Calibration of score based likelihood ratio estimation in automated forensic facial image comparison

Macarulla Rodriguez, A.; Geradts, Z.; Worring, M.

DOI
10.1016/j.forsciint.2022.111239

Publication date
2022

Document Version
Final published version

Published in
Forensic Science International

Forensic Science International 334 (2022) 111239

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Calibration of score based likelihood ratio estimation in automated forensic facial image comparison

Andrea Macarulla Rodriguez[a,b,*], Zeno Geradts[a,b], Marcel Worring[a,b]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497GB, The Netherlands
[b] University of Amsterdam, Science Park 904, Amsterdam 1098XH, The Netherlands

ARTICLE INFO

Article history:
Received 8 September 2021
Received in revised form 15 January 2022
Accepted 22 February 2022
Available online 7 March 2022

Keywords:
Facial image comparison
Calibration
Deep learning
Forensic science
Score based likelihood ratio

ABSTRACT

Forensic facial image comparison lacks a methodological standardization and empirical validation. We aim to address this problem by assessing the potential of machine learning to support the human expert in the courtroom. To yield valid evidence in court, decision making systems for facial image comparison should not only be accurate, they should also provide a calibrated confidence measure. This confidence is best conveyed using a score-based likelihood ratio. In this study we compare the performance of different calibrations for such scores. The score, either a distance or a similarity, is converted to a likelihood ratio using three types of calibration following similar techniques as applied in forensic fields such as speaker comparison and DNA matching, but which have not yet been tested in facial image comparison. The calibration types tested are: naive, quality score based on typicality, and feature-based. As transparency is essential in forensics, we focus on state-of-the-art open software and study their power compared to a state-of-the-art commercial system. With the European Network of Forensic Science Institutes (ENFSI) Proficiency tests as benchmark, calibration results on three public databases namely Labeled Faces in the Wild, SC Face and ForenFace show that both quality score and feature based calibration outperform naive calibration. Overall, the commercial system outperforms open software when evaluating these Likelihood Ratios. In general, we conclude that calibration implemented before likelihood ratio estimation is recommended. Furthermore, in terms of performance the commercial system is preferred over open software. As open software is more transparent, more research on open software is urged for.

1. Introduction

When face images are presented as evidence in court, the target most often is to interpret the result of the comparison between trace and suspect images. No standard method is, however, available for that task. The comparison technique, whether it is performed manually or using an automatic system, must meet legal requirements which vary per country [9,16,45]. Although the use of automatic systems is increasingly studied in the field of facial image comparison, for legal deployment it lacks standardization and validation. This is one of the reasons why cases of facial image comparison in court are currently still carried out manually by specialized facial image comparison experts [16,45]. Having a unified and validated method for interpreting scores by experts and machine can provide the standardization needed in court.

The Likelihood Ratio (LR) comes as a possible solution [26,36] for standardization, expressing the decision as the ratio of the probability given the evidence of a match against the probability of a non-match. Forensic experts endorse its use due to its compliance with the requirements of evidence-based forensic science: it is scientifically sound in particular it has transparent procedures, is testable, and it clearly separates the responsibilities of the forensic examiner and the court [1,39]. For evidence in speaker recognition, fingerprints and DNA analysis, a distance or similarity based biometric Score likelihood ratio (SLR) is being studied and used [4,23,31]. Here, we aim to realize a similar approach for facial image comparison. As explored in [40] and [3], automated systems for facial image comparison (especially when based on deep learning) combined with score-based likelihood ratio estimation have a great potential to help the forensic expert in the evaluative process [16].

In this paper, we make a number of contributions. We develop a pipeline that given a score produces an LR estimation that can be

* Corresponding author at: Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497GB, The Netherlands.
E-mail addresses: a.macarullarodriguez@uva.nl (A.M. Rodriguez), z.geradts@nfi.nl (Z. Geradts), m.worring@uva.nl (M. Worring).

https://doi.org/10.1016/j.forsciint.2022.111239
0379-0738/

compared to forensic experts and ENFSI participants. This serves as an SLR evaluation and validation for both open and commercial software. Thus we explore their differences and determine whether there is room for improvement in open software automated systems. Secondly, we estimate the influence of different LR calibrations in relation to resolution and image features, based mainly on surveillance images which is a major source of evidence in forensic cases. Calibration has been researched and used in speaker comparison [4,28] for similar types of voices. As identified in [3,16] similar treatment in faces has yet to be researched. Thirdly, we compare the Likelihood Ratio estimation from both open software as well as commercial software to a set of forensic experts in the ENFSI Proficiency Face Comparison test (which include case work related images such as surveillance) using the statistic elements of Cost Log Likelihood Ratio (Cllr) [5,6].

Fig. 1 gives an overview of the main topics presented in this paper.

2. Related work

We study related work by first considering how likelihood ratios are used in forensic fields other than facial image comparison. From there we consider how facial image comparison is currently being done. Finally, we look at the core step in standardization namely the calibration.

