Two-step calibration method for multi-algorithm score-based face recognition systems by minimizing discrimination loss

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

DOI
10.1109/ICB.2016.7550094

Publication date
2016

Document Version
Final published version

Published in
2016 International Conference on Biometrics (ICB)

Two-step Calibration Method for Multi-algorithm Score-based Face Recognition Systems by Minimizing Discrimination Loss

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

We propose a new method for combining multi-algorithm score-based face recognition systems, which we call the two-step calibration method. Typically, algorithms for face recognition systems produce dependent scores. The two-step method is based on parametric copulas to handle this dependence. Its goal is to minimize discrimination loss. For synthetic and real databases (NIST-face and Face3D) we will show that our method is accurate and reliable using the cost of log likelihood ratio and the information-theoretical empirical cross-entropy (ECE).

1. Introduction

The likelihood ratio (LR) approach of evidence evaluation is increasingly accepted in forensic science [7]. The LR of evidence e is defined as the ratio between the probability of the evidence given *prosecution* and *defense* hypotheses, i.e,

$$LR(e) = \frac{P(e|H_{\rm p})}{P(e|H_{\rm d})} \tag{1.1}$$

where $H_{\rm p}$ and $H_{\rm d}$ are two mutually exclusive hypotheses respectively supporting whether or not the suspect is the donor of the biometric trace. This quantitative value is computed by a forensic scientist and can be used to support the fact finder (judge/jury) in court to make an objective decision. The Bayesian framework explains elegantly how the LR supports the decision via relation

$$\frac{P(H_{\rm p}|e)}{P(H_{\rm d}|e)} = \frac{P(e|H_{\rm p})}{P(e|H_{\rm d})} \times \frac{P(H_{\rm p})}{P(H_{\rm d})}. \tag{1.2}$$

This means that the LR can be interpreted as a multiplicative factor for the information before analyzing the evidence (*prior odds*) to get the new information after taking the evidence into account (*posterior odds*).

In this paper we are studying *multi-algorithm score-based face recognition systems*, in which two or more different algorithms compute a similarity score for any pair of face images. This means that the evidence e is a vector of scores in which every score describes the similarity of the image found at the crime scene and an image of the suspect. It is intuitively understandable that combining several algorithms might be advantageous. For instance, every individual algorithm can be selected for its good performance under a specific condition, such as varying pose, illumination, or robustness. Therefore, an appropriate combination is hoped to integrate the complementary information of the individual algorithms. Indeed, several studies [12, 22, 23] show that a multi-algorithm method might enhance the recognition performance.

Several methods of deriving the LR from a biometric comparison score, which is also called *calibration*, have been proposed and evaluated for single-algorithm face recognition systems; see [1, 2] for a survey of these methods. In contrast, to the best of our knowledge, there is no method of combining two or more face recognition systems for forensic evidence evaluation. In this paper, we propose such a method, which we will call the two-step calibration method, for calibrating multi-algorithm face recognition systems via parametric copulas. We will compare our method to the *linear* logistic regression (logit) method, which is commonly used in the field of speaker recognition [15, 16], and also to the Gaussian Mixture Model (GMM) [17] and the simple Product of Likelihood Ratios (PLR) [25], which are originally proposed in biometric fusion for person authentication. We also show through simulation and real data that the logistic regression method used in the field of speaker recognition [15, 16] is not recommendable for use in forensic face scenarios.

The rest of this paper is organized as follows. Section 2 reviews the cost of log likelihood ratio and the ECE plot, which measure the accuracy and re-

978-1-5090-1869-7/16/$31.00 ©2016 IEEE

liability of calibration methods. Our two-step calibration method is presented in Section 3. Section 4 demonstrates the excellent performance of our method for both synthetic and real databases. Finally, our conclusions are presented in Section 5.

2. Performance Evaluation of Likelihood Ratio Computation

There are two types of measures for the reliability of calibration methods: *application-dependent* [5, 14] and *application-independent* [4, 26, 19, 20] measures. Since forensic scientists do not have access to the prior odds, we will focus on application-independent ones.

