreys method to construct a credible interval (CI) from this. It is one of the more trusted ways to obtain a CI here [2, 3]. In conclusion, for a given significance level $0 < \varepsilon << 1$, we will have the $100(1-\varepsilon)\%$ Jeffreys CI $[L, U]$ where

$$L = B(\varepsilon/2; \beta_1, \beta_2) \tag{2.8}$$

and

$$U = B(1 - \varepsilon/2; \beta_1, \beta_2) \tag{2.9}$$

with

$$\beta_1 = n_{\text{gen}}\widehat{\text{TPR}} + \frac{1}{2} \text{ and } \beta_2 = n_{\text{gen}}(1-\widehat{\text{TPR}}) + \frac{1}{2}.$$

Here, $B(\varepsilon; p_1, p_2)$ denotes the ε quantile of a Beta(p_1, p_2) distribution. This means that it is approximately $100(1-\varepsilon)\%$ certain that the true TPR is in-between L and U.

3. Fixed FAR correction factor

According to the Neyman-Pearson lemma [17], the optimal fusion is the likelihood-ratio-based method, i.e., by taking $\psi = \text{LR}$ where

$$\text{LR}(\mathbf{s}) = \frac{f_{\text{gen}}(\mathbf{s})}{f_{\text{imp}}(\mathbf{s})} \tag{3.1}$$

where f_{gen} and f_{imp} are the densities of genuine and impostor scores, respectively, which are unknown in practice. Therefore, we have to estimate the LR from data.

3.1. Correction factor

A copula is a distribution function on the unit cube $[0,1]^d$, $d \geq 2$, of which the marginals are uniformly distributed. Susyanto et al. [21] use a specific copula called Gaussian copula to handle dependence between classifiers in biometric fusion. However, since the Gaussian copula is appropriate for only a limited number of biometric data sets, we will use a family of well-known parametric copulas from the collection of elliptic and Archimedean copulas.

For any continuous multivariate distribution function there exists a copula function [18].

Theorem 3.1 (Sklar (1959))**.** *Let $d \geq 2$, and suppose H is a distribution function on $\mathbb{R}^d$ with one dimensional continuous marginal distribution functions $F_1, \cdots, F_d$. Then there is a unique copula function $C : [0,1]^d \to [0,1]$, so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3.2}$$

for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$.

The joint density function can be computed by taking the d-th derivative of (3.2):

$$h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times \prod_{i=1}^{d} f_i(x_i) \tag{3.3}$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. Note that according to (3.3), we can estimate separately the dependence structure represented by the copula density c and the individual densities f_i in order to get the joint density h. If C_α is determined by a finite dimensional Euclidean parameter α then it is called a parametric copula. In this case, we can estimate the dependence parameter α based on i.i.d. observations

$$\mathbf{X}_1, \ldots, \mathbf{X}_n$$

with

$$\mathbf{X}_i = (X_{1i}, \ldots, X_{di}) \quad \forall i = 1, \ldots, n$$

by the pseudo-maximum likelihood estimator (PMLE). Mathematically, the PMLE of α has to maximize

$$\frac{1}{n}\sum_{i=1}^{n} \log c_\alpha \left(\hat{F}_1(X_{1i}), \ldots, \hat{F}_d(X_{di})\right) \tag{3.4}$$

where $\hat{F}_j$ is the modified empirical distribution function as defined in (2.7) based on $X_{j1}, \ldots, X_{jn}$ for $1 \leq j \leq d$ and c_α is the copula density.

Let C_{gen} and C_{imp} be the copula corresponding to genuine and impostor scores with copula densities c_{gen} and c_{imp}, respectively. In view of (3.1) and (3.3), the likelihood ratio at score $\mathbf{s} = (s_1, \cdots, s_{\text{d}})$ can be written as

$$\text{LR}(\mathbf{s}) = \text{PLR}(\mathbf{s}) \times \text{CF}(\mathbf{s})$$

where

$$\text{PLR}(\mathbf{s}) = \prod_{i=1}^{d} \text{LR}_i(s_i) \tag{3.5}$$

is the product of the individual likelihood ratios and

$$\text{CF}(\mathbf{s}) = \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1), \cdots, F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1), \cdots, F_{\text{imp},d}(s_{\text{d}}))} \tag{3.6}$$

is the copula density ratio that will be called the *correction factor*. Here, $F_{\text{gen},i}$ and $F_{\text{imp},i}$ denote the distribution functions of genuine and impostor scores, respectively.