2.1. Likelihood ratio in forensics

Using a Bayesian probabilistic framework has been proposed in recent years as a logical and appropriate way to report evidence to a court of law [2,37,39]. The work of [42] states the requirements of evidence-based forensic science, which are: adoption of a basic-research model, design of experiments that test said model and the ability of experts to inform court about the relative strengths and weaknesses and suggestion on how that knowledge applies to individual cases. They also recommend that for machine learning data should be collected based on the frequency with which markings and attribute variations occur in different populations. The Likelihood Ratio has been proposed in recent decades as a method which addresses these requirements by providing transparent procedures and being testable, as indicated in the introduction [14,39]. When computed for a certain benchmark, different methods such as Cllr and ECE can be used to assess its predictions, see section 2.3 for more information about these methods. Score based procedures for the calculation of forensic likelihood ratios are popular across different branches of forensic science [29] especially in DNA [33], and speaker comparison [4,23,28]. They have two stages, first a function or model which takes measured features from known-source and questioned-source pairs as input and calculates scores as output, then a subsequent model which converts scores to likelihood ratios [29]. LR based on biometric similarity scores is referred to as Score based Likelihood Ratio (SLR) and defined as:

$$SLR = \frac{P(s|H_p, I)}{P(s|H_d, I)}, \tag{1}$$

where H_p is the null hypothesis or the prosecution hypothesis (evidence originates from the same source) and H_d is the alternative hypothesis or defense hypothesis (evidence originates from a different source). The value s is the score returned by the biometric system and I is the background information available in the case apart from the evidence. Although LR can be used for any type of forensic evidence (such as DNA or fingerprints), in our work it corresponds to face evidence.

According to [48], efforts to model or compensate the effects of adverse conditions in likelihood ratio computation should be improved. They evaluate the impact of these adverse conditions on glass samples. The analysis of [48] shows that integration of advanced machine-learning algorithms for the compensation of adverse conditions into forensic evaluation helps in this direction. They find this impact greatly affects calibration performance. There is a lack of a similar study in case of facial image comparison.

In [38] and [26], different LR validation methods are explored and analysed. The first question to consider is what and how to validate? In both papers, Cost Log Likelihood Ratio and ECE plot validation [6,39] are proposed as promising characteristics. ECE is exposed in [35] as a method which measures both discrimination and calibration, and shows its potential. It also describes how other related measures such as Confusion Entropy (CEN) or Matthews Correlation Coefficient (MCC) work with decision errors rather than probabilities. This implies the selection of a threshold and therefore they do not consider performance at different prior probabilities either. Other metrics are considered in [16], such as Tipett plots, Detection error trade-off (DET) and equal error rate (EER). They present an overview table summarizing the use and adequacy of these metrics for the assessment of model performance. In this overview, the graphical representation that scores the highest for both discrimination and calibration is again the ECE plot. In conclusion, for this work and according to the studied literature, the best indicators of both discrimination and calibration performances are Cllr and the ECE plots (explained in 3.3.1) [36,38] which give a good view of both the calibration and discrimination power of the forensic experts and the automated systems.

2.2. Facial image comparison in forensics

Facial image comparison in Forensics has been largely studied from a manual point of view [14]. There have been tentative approaches on automated systems performing this task, whether for intelligence, investigation, or evaluative purposes Zeinstra et al. [49] Ali [3], Tistarelli and Champod [45]. And facial image comparison has proven to have potential to help the forensic expert if the likelihood ratio estimation method is properly standardized and validated [40]. In manual comparison, four methods are typically used to analyse and compare faces: holistic, morphological and photo-anthropometric processes, along with direct superposition of the images [14].

These methods are not exclusive and can be combined in order to carry out the most exhaustive analysis with regard to the information available on the image. Recommendations in ENFSI practices are: out of these four methods, holistic comparison is only recommended when other more effective methods are not available, morphological (feature comparisons) is useful and recommended for facial image comparison. Both photo-anthropometric comparison and superposition are not recommended when using uncontrolled imagery.

Current face recognition systems [8,24,43], already reach very high levels of accuracy in public non-forensic benchmarks, and it is expected that in the coming years they will keep improving. If this improvement is accompanied with a standardization and proper validation in their decisions, they could become a powerful tool in Forensic Science [26]. An enforcement of this idea can be found in [16], where there is an extensive survey on the role of these automated system nowadays in the forensic field. They propose to improve the discussion between forensic expert, investigators and legal practitioners to best develop this method with respect to the needs and constraints of each.

2.3. Calibration in forensics

The calibration state of a model refers to the closeness of the computed value to the known value. Therefore, the calibration measures the extent to which the SLR points towards the correct

proposition. It has been used in other fields of Forensic Science such as speaker comparison, DNA analysis or fingerprints [6,38,39]. In [36], the problem of incorrect selection of databases is put forward. This problem is tackled in [30] for the speaker comparison case. It discusses what should be the implications of a good calibration and proposes ECE as the preferred method of validation. For facial image comparison this implies that ECE methods for evaluation are adequate for detection if the performance of both the automated model and the forensic participant are affected in the same way by the chosen calibration population.

In literature, the term -calibration- is used to describe two different processes. It usually refers to SLR as described in the introduction, or it can more specifically point to the subsequent process to adapt models which have high discriminating power but are poorly calibrated [38]. As this second step is essential to enhance the overall performance of a model, [16] poses that the term -calibration- should not differentiate between the steps of score-to-SLR and SLR-to-calibrated SLR. Instead methods should cover every computation used from the initial score to the final reported SLR regardless of the number of treatment steps needed. In this work, we evaluate the effects of selecting the database to perform said calibration, for which the second step is not required. We evaluate the first interpretation of the term, so score to likelihood ratio with no subsequent computations, as they do in the work of [38,39].

In [4], calibration on information extracted from speech is explored. It addresses the main issues in calibrating data: limited training data and dataset shift when score distributions change between calibration and test sets. Calibration in speaker recognition is based on features namely duration of audio, distance, language, and gender [4]. The work of [28] studies the impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings.