2.1. Cost of log likelihood ratio

The cost of log likelihood ratio (C_{llr}) is introduced by Brümmer and du Preez[4] in the field of speaker recognition, is based on a generalization of cost evaluation metrics, and is used in forensic face scenarios in [13]. This measure may be interpreted as a summary of a LR computation [3]. Note that a face recognition system does not necessarily produce a similarity score as an LR value. Thus, a calibration is needed to make this *original score* interpretable as an accepted measure of strength of evidence in court by mapping it into LR value, which we also call *LR score*. A score is called *genuine* if it is associated to 2 images of the same person, and is called an *impostor* score if it involves 2 images of two different persons. Let $\mathcal{M}$ denote a method to calibrate original scores into LR values. Given a set of scores, let $\mathcal{LR}_{\mathrm{p}}$ denote the set of N_{gen} genuine $\mathcal{M}$-calibrated scores, which correspond to the hypothesis of the prosecution, and $\mathcal{LR}_{\mathrm{d}}$ the set of N_{imp} impostor $\mathcal{M}$-calibrated scores, which correspond to the hypothesis of the defense. The cost of log likelihood ratio C_{llr} is defined by

$$C_{\mathrm{llr}} = \frac{1}{2N_{\mathrm{gen}}} \sum_{LR \in \mathcal{LR}_{\mathrm{p}}} \log_2 \left(1 + \frac{1}{LR}\right) + \frac{1}{2N_{\mathrm{imp}}} \sum_{LR \in \mathcal{LR}_{\mathrm{d}}} \log_2 \left(1 + LR\right). \quad (2.1)$$

To explain the name of this metric we note that LR in formula (2.1) may be rewritten in terms of the logarithm of LR. Interestingly, this metric can be decomposed into a *discrimination* and *calibration* form via relation

$$C_{\mathrm{llr}} = C_{\mathrm{llr}}^{\mathrm{min}} + C_{\mathrm{llr}}^{\mathrm{cal}}. \quad (2.2)$$

Here, $C_{\mathrm{llr}}^{\mathrm{min}}$ and $C_{\mathrm{llr}}^{\mathrm{cal}}$ denote the discrimination and calibration loss, respectively. Discrimination loss is the opposite of discrimination power (the ability of the system to distinguish between genuine and impostor scores). The smaller the value of this quantity, the higher the discrimination power. The $C_{\mathrm{llr}}^{\mathrm{min}}$ is defined as the minimum C_{llr} value under evaluation by preserving the discrimination power which is attained by the Pool-Adjacent-Violators (PAV) algorithm as proved in [4]. Therefore, the $C_{\mathrm{llr}}^{\mathrm{min}}$ is computed by plugging the $\mathcal{M}$-calibrated scores after PAV transformation into (2.1). On the other hand, calibration loss indicates the calibration performance on a separate evaluation set.

2.2. ECE plot

The Empirical Cross Entropy (ECE) plot is an application-independent method of measuring the reliability of calibration with an information theoretical interpretation [19]. The ECE function is defined as a function of the log prior odds by

$$ECE(lp) = \frac{1}{2N_{\mathrm{gen}}} \sum_{LR \in \mathcal{LR}_{\mathrm{p}}} \log_2 \left(1 + \frac{1}{LR \times e^{lp}}\right) + \frac{1}{2N_{\mathrm{imp}}} \sum_{LR \in \mathcal{LR}_{\mathrm{d}}} \log_2 \left(1 + LR \times e^{lp}\right) \quad (2.3)$$

for every $lp \in (-\infty, \infty)$. Clearly $C_{\mathrm{llr}} = ECE(0)$ holds, which shows that the ECE generalizes the cost of log likelihood ratio.

Figure 1 is an example of the ECE plot of a system. The solid red curve represents the performance of the calibration, the dashed blue curve is the minimum ECE value under evaluation by preserving the discrimination power which is attained by PAV transformation, and the dashed black curve is the entropy of the neutral system without considering the evidence, i.e., all LR values equal to 1. The difference between the solid red and dashed blue curves is the calibration loss. Since the ECE value can be interpreted as the average information loss by taking the system into account, we can see that the system will lose more information than the neutral system for log prior odds greater than 2. Therefore, forensic scientists should provide the usual LR and also explain to the fact finder that he should not use the forensic system if his log prior odds are greater than 2.