Note that for every i-th component of score $\mathbf{s} = (s_1, \cdots, s_{\text{d}})$, the posterior probability $P(H_1|s_i)$ can be estimated optimally by the Pool-Adjacent-Violators (PAV) algorithm as shown in [23] where H_1 correspond to a genuine user. Therefore, from the Bayesian relation

$$\frac{P(H_1|s_i)}{P(H_0|s_i)} = \frac{P(s_i|H_1)}{P(s_i|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

where H_0 corresponds to an impostor user, we can estimate LR_i optimally by

$$\widehat{\text{LR}}_i = \frac{P(H_1|s_i)}{1-P(H_1|s_i)} \times \frac{n_{\text{imp}}}{n_{\text{gen}}} \tag{3.7}$$

as used in [4] for calibrating scores in the field of speaker recognition. Therefore, we only need to estimate the correction factor CF.

3.2. Fixed FAR fusion

Estimating CF can be done by estimating c_{gen} and c_{imp} separately. Of course we will not estimate these copula densities nonparametrically since it will lead to the same problems as when estimating the original density functions directly. We will approximate CF by the following parametric copulas: Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), and Gumbel (Gu). We also include the independence copula (ind) to guarantee that our fusion is better than the PLR method. Readers interested in copulas are referred to [9] for a more detailed explanation. To have more dependence models and because the Clayton and Gumbel copulas are not symmetric, their flipped forms (flipped Clayton (fCl) and flipped Gumbel (fGu)) will be included as well (if U has copula C then $1 - U$ has copula flipped C). Therefore, the copulas c_{gen} and c_{imp} are chosen from the copula family

$$\mathcal{C} = \{\text{ind}, \text{GC}, \text{t}, \text{Fr}, \text{Cl}, \text{Gu}, \text{fCl}, \text{fGu}\}.$$

Note that the best copula pair must have the best performance among other pairs in the sense that it has the highest TPR at fixed FAR. Applying a goodness-of-fit test as provided in [7] will only give the copula pair that is closest to the pair $(c_{\text{gen}}, c_{\text{imp}})$, but whose ratio is not necessarily closest to the ratio $c_{\text{gen}}/c_{\text{imp}}$. Therefore, we propose to choose the best copula pair directly by maximizing the empirical TPR at the given FAR$= \alpha$ as explained in Section 2. Given a fixed FAR $= \alpha$, a set $\mathcal{C}$ of n_c candidate copulas and a training set, our fixed FAR fusion is very simple. The first step is computing PLR by the PAV algorithm and multiplying it by each of all copula pairs $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$ in which the dependence parameters have been estimated by the PMLEs as defined in (3.4). Of the $n_c \times n_c$ resulting different combined scores we choose the one that maximizes the TPR.

4. Experimental Results

To study the performance of our fixed FAR fusion in improving the simple PLR method we apply it to synthetic and real databases, which are split up into training and testing sets. Given a training set, we will compute the product of the individual likelihood ratios and select the best copula pair. The corresponding testing set is used for evaluation only. We compare our fixed FAR fusion to the linear Logit fusion explained in [13] and the GMM fusion proposed in [14] at FAR$= 0.01\%$ for all experiments. The Jeffreys CIs of all fusions are computed at significance level 0.01 and the improvement of fusion ψ compared to PLR fusion in TPR at 0.01% FAR is defined by $[L_\psi - U, U_\psi - L]$ where $[L_\psi, U_\psi]$ and $[L, U]$ are the 99% Jeffreys CIs of fusion ψ and PLR fusion, respectively, as explained in Section 2.