A problem that could arise when calibrating [46,47] is data scarcity. The references indicate that the use of simulated data gives a big improvement in data scarcity situations, but the testing of the validity of simulated databases for the operational use of systems in a real setup is still controversial. For face images this would imply that a solution for data scarcity could be Generative Adversarial Networks (GAN) that generate realistic fake face images [18]. However, forensic implications of simulating face data should be evaluated.

3. Materials and methods

We aim for the validation of automated facial image comparison systems computing an SLR. Referring again to Fig. 1, this validation has two parts. First is the SLR system itself, which consists of a scorer and a calibrator. In this case the scorer is the biometric system, i.e.,

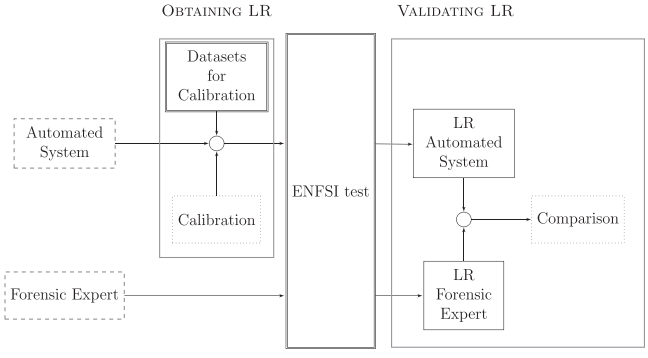

Fig. 1. Overview of the main topics addressed in this paper. Dashed boxes correspond to evaluating agents. Dotted boxes represent operations and double framed boxes correspond to data.

the facial image comparison automated system that will return either a distance or a similarity which will be treated as a score. The other element, the calibrator, will take a set of scores that either correspond to a group of facial images of comparisons within the same person (within source variability or WSV) or comparisons in a set of face images amongst different persons (between source variability or BSV). Having a set with different people, each of them with several images of themselves and using the two sets of comparisons defined, a likelihood ratio can be estimated. Once the SLR is obtained, it must be calibrated. A well calibrated LR will be accurate with its own predictions [36]. In the final step, LR estimation will be validated. This validation is done using three measures, namely Cost Log Likelihood Ratio (Cllr) [5,6], Minimum Cost Log Likelihood Ratio (Cllr min) and Empirical Cross-Entropy (ECE) [36,38]; and compared to experts that have estimated a likelihood ratio for a series of tests issued each year [11–13,41,25,27,17].

3.1. Materials

3.1.1. Calibration of datasets: LFW, SC face and forenface

The Likelihood Ratio is the ratio of two probabilities. As the probability functions of WSV and BSV are unknown, it is necessary to obtain them empirically. Using the scorer to generate multiple intermediate scores of both populations in which the ground truth is known, histograms can be computed. Subsequently, the histograms are approximated with probability functions through different methods, namely Isotonic Regression [22], Kernel Density Estimation [19] and Logistic Regression [20]. There has been some discussion on which type of datasets are optimal for calibration [28,39], where there are some recommendations such as defining the WSV set with pairs that are highly similar (small distance between their embeddings) or choose a WSV set with the same features as the comparison at hand. The discrimination is robust independently from the dataset the system was calibrated with, but calibration itself is highly dependent on the conditions [31]. In particular, in the work of [31] the effect of duration, distance, language, and gender in speaker comparison by using a variety of datasets makes a difference in the calibration results. Intuitively, the higher the number of comparisons and the more similar the dataset is to the tested data, the better the calibration will be. In our setting the datasets used, in which surveillance images predominate to be compliant with the forensic nature of the tests, are described in Table 1[21,15,50].

3.1.2. ENFSI proficiency test

The ENFSI proficiency tests are issued every year by the European network of Forensic Science Institutes (ENFSI) Digital Imaging Working Group (DIWG). They are tests in Facial Image Comparison (FIC) for quality assurance purposes, to examine how well the institutes perform their facial image comparisons. In Table 2, explanations and descriptions of the ENFSI tests are given.

3.2. Methods

3.2.1. Obtaining the SLR

Following similar procedures as in DNA and speaker comparison [4,23] and other face recognition works in forensics such as [3,40], the score obtained when comparing two faces is transformed to a Likelihood Ratio. Although the process of calibration has been studied and analysed in speaker comparison works such as [28] or [31], [3] and in facial image comparison in [40] those studies in facial image comparison did not take into account how different calibration characteristics such as features affect the results. It is for that reason that in this work we select different calibration types based on the work of speaker comparison and test them against ENFSI tests. The following section gives details on how this process is carried out.

Table 1
Dataset description.

Dataset	Number of images	Characteristics						
		Age	Gender	Ethnicity	Occlusions	Pose	Illumination	Resolution
Labeled Faces in the Wild (LFW)	13,233 images 3233 labeled 5749 subjects	Mostly adults Few children	Mostly men	Mostly Caucasian	Not many	Mild	Acceptable	Mostly high
SC Face	images 4160 images 130 subjects	From 20–75	115 males 15 females	All Caucasians	Beard Mustache Glasses	from −90 to +90 in equal steps of 22.5 degrees	Uncontrolled	Variable
ForenFace	Mostly videos 97 subjects	Contains both facial images and videos	Mixed	Mostly Caucasian	Caps Beards Glasses	Different	Acceptable	High, Low

Table 2
Summary of ENFSI tests and their characteristics.