3. Evidential Value Computation of Multi-algorithm Systems by Minimizing Discrimination Loss

This section explains how to get calibrated scores for d-algorithm face recognition systems, i.e, computing the LR at evidence $e = (s_1, \cdots, s_{\mathrm{d}})$. Of course, if the joint density functions of the evidence under both hypotheses, which will be denoted by f_{gen} and f_{imp} for genuine and impostor scores, respectively, are known then the exact LR can be easily

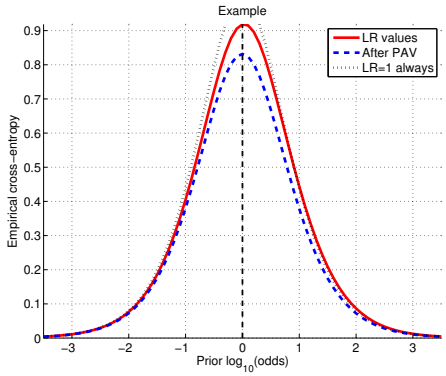

Figure 1: Example of ECE plot

obtained. However, in practice, these density functions have to be estimated from data. This classical problem in statistics can be solved by parametric (e.g., normal distribution, Weibull distribution) and nonparametric (e.g., histogram, kernel density estimation) models. However, the choice of an appropriate parametric model is sometimes difficult while nonparametric estimators suffer from the difficulty that they are sensitive to the choice of the bandwidth or of other smoothing parameters, especially for our multivariate case. Therefore, it is natural to approach our estimation problem semiparametrically, modelling the marginal densities nonparametrically and the dependence between them by parametric copulas.

3.1. Dependence through Copula

Mathematically, a copula is a distribution function on the unit cube $[0,1]^d$, $d \geq 2$, of which the marginals are uniformly distributed. In practice, it is widely used to describe the dependence of random variables; see e.g. [6, 8] for application in econometrics and finance. In biometric fusion, Susyanto et al. [24] use a specific copula called Gaussian copula to handle the dependence between classifiers. A classical result of Sklar [21] relates any continuous multivariate distribution function to a copula.

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 C so that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \quad (3.1)$$

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 (3.1):

$$h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times \prod_{i=1}^{d} f_i(x_i) \quad (3.2)$$

where c is the copula density and f_i is the i-th marginal density for every $i = 1, \cdots, d$. We can see that the density h is a product of the copula density depending only on the marginal distributions $F_1, \cdots, F_d$ and its marginal densities. It means that 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 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

$$\frac{1}{n} \sum_{i=1}^{n} \log c_\alpha \left(\hat{F}_1(X_{1i}), \ldots, \hat{F}_d(X_{di}) \right) \quad (3.3)$$

where

$$\hat{F}_j(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{\{X_{ji} \leq x\}}, \quad \forall 1 \leq j \leq d$$

is a *modified* empirical distribution function 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 (1.1) and (3.2), the likelihood ratio at $e = (s_1, \cdots, s_{\text{d}})$ can be written as

$$LR(e) = \frac{c_{\text{gen}}(F_{\text{gen},1}(s_1), \cdots, F_{\text{gen},d}(s_{\text{d}}))}{c_{\text{imp}}(F_{\text{imp},1}(s_1), \cdots, F_{\text{imp},d}(s_{\text{d}}))} \times \prod_{i=1}^{d} LR_i(s_i) \quad (3.4)$$

where $F_{\text{gen},i}$ and $F_{\text{imp},i}$ denote the distribution functions of genuine and impostor scores, respectively. The i-th individual LR can be computed for each $i = 1, \ldots, d$ by the PAV algorithm, which is optimal for calibrating 1-dimensional scores. Therefore, we only need to estimate the first factor at the right hand side of (3.4): the copula part.