Given genuine and impostor scores

$$\mathbf{W}_1, \ldots, \mathbf{W}_{n_{\text{gen}}}$$

and

$$\mathbf{B}_1, \ldots, \mathbf{B}_{n_{\text{imp}}}$$

in the training set, our procedure to choose the best copula pair is simple. We randomize the genuine (impostor) scores and take two disjoint subsets with size

$$n_b = \min\{10,000; \lfloor n_{\text{gen}}/2 \rfloor\}$$

and

$$n_w = \min\{10,000; \lfloor n_{\text{imp}}/2 \rfloor\}.$$

This re-sampling method is aimed at increasing the computation speed because it will be repeated 100 times to see the consistency. Once the product of the individual likelihood ratios is computed, it is multiplied by the 64 copula pair estimates $\hat{c}_{\text{gen}}/\hat{c}_{\text{imp}}$. After all 64 combined scores are obtained using the first subset, the empirical TPR at 0.01% FAR is then computed. The final TPR for each copula pair is the average over all 100 experiments. The best copula pair is the pair having the highest average of the TPR values. If there are several pairs having the same averages, we choose the pair with the smallest variance. If there is still more than one pair having the smallest means and variances then we choose one of them at random.

4.1. Synthetic Data

To get synthetic data that behave like real data, we take two algorithms presented in [20]. The first algorithm measures the similarity of the left half of the face between two images and the second one the similarity of the right half. The density and distribution functions of the genuine and impostor scores for each algorithm are estimated by a mixture of logconcave densities [5]. We choose this estimation method because it is more general than a Gaussian mixture and more robust for handling skewness. To obtain scores with *explicit* dependence that can be represented by a copula C, we generate random samples of the copula C and apply the inverse transform technique, using the estimates of the two marginal distribution functions. In this way the generated scores have as marginal distribution functions these estimates of the distribution functions of data generated by the two algorithms. Recall that if F is a continuous distribution function then U is uniformly distributed if and only if $F^{-1}(U)$ has distribution function F.

In our experiment, we generate 10,000 genuine and 1,000,000 impostor scores in the way as explained above. The dependence is made by putting 4 different copula pairs

$$\{(\mathrm{GC}, \mathrm{GC}), (\mathrm{t}, \mathrm{fCl}), (\mathrm{fGu}, \mathrm{GC}), (\mathrm{Cl}, \mathrm{Gu})\}$$

completed with 9 dependence level pairs obtained from the cross pairs

$$\{\mathrm{low}, \mathrm{moderate}, \mathrm{high}\}.$$

In order to know the effect of dependence in biometric fusion, the low, moderate, and high dependence levels are set to have correlation values 0.1, 0.5, and 0.9 for Gaussian and Student's t copulas while for other copulas we put parameters 1, 10, and 50. Student's t copula has 3 degrees of freedom for all experiments.

By following our procedure, we get that the best copula pair is the true one for every experiment. Then, the fixed FAR fusion is compared to the PLR fusion to see the gain of considering dependence in biometric fusion. Figure 1 shows the improvement by the fixed FAR fusion compared to the PLR fusion. We can see that we really have to take the dependence into account when the dependence between the impostor scores is higher than between the genuine ones. Moreover, the dependence between classifiers should be taken into account even for low levels of dependence.

4.2. FVC2002-DB1 database

This data set [12] consists of 100 fingers with 8 impressions per finger. We will use the same experimental set up as used in [15] by putting the first two impressions as templates and the remaining ones as queries. Two 600×100 scores matrices are obtained by matching each query to both the templates using a minutiae matcher [1]. The purpose of this experiment is to see the improvement in using our fixed FAR method for multi-instances scenarios. To have a big enough testing set so that the CIs are not too large, we did 1,000 experiments. In every experiment, we randomized the 100 subjects, and took 70 subjects for training and the remaining 30 for testing. Our fixed FAR and benchmark fusion methods were trained on the first subset and evaluated on the second subset. As a result, each fusion method has 180 genuine and 5,220 impostor scores for every experiment. The average TPR is computed by pooling all genuine scores from the 1,000 experiments in one set and all impostor scores in the other set [6]. Therefore, we have 180,000 genuine and 5,220,000 impostor scores in total.