Year	Number of images	Query		Reference
		Constant	Different	
2011	60	Illumination, Angle of view	Qualities, compression, distance to the camera.	Frontal acceptable quality
2012	60	Image quality, Distance to the camera	Angle of view	Acceptable quality
2013	80	Distance to the camera, angle of view	Distracting headwear (e.g. cap or a scarf)	Same as query image
2017	70	–	Pose, illumination, quality	Frontal acceptable quality
2018	40	Illumination	Cameras, pose, quality	Same as query image
2019	40	Various camera types, facial image comparison of children		Image quality was high, good enough to show moles and scars
2020	40	Imitates id document photo, full frontal		Some images were close relatives which includes twins/siblings.

3.2.2. The scorer

The scorer is the system or person whose goal is to provide an estimation of a Likelihood Ratio, possibly through the intermediate determination of a distance or similarity. This scorer can e.g., be a pre-trained neural network which is calibrated so the intermediate score can be transformed to a Likelihood Ratio or a forensic expert who directly provides an estimated likelihood ratio based on the visual comparison of the face features [14]. The scorers used in this work are as follows: .

The automated system scorer compares two facial images and returns either a distance or a similarity as intermediate score. The scores group in two sets. As mentioned in Section 3.2.1, the first set is for estimating WSV in which two images corresponding to the same person are compared and the second set in which the comparisons correspond to different persons for estimating the BSV. Our open-source scorer uses Deepface state-of-the-art face recognition built with Deep Learning [44]. According to [44], the supported models FaceNet-512 got 99.65%; ArcFace got 99.41%; Dlib got 99.38% and VGG-Face got 98.78%; accuracy scores on Labeled Faces in the Wild benchmark whereas human beings could have just 97.53%. The commercial automated system we use is FaceVACS version 5.5.2 [7] from Cognitec. This commercial system only provides the final similarity score between two facial images. Open software exposes the architecture and weights that output the representation of each of the facial images in the n-dimensional space, which gives flexibility for tasks such as clustering or comparison. Also, open software allows to change the method to compute the similarity score between facial images. While the similarity score of Cognitec is a number between 0 and 1, but not disclosed how it is exactly computed, open software has different distance functions such as euclidean distance or cosine similarity which can be computed and compared.

The forensic participants are members of the European Network of Forensic Science Institutes (ENFSI). Each year, a Proficiency Face Recognition test is distributed among laboratories within the organization and experts can assess which factors affect face recognition and their own assessments on Likelihood Ratio estimation. The manual forensic facial comparison process is a pair by pair comparison in which the experts estimate the likelihood ratio based on facial image features. The experts use a structured method to reach matching/non-matching conclusions for an image pair.

3.2.3. Calibration

As mentioned, calibration is the process of obtaining a Likelihood Ratio from a score. Likelihood Ratio is defined in Section 2.1.

Now, there are two questions that need to be addressed according to similar studies where Score-based Likelihood Ratio is used for comparison assessment. First, which images to use for calibration? The whole dataset or just a subset having the most relevant features? Second, how to model the WSV and BSV distributions given the available data [16,39,45]? Given that the performance of facial image comparison highly depends on the quality of the data that a model is built with, the author in [34] suggests to use images having similar conditions to the real life facial image comparisons. Regarding the BSV modeling, [3] uses what is known as "pseudo-traces", that is using several pictures of the reference individual in the comparison instead of generic pictures of the same person not related to the case at hand. In their results, on average 59,2% of the cases using this approach were more effective than the generic approach. In the case of BSV, no modeling other than generic between-source comparisons has been done [3,40]. However, taking this approach is paramount due to the importance of choosing the relevant population to obtain a suitable $P(E|H_d, I)$. According to [16], no study has yet shown the impact of variations in the choice of the relevant population for automatic face recognition. Moreover, in speaker comparison, in works such as [31], they calibrate according to divisions of the dataset with the same features, e.g., age or gender. It is for that reason, that in this paper, three types of BSV calibration were carried out attending the methods practiced in other forensic disciplines.

With naive calibration, SC Face and ForenFace datasets image pairs were used indistinctly. In this dataset, no filters according to scores or features (as done in [30,48]) were applied when choosing the pairs for both WSV and BSV distributions. This approach is considered the "generic" approach.

Quality score calibration is an attempt of detecting how rare or frequent it is to find a face similar to the suspect's face in the relevant population, also known as "typicality". The calibration is performed in the following way: first, each image of the SC Face and ForenFace Dataset is compared against 1000 randomly chosen images from Labeled Faces in the Wild. As all the identities in ForenFace and SC Face with respect to Labeled faces in the Wild correspond to a different person, all the scores obtained will belong to the BSV distribution. What we will call a "Quality Score" is the average of the ten highest score mismatches from both SCFace and ForenFace with respect to Labeled faces in the wild. The higher that score, said face (from either SCFace or ForenFace) is more easily confused against a "standard" dataset (LFW) than another image with a lower score. This "Quality score" will be used to create different sets of calibration BSV corresponding to the Quality Score of the compared test faces. In other words, later in the validation part of the pipeline, faces will be compared in pairs. Each image of these pairs will be contrasted against LFW and a quality Score will be assigned to said test pair. Then this pair will only be calibrated with images having the same "Quality Score". For example, a test pair with "Quality Score" of 7 and 8 respectively, will generate a BSV in which the comparison scores have been obtained with calibration pairs that are also a 7 and 8 in "Quality Scores".

For feature calibration, more intuitive than the former, all images in the test pairs were labeled according to if they contain headgear, beard, glasses, yaw, pitch, resolution or other occlusions. The databases SCFace and ForenFace have already this type of labeling so the BSV population was generated with only the images that presented the same features as the test images.