3.2. Two-step Calibration Methods

As noted before, the density functions f_{gen} and f_{imp} have to be estimated, which implies that the copula densities c_{gen} and c_{imp} must be estimated as well. Estimating copula density functions nonparametrically will lead to the same problems as when estimating the original density functions directly. Therefore, we will approximate the copula part of (3.4) by some well-known parametric copulas. We will use the following copulas: Gaussian copula (GC), Student's t (t), Frank (Fr), Clayton (Cl), and Gumbel (Gu). We also include the independent (ind) to guarantee that our combined system is better than the simple product of likelihood ratios. Readers interested in copulas are referred to [11] 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

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

We can choose the best copula for c_{gen} and c_{imp} from the family $\mathcal{C}$ for f_{gen} and f_{imp} by a goodness-of-fit test as provided in [9]. However, this method only guarantees that the selected copulas are closest to c_{gen} and c_{imp}, but not necessarily good enough to model $c_{\text{gen}}/c_{\text{imp}}$. Therefore, we propose to choose the best copula pair directly by minimizing the discrimination loss as explained in Section 2.1. Since the estimated copula pair is not the true copula and only minimizing the discrimination loss among other pairs, the combined scores can be poorly calibrated. To solve this problem, we apply the PAV algorithm once combined scores have been obtained via the best copula pair.

Given a set $\mathcal{C}$ of n_c candidate copulas and a training set, our two-step calibration method is very simple. The first step is computing the product of the individual likelihood ratios by the PAV algorithm and multiplying this product 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 (3.3). Of the $n_c \times n_c$ resulting different combined scores we choose the one that minimizes the discrimination loss. The second step is transforming the combined scores by the PAV algorithm so that the final scores have high discrimination power and are also well-calibrated.

4. Experimental Results

To study the performance of our two-step calibration 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, select the best copula pair, and calibrate the combined scores. The corresponding testing set is used for evaluation only. The ECE plot is chosen for evaluation because it is more general than the cost of log likelihood ratio. On the real databases, beside plotting the ECE curves, we also highlight the discrimination loss and the ECE values for log prior odds -2, 0, and 2; see Table 1. We compare our two-step method to the logit method studied in [16] and the GMM method where the number of the mixture components is automatically estimated by the minimum message length criterion as proposed in [10]. For all experiments, the maximum value of the number of the mixture components is 20.

Given genuine and impostor scores

$$W_1, \ldots, W_{n_{\text{gen}}}$$

and

$$B_1, \ldots, B_{n_{\text{imp}}}$$

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

$$n_b = \min\{10000, \lfloor n_{\text{gen}}/2 \rfloor\}$$

and

$$n_w = \min\{10000, \lfloor n_{\text{imp}}/2 \rfloor\}.$$

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 discrimination loss is then computed. The final discrimination loss for each copula pair is the average over all 100 experiments. The best copula pair is the pair having the smallest average of the discrimination loss 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 [23]. 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 kernel density estimation. To obtain scores with *explicit* dependence

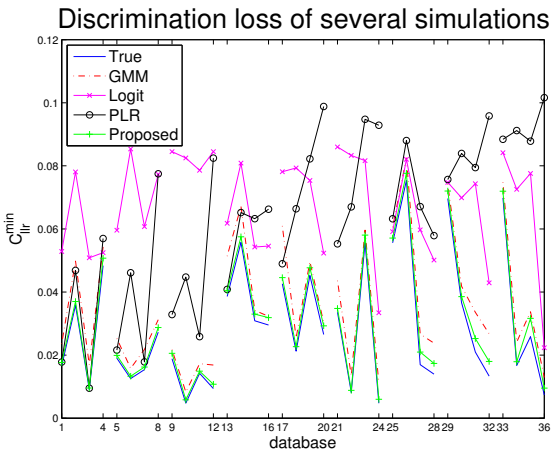

Figure 2: Performance on synthetic data. On the x-axis the databases are indicated in 9 groups of 4, each group having the same dependence level pair for each of the 4 chosen copula pairs. Database 1-4 has low and low dependence levels for genuine and impostor scores, 5-8 low and moderate, 9-12 low and high, 13-16 moderate and low, etc.

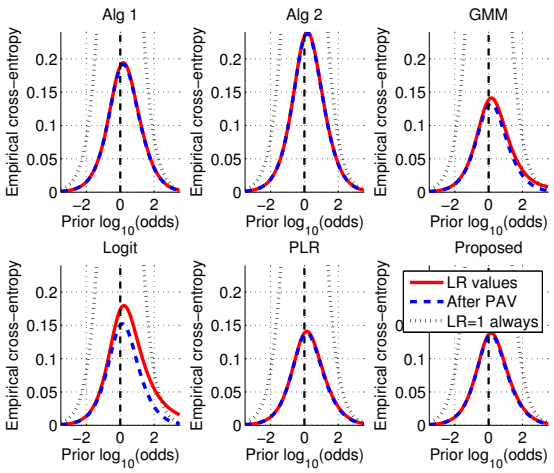

Figure 3: ECE plot of NIST-face

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

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

completed with 9 dependence level pairs obtained from the cross pairs $\{low, moderate, high\}$. 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 1, 10, and 50. Student's t copula has 3 degrees of freedom for all experiments.