For every experiment, we train our fixed FAR fusion method by following the procedure explained at the beginning of this section and the pair (ind,fCl) is obtained as the best copula pair. The difference of the area under ROC of our fixed FAR and the PLR fusion is around 0.1%, which is relatively small. At first sight it is consistent with the results in [15] , which claims that considering dependence will not improve the PLR fusion significantly. However, if we highlight the TPR at FAR= 0.01% (see Figure 2), we can see that the improvement is significant. Detailed TPR values for our fixed FAR and benchmark fusions are provided in Table 1. On this database, our fixed FAR fusion is slightly better than the GMM fusion and both of them improve the PLR fusion at significance level 0.01. On the other hand, the Logit and PLR fusions have almost the same performances.

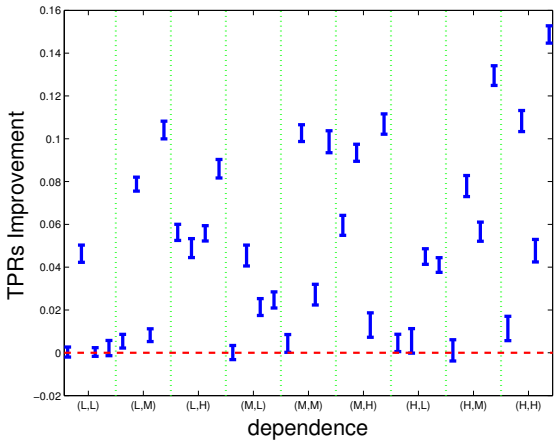

Figure 1: Gain of considering dependence between classifiers. The blue thick lines are the 99% Jeffreys CI of fixed FAR fusion compared to PLR fusion. The blue thick lines that do not intersect the red dashed line, mean that the gain of considering dependence is significant. On the x-axis the databases are indicated in 9 groups of 4, each group having the same dependence level pair for each of the 4 chosen copula pairs. Database (L,L) has low and low dependence levels for genuine and impostor scores, (L,M) low and moderate, (L,H) low and high, etc.

4.3. NIST-face database

The NIST-face BSSR1 database is published by the National Institute of Standards and Technology [16]. The data contain similarity scores from two face algorithms run on images from 3,000 subjects with each subject having two probe images and one gallery image. To evaluate the performance of our benchmark fusion strategies, we randomize the subjects and split the set into two disjoint sets with size 1,500 each. Each fusion strategy is trained on the first subset and evaluated on the second subset. This procedure is repeated 10 times. Then, we col-

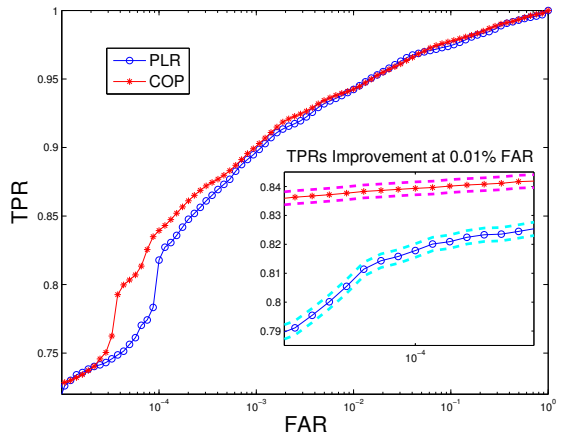

Figure 2: Comparison between the PLR and our fixed FAR fusion methods on FVC2002-DB1 database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

Table 1: PERFORMANCES AT 0.01% FAR ON FVC2002-DB1.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	77.5%	N/A
PLR	81.8%	N/A
Logit	81.9%	[−0.4%, 0.6%]
GMM	83.6%	[1.3%, 2.3%]
FFF	**83.9%**	[1.7%, 2.6%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion. The bold number is the best one and the underlined number is the worst one.