Regarding the WSV population, [3] uses images from the same subject as the test pair to generate the test WSV population. but in our work, the usecase is that only one image of the suspect is available, as the suspect is not yet convicted. This is the case presented in the ENFSI tests used to evaluate. Because of that, a generic approach was taken by generating the same WSV for each calibration using pairs from the databases LFW, SCFace and ForenFace with the same identity.

To obtain the Likelihood Ratio from a score, in this paper we follow three types of statistical methods to fit the WSV and BSV distributions. Three calibration methods were evaluated, Isotonic Regression, Kernel Density Estimation and Logistic Regression. They were chosen as two non-parametric (Isotonic regression and KDE), and one parametric (Logistic Regression) method. The Logistic Regression was chosen in the first place because it can assume the characteristics of many different types of distributions. It is flexible enough to model a variety of datasets. It can adapt to both skewed data and symmetric data. It is a parametric distribution, which assumes parameters (defining properties) of the population distribution from which the calibration data are drawn. Because of that, the second choice is a kernel density estimation (KDE), which is a non parametric test that does not make such assumptions. The third method chosen is Isotonic regression commonly used in machine learning models for statistical inference. The choice of one method or another doesn't seem to have a correlation with the performance of the different models of Likelihood Ratio estimation. The software used for calibration computations was from [32].

A. Isotonic regression is a free-form linear model that can be fit to sequences of observations [22] and then used for prediction. A common algorithm to obtain the isotonic regression is pool-adjacent-violators algorithm (PAVA). If we have the data $(x_1, y_1), ...,(x_n, y_n) \in \mathbb{R}$, isotonic regression looks for $\beta_1, ...,\beta_n \in \mathbb{R}$ such that the β_i approximate the y_i while being monotonically non-decreasing.

$$\text{minimize}_{\beta_1,...,\beta_n} \sum_{i=1}^{n} (y_i - \beta_i)^2 \tag{2}$$

For the Likelihood Ratio estimation, x_i represents the score and $y_i = 0$ if it is a mismatch or $y_i = 1$ if the pair comparison is a match. Applying the PAVA algorithm, proceeds as follows: going from low values of x_i to high values of x_i, we set $\beta_i = y_i$. If this causes a violation of monotonicity ($\beta_i = y_i < y_{i-1} = \beta_{i-1}$), replace both β_i and β_{i-1} with the mean $\frac{y_{i-1}+y_1}{2}$. This could result in earlier violations. If this happens, we average β_{i-1} and β_{i-2}.

B. Kernel Density Estimation is a non-parametric density estimator. It is an algorithm which seeks to model the probability distribution that generated a dataset [19]. To fit this distribution, it makes use of two parameters, which are the kernel, which specifies the shape of the distribution placed at each point, and the kernel bandwidth, which controls the size of the kernel at each point.

C. Logistic regression models the probability of a certain class or event existing [20]. Logistic Regression is used when the dependent variable (target) is categorical. The dependent variable is a binary variable that contains data coded as 1 (match) or 0 (mismatch). In other words, in this paper, the logistic regression model predicts the probability of match given a score $P(Y = 1)$ as a function of X.

3.3. Validating LR

The validation (see Section 2.3) for Likelihood Ratio assessments has been discussed in [5,28]. There three metrics are introduced that consider not only if the decision taken by the automated system was correct, but also penalizes if the system provides an inconclusive answer. The metrics are Cllr, Cllr Min and ECE plot [10,38]. Compared to equal error rate or ROC curves, these metrics provide a better representation of both the discrimination power of the model and its calibration performance. These metrics can be used to evaluate any set of Likelihood Ratio estimations, both for the automated systems and the forensic experts. In this paper we will use them to evaluate their performance on the ENFSI Proficiency tests.

3.3.1. Evaluation criteria

Forensic experts and automated system are compared with respect to their estimated Likelihood Ratios. As explained in 1, validation requires both assessment of discrimination and calibration. In [36], proposes both Log-Likelihood Ratio cost (Cllr) and Empirical Cross-Entropy (ECE) as adequate metrics for validating calibration on an incorrect selection of databases, a bad choice of statistical models, low quantity and bad quality of the evidence. There are several methods that evaluate the model performance on discrimination and calibration, such as EER, DET, Tipett plots (see Section 2). However, according to [16] and [36], the ones that condense this information better are Cllr, Cllr min and ECE plots, which are described in Section 2.1.

The Cost likelihood ratio is defined as:

$$C_{llr} = \frac{1}{2N_p} \sum_{i_p} \log_2(1 + \frac{1}{SLR_{i_p}}) + \frac{1}{2N_d} \sum_{j_d} \log_2\left(1 + SLR_{j_d}\right), \tag{3}$$

where the indices i_p and j_d respectively denote summing over the computed LR scores for each face pair comparison where each proposition (respectively prosecutor or defense) is true. Minimizing the value of Cllr implies an improvement of both discrimination and calibration performance of the automated system [39]. The value ranges from zero (perfect decision making), to infinity (completely wrong). A value of one indicates the system makes a random selection. A value larger than one indicates that the system is making a decision worse than random, i.e. supporting the prosecution hypothesis when it should be supporting the defense hypothesis or vice versa.

Empirical Cross-Entropy in terms of prior odds and the SLR is given by [39]:

$$ECE(O(H_p), SLR) = \frac{P(H_p \mid I)}{N_p} \sum_{i_p} \log_2\left(1 + \frac{1}{SLR_{i_p} \times O\left(H_p\right)}\right) + \frac{P(H_d \mid I)}{N_d} \sum_{i_d} \log_2(1 + SLR_{i_d} \times O(H_p)), \tag{4}$$

where s_{i_p} and s_{i_d} denote the scores from the same subject and different subject scores in each of the facial image comparisons, where H_p or H_d is respectively true. $O(H_p)$ is the value of the prior odds.