Figure 2 is the plot of discrimination loss of our 36 simulated databases. The true LR can be computed exactly because the underlying distributions are known. We can see that our two-step method outperforms the others. As expected, the PLR method performs poorly when the dependence between algorithms is moderate or high because much information will lose by assuming the independence between algorithms. The GMM method is the second best method but the computation time is much longer than for the two-step method. We can also see that the logistic regression method, which is commonly used in the field of speaker recognition, has the worst performance among all methods.

4.2. NIST-face BSSR1 database

The NIST-face BSSR1 database is published by National Institute of Standards and Technology [18]. The data contain similarity scores from two face systems run on images from 3000 subjects with each subject having two probe images and one gallery image. We only take 2992 subjects because the scores of the other 8 subjects on the first system are always -1. It is reported that these images are not accepted by the facial recognition system and are therefore excluded. To evaluate the performance of our benchmark calibration methods, we randomize the subjects and split the set into two disjoint sets with size 1496. We follow the procedure explained at the beginning of this section and the pair (ind, GC) is obtained as the best copula pair. We repeat this experiment 20 times and all of them give almost the same result. Therefore, we decided to show only one of these results.

The GMM, PLR, and our two-step method have almost the same performance as seen from the ECE plot in Figure 3. Although the logit method performs reasonably well for small values of the log prior odds, it has the highest calibration loss among all methods and it is even dangerous for use in forensic face scenarios for large values of the log prior odds (greater than 2). By only considering the discrimination loss, Table 1 of the NIST-face part tells us that the GMM is the best calibration method. However, the ECE values show that the two-step method is actually the best one.

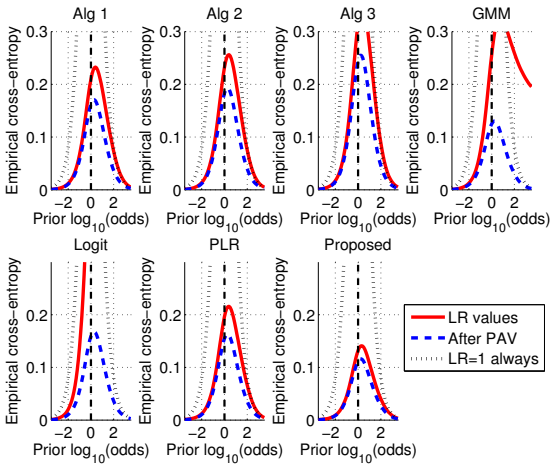

Figure 4: ECE plot of Face3D

4.3. Face-3D database

This database is used in [22, 23] for 3D face recognition. It is quite realistic for the forensic face problem, 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 classifiers operating on 30 different facial regions with 2 different image registration methods. In our experiment, we only take 3 classifiers out of these 60: similarity of the full face, the left and the right half. The results of these 3 classifiers are rather correlated, of course. This choice is made to see the performance of our benchmark methods in handling the dependence among classifiers. Although they are not different algorithms, we use these 3 classifiers to have different types of data and see the performance our two-step method for multi-classifiers scenario as well. By following our procedure, we get as the best copula pair (ind, t). The performances on this database are provided by Figure 4 for the ECE plot and Table 1 for the discrimination loss and some values of the ECE.

We can see that our two-step method is the best one using all evaluation metrics ($C_{\text{llr}}^{\text{min}}$, C_{llr}, ECE at small prior odds, and ECE at high prior odds). As before, the logit method performs poorly on this database since the underlying distributions are not gaussian. Surprisingly, the GMM method also has poor performance on this database; it is even worse than the simple PLR. This may be because the number of the mixture components is more than the the maximum value that we set. However, if we increase the number of components then we will have a problem with the limitation of the sample size.