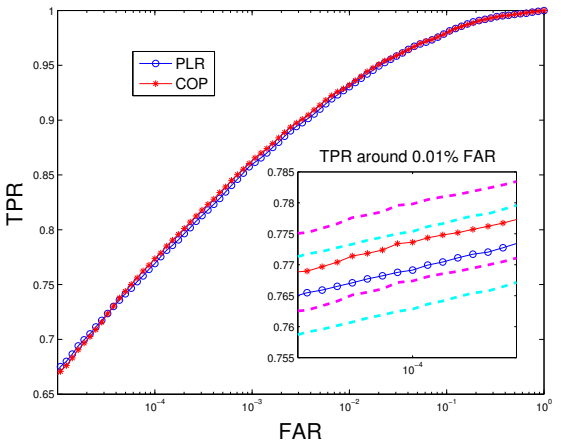

Figure 3: Comparison between the PLR and our fixed FAR fusion methods on NIST-face database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

Table 2: PERFORMANCES AT 0.01% FAR ON NIST-FACE DATABASE.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	71.2%	N/A
PLR	76.9%	N/A
Logit	76.1%	[−2.0%, 0.5%]
GMM	76.8%	[−1.4%, 1.1%]
FFF	77.4%	[−0.8%, 1.7%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion.

lect all genuine scores from all 10 experiments in one set and all impostor scores in another set resulting in 30,000 genuine and 44,970,000 impostor scores.

Figure 3 shows that the ROC of our fixed FAR fusion method almost coincides with the ROC of the PLR fusion. Although our fixed FAR fusion has the highest TPR, we should not conclude that it is the best one because all 99% Jeffreys CIs are overlapping (see Table 2). This means that on this database, the simple PLR fusion method is comparable to other fusion methods that take dependence into account.

4.4. Face3D database

This database is used in [19, 20] for 3D face recognition. It is quite realistic for biometric verification because both the training and the testing set contain very different images (taken with different cameras, backgrounds, poses, expressions, illuminations and time). In his papers, the author proposes 60 different classifiers by measuring the similarity of different regions. In our experiment, we only take 5 regions out of these 60: similarity of the full face, the left half, the right half, the bottom part, and the upper part. The results of these 5 algorithms are rather correlated, of course. This choice is made to see the performance of our benchmark methods in handling the dependence between classifiers. By following our procedure, we get as the best copula pair (ind,Fr).

Figure 4 shows clearly that considering dependence can improve the performance significantly. We can see that our fixed FAR fusion method is the only fusion strategy that can handle the dependence on this database as given in Table 3. While our fixed FAR fusion performs very well in handling the dependence, the GMM fusion is even worse than the best single matcher. This happens because the estimated number of components in the GMM is equal

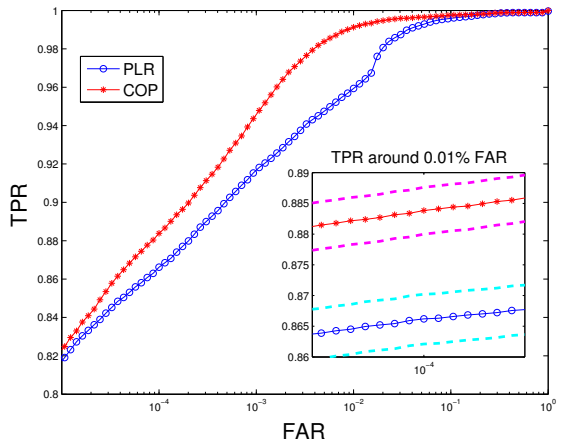

Figure 4: Comparison between the PLR and our fixed FAR fusion methods on Face3D database. The small box contains the highlighted performance at around 0.01% FAR. The dashed lines are the 99% Jeffrey CIs.

Table 3: PERFORMANCES AT 0.01% FAR ON FACE3D DATABASE.

Methods	TPR	99% Jeffreys CI compared to PLR in TPR at 0.01% FAR
BSM	84.9%	N/A
PLR	86.6%	N/A
Logit	87.6%	[0.1%, 1.7%]
GMM	81.2%	[−6.3%, −4.5%]
FFF	**88.4%**	[1.0%, 2.6%]

BSM: Best Single Matcher, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios, FFF: our fixed FAR fusion. The bold number is the best one and the underlined number is the worst one.

to the the maximum value (20) of it when being estimated by the minimum message length criterion as proposed in [8]. It means that the number of components may be more than 20. However, if we increase the number of components then the estimator becomes less reliable.