To be more precise, the meaning of the ECE plot is as follows [39]: .

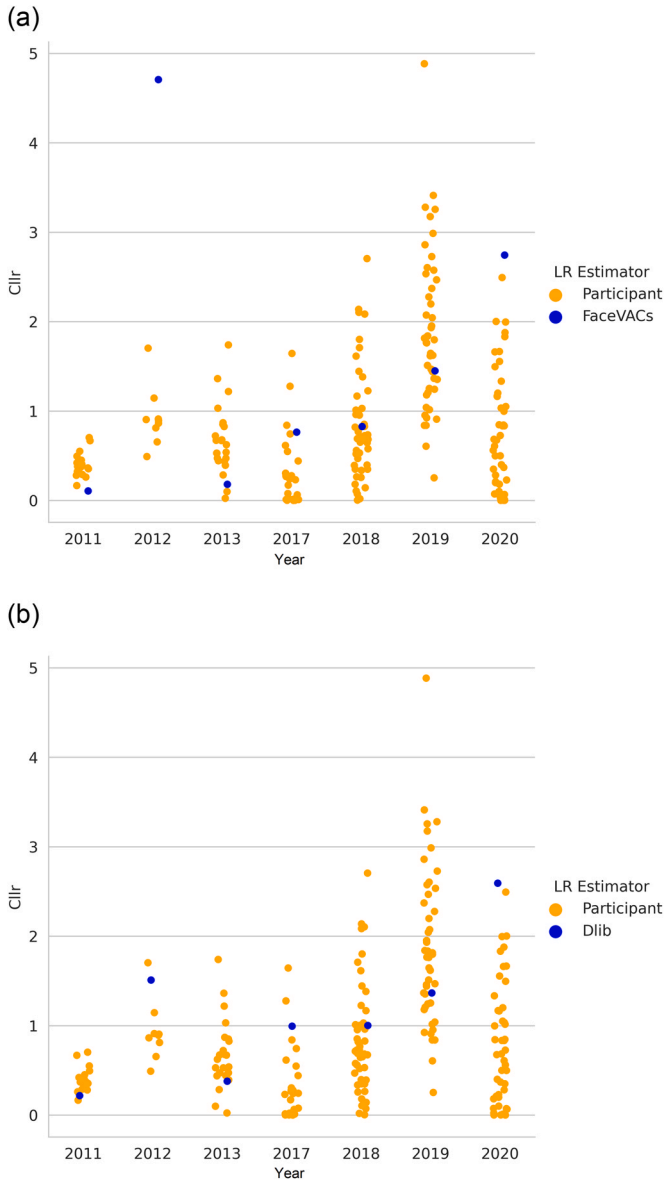

Fig. 2. Cllrs for naive calibration with Dlib and FaceVACs.

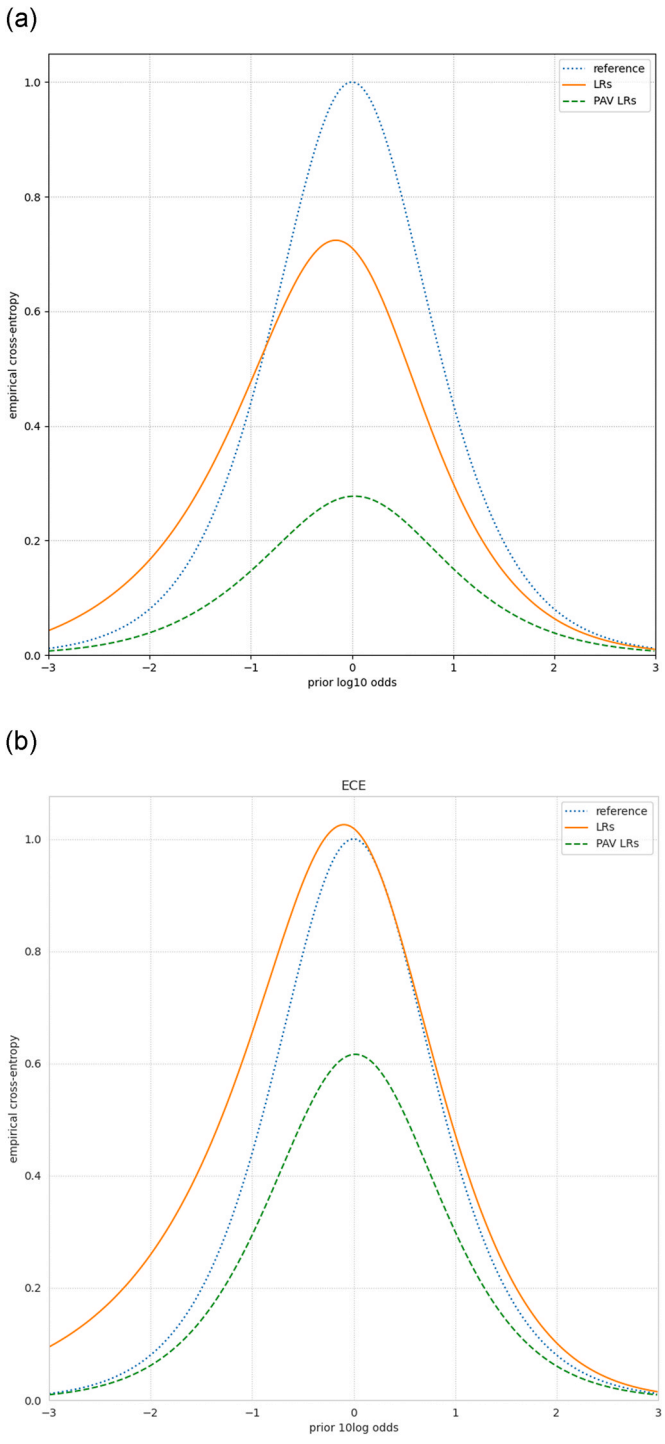

Fig. 3. ECE plots for naive calibration with Dlib and FaceVACs.