Methods	NIST-Face				Face 3D			
	$C_{\text{llr}}^{\text{min}}$	ECE			$C_{\text{llr}}^{\text{min}}$	ECE		
		-2	0	2		-2	0	2
BSS	0.187	0.153	0.189	0.037	0.162	0.012	0.210	0.073
GMM	**0.131**	0.123	0.139	0.033	0.125	0.013	0.256	0.260
Logit	0.150	0.138	0.174	0.055	0.159	0.020	0.598	0.828
PLR	0.137	0.123	0.139	0.028	0.155	0.012	0.197	0.066
Proposed	0.134	**0.120**	**0.136**	**0.027**	**0.112**	**0.009**	**0.132**	**0.040**

Table 1: Discrimination loss and ECE of different methods on the real databases. BSS: Best Single System, GMM: Gaussian Mixture Model, Logit: Logistic Regression, PLR: Product of Likelihood Ratios. The bold number is the best one in every column.

5. Conclusion

We propose a two-step calibration method to compute the likelihood ratio of multi-algorithm score-based face recognition systems in forensic evidence evaluation. The first step of the two-step method is computing the product of the individual likelihood ratios multiplied by the density ratio of the best copula pair determined by minimizing discrimination loss. The simple second step is applying the PAV algorithm in order to get well-calibrated scores. Using several synthetic data sets, we have shown that our approach performs very well in handling all dependence levels (low, moderate, and high). We also see that our two-step method on the real databases NIST-face BSSR1 and Face3D. We conclude that the GMM method, which works quite well in biometric fusion for person authentication, can somehow perform poorly in forensic face scenarios. We also recommend to avoid the logistic method, which is commonly used in the field of speaker recognition, to compute the likelihood ratio in forensic face recognition because it has high discrimination loss and sometimes it is much worse than the neutral system.

References

[1] T. Ali, L. J. Spreeuwers, and R. N. J. Veldhuis. Forensic face recognition: A survey. In A. Quaglia and C. M. Epifano, editors, *Face Recognition: Methods, Applications and Technology*, Computer Science, Technology and Applications, page 9. Nova Publishers, 2012.

[2] T. Ali, L. J. Spreeuwers, R. N. J. Veldhuis, and D. Meuwly. Effect of calibration data on forensic likelihood ratio from a face recognition system. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Washington, DC, U.S.A.*, IEEE explore digital library, pages 1–8, United States, September 2013. IEEE.

[3] N. Brümmer and E. de Villiers. The BOSARIS toolkit: Theory, algorithms and code for surviving

the new DCF. *CoRR*, abs/1304.2865, 2013.

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

[5] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. The NIST speaker recognition evaluation overview, methodology, systems, results, perspective. *Speech Communication*, 31(23):225 – 254, 2000.

[6] P. Embrechts. Copulas: A personal view. *Journal of Risk and Insurance*, 76(3):639–650, 2009.

[7] O. E. Facey and R. J. Davis. Re: Expressing evaluative opinions; a position statement. *Science & Justice*, 51(4):212 –, 2011.

[8] Y. Fan, , and A. J. Patton. Copulas in econometrics. *Annual Review of Economics*, 6:179–200, 2014.

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

[10] 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.

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

[12] X. Lu, Y. Wang, and A. Jain. Combining classifiers for face recognition. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 3, pages III–13–16 vol.3, July 2003.

[13] M. I. Mandasari, M. Gunther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen. Score calibration in face recognition. *IET Biometrics*, 3(4):246–256, 2014.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. pages 1895–1898, 1997.

[15] G. S. Morrison. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus gaussian mixture model and universal background model (GMM-UBM). *Speech Communication*, 53(2):242 – 256, 2011.

[16] 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.

[17] 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.

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

[19] D. Ramos and J. Gonzalez-Rodriguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(13):156 – 169, 2013. EAFS 2012 6th European Academy of Forensic Science Conference The Hague, 20-24 August 2012.

[20] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken. Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58(6):1503–1518, 2013.

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

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

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

[24] 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.

[25] 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.

[26] D. A. van Leeuwen and N. Brümmer. Speaker classification i. chapter An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pages 330–353. Springer-Verlag, Berlin, Heidelberg, 2007.