5. Conclusion

We have proposed and used an alternative method for evaluating the performance of biometric fusion methods at fixed FAR using Jeffreys credible intervals. We have also proposed a fixed FAR fusion method to improve via parametric copulas the PLR fusion strategy. From a simulation study with synthetic data, we have concluded that it is always useful to take the dependence into account even for low dependence levels. It has also been shown that our fixed FAR fusion method is the best method on real databases compared to the GMM and Logit fusion methods, which are also designed to handle dependence. Instead of providing a "rule of thumb" whether the dependence in biometric fusion has to be taken into account or not, we propose to always check whether our fixed FAR method improves on the PLR fusion method by a simple test as follows: define relevant training and testing sets, follow our procedure in choosing the best copula pair on the training set, and finally check the significance improvement using our evaluation method on the testing set. We can see from the FVC2002-DB1 database that the existing rule of thumb concludes the unimportance in considering dependence. However, when the FAR value is fixed (0.01%), we get a significant improvement of around 82% to 84% (around 2%). Although it is a relatively small improvement, our fixed FAR fusion method reduces the number of people that have to be checked manually from 18 to 16 for every 100 people. This means that if the manual checking needs 10 minutes per person then we save 20 minutes for every 100 people.

Acknowledgements

This research was supported by the Netherlands Organisation for Scientific Research (NWO) via the project Forensic Face Recognition, 727.011.008.

References

[1] J. Abraham, J. Gao, and P. Kwan. *Fingerprint Matching Using A Hybrid Shape and Orientation Descriptor*. INTECH Open Access Publisher, 2011.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statist. Sci.*, 16(2):101–133, 05 2001.

[4] N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2):230–275, 2006.

[5] G. T. Chang and G. Walther. Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, 51(12):6242 – 6251, 2007.

[6] T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[7] J.-D. Fermanian. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1):119 – 152, 2005.

[8] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[9] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1997.

[10] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[11] Y. Ma, B. Cukic, and H. Singh. A classification approach to multi-biometric score fusion. In *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication*, AVBPA'05, pages 484–493, Berlin, Heidelberg, 2005. Springer-Verlag.

[12] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 811–814 vol.3, 2002.

[13] G. S. Morrison. Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197, 2013.

[14] K. Nandakumar, Y. Chen, S. Dass, and A. Jain. Likelihood ratio-based biometric score fusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):342–347, Feb 2008.

[15] K. Nandakumar, A. Ross, and A. K. Jain. Biometric fusion: Does modeling correlation really matter? In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1–6, Sept 2009.

[16] National Institute of Standards and Technology. Nist biometric scores set - release 1, 2004. Available at http://www.itl.nist.gov/iad/894.03/biometricscores.

[17] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, 1933.

[18] M. Sklar. *Fonctions de Répartition À N Dimensions Et Leurs Marges*. Université Paris 8, 1959.

[19] L. Spreeuwers. Fast and accurate 3D face recognition. *International Journal of Computer Vision*, 93(3):389–414, 2011.

[20] L. Spreeuwers. Breaking the 99% barrier: optimisation of three-dimensional face recognition. *Biometrics, IET*, 4(3):169–178, 2015.

[21] N. Susyanto, C. A. J. Klaassen, R. N. J. Veldhuis, and L. J. Spreeuwers. Semiparametric score level fusion: Gaussian copula approach. In *Proceedings of the 36th WIC Symposium on Information Theory in the Benelux, Brussels*, pages 26–33, Brussels, May 2015. Université Libre de Bruxelles.

[22] Q. Tao and R. N. J. Veldhuis. Robust biometric score fusion by naive likelihood ratio via receiver operating characteristics. *IEEE Transactions on Information Forensics and Security*, 8(2):305–313, February 2013.

[23] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 694–699, New York, NY, USA, 2002. ACM.