LRs: This curve is the ECE of the LR values in the validation set, as a function of the prior log-odds. The lower this curve, the more accurate the method. This curve shows the overall performance of the LR method.
PAV LRs: This curve is the ECE of the validation set of LR values after the application of the PAV algorithm. This shows the best possible ECE in terms of calibration, and it is a measure of discriminating power.
Reference: This curve represents the comparative performance of a so-called neutral LR method, defined as the one which always delivers $LR = 1$ for each forensic case in the set of LR values. This

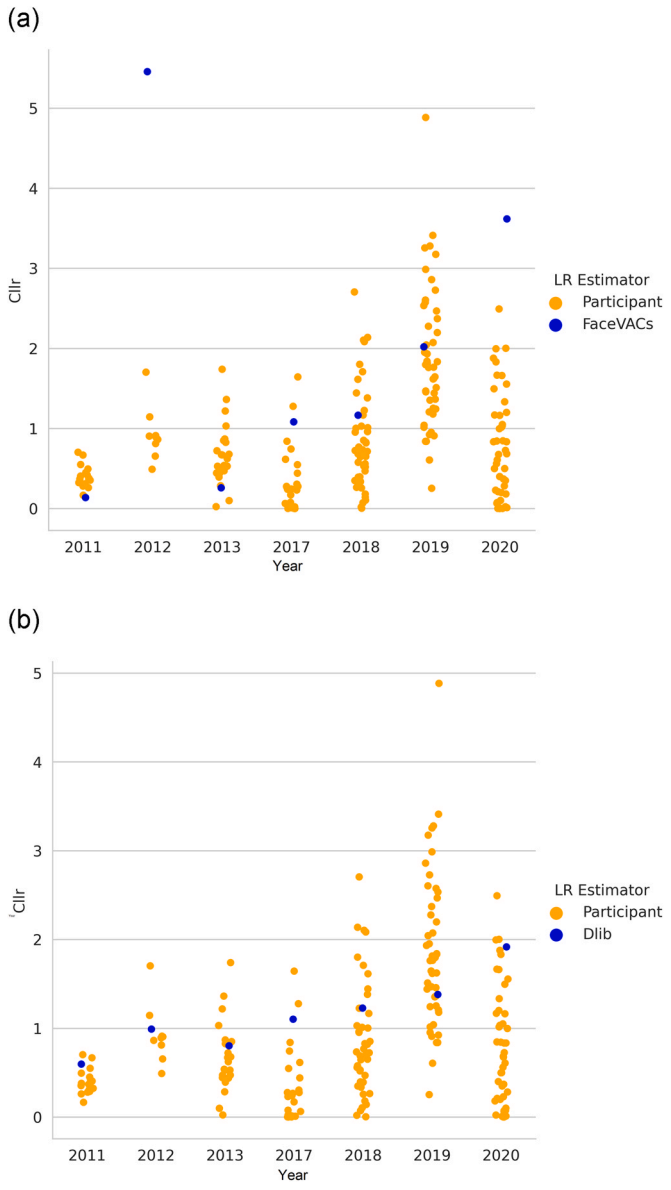

Fig. 4. Cllrs for same quality score calibration with Dlib and FaceVACs.

neutral method is taken as baseline performance: the accuracy should always be better than the neutral reference. Therefore, the solid curve in an ECE plot should be always be lower than the reference curve, for all represented values of the prior log-odds.

4. Results

We used the three following three types of calibration: naive, quality score and same features calibration (see Sections 4.1, 4.2 and 4.3).

4.1. Naive calibration

Calibration was performed with three datasets (LFW, SC face and ForenFace) with no filters related to the testing ENFSI tests. From the three datasets chosen, 10,000 random pairs were selected as a representative sample. In Figs. 2a and 2b the Cllr from both face recognition and FaceVACs can be seen. In Figs. 3a and 3b ECE plots for the naive calibration can be seen.

We can appreciate that the year 2020 has a very poor Cllr, which approximates 4. This could be due to that year having identical twins in the ENFSI tests, which confused the algorithm and led it to classify

as matches what should have been mismatches. For the year 2019, the Cllr is quite high, which indicates a poor performance, but it is in the same interval as the forensic experts. This year the comparison of faces was among children so both the algorithm and the experts had difficulties with the images. For the years 2017 and 2018 the Cllr is approximately 1, which indicates the power decision of a random algorithm. On the other hand, human participants managed to have their cllr below one in year 2011 (except 2 participants) and about two thirds of them had a cllr below 1 in year 2018. For the rest of the years 2011, 2012 and 2013, both the commercial software FaceVACs and the open software Face recognition present results comparable to the best performing experts. Regarding the ECE plots, both FaceVACs and Face Recognition seem to make less errors in the prosecution priors than in the defense priors.

4.2. Quality score calibration

For each WSV pair of the calibration datasets (LFW, Sc and ForenFace) the corresponding BSV (i.e. pairs that correspond to a mismatch) is chosen according to a 'quality-score'. Through experiments, it can be seen that in higher resolutions there is a clearer threshold in which the system distinguishes which comparisons are a match and which ones are a mismatch. When the size of the image (measured in megapixels) is above 0.3, the similarity of matched pairs is close to 1, and close to 0 in the case of mismatches. As resolution of the images decreases, similarity for matched images also decreases and similarity for mismatches rises for some cases.

In Figs. 4a, 4b, 5a, and 5b, the validation of the automated systems against experts is checked. The results are shown for the years 2011, 2012, 2013 and 2017, 2018, 2019 and 2020 and both Cllr and Cllr min are plotted.

4.3. Feature calibration

The feature calibration was performed with pairs of the two datasets (SC face and ForenFace). For each test pair (from ENFSI tests), the set of features of image one and the set of features of image two are considered to calibrate only with the pairs of the calibration dataset that have the same set of features as these two images. The features to be considered were: glasses, beard, headgear, other occlusions, and low quality. The datasets were manually annotated.

In Figs. 6a, 6b, 6c, and 6d, it can be seen that Cllr calibrating the system with comparisons that have the same features has improved results with respect to Cllr calibrated with comparisons of the same quality score.

In Figs. 7a and 7b ECE plots for both automated systems are plotted.

4.4. Overview

An overview of the results can be seen in Table 3. For this results, the open source Dlib, is compared to the commercial software FaceVACS and to the ENFSI participants. The different results can be seen where the filters applied improve with respect to naive calibration. FaceVACS performs better than the open software system. The calibrator chosen for the results in the table was Isotonic Calibrator, although calibrating with any of the three would turn out to be similar to Cllr, the Isotonic seemed to outperform a bit with respect to Logistic Regression and Calibration. However, further work is necessary to make any recommendations on which cases each of the three calibration methods should be used.

5. Discussion

As we can see in the Cllr and ECE plots, the commercial software FaceVACs outperforms both the open software face recognition and the experts for full frontal images. However, the quality of images presented in the tests are easier for the automated system than the material that is normally handled in cases. Most of the images (especially in the year 2017) are frontal with little pose variation, which facilitates the task for the automated system. Most of the wrong assessments provided by the automated system were due to occlusions in the test images (caps, mics, scarfs) or to illumination. On the year 2011 dataset, where the illumination was constant but the images had different resolution and compression, there was not significant improvement neither in FaceVACs nor face recognition with respect to naive calibration or quality score and filtered base

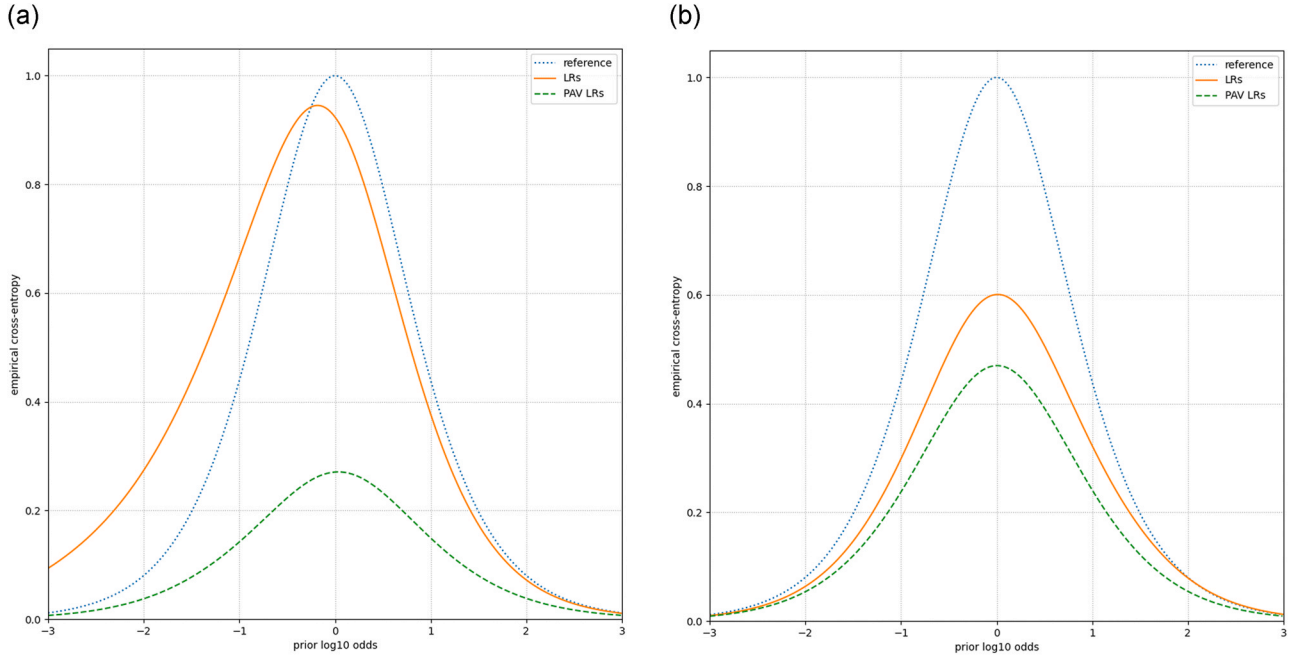

Fig. 5. ECE plot for quality score calibration with Dlib and FaceVACs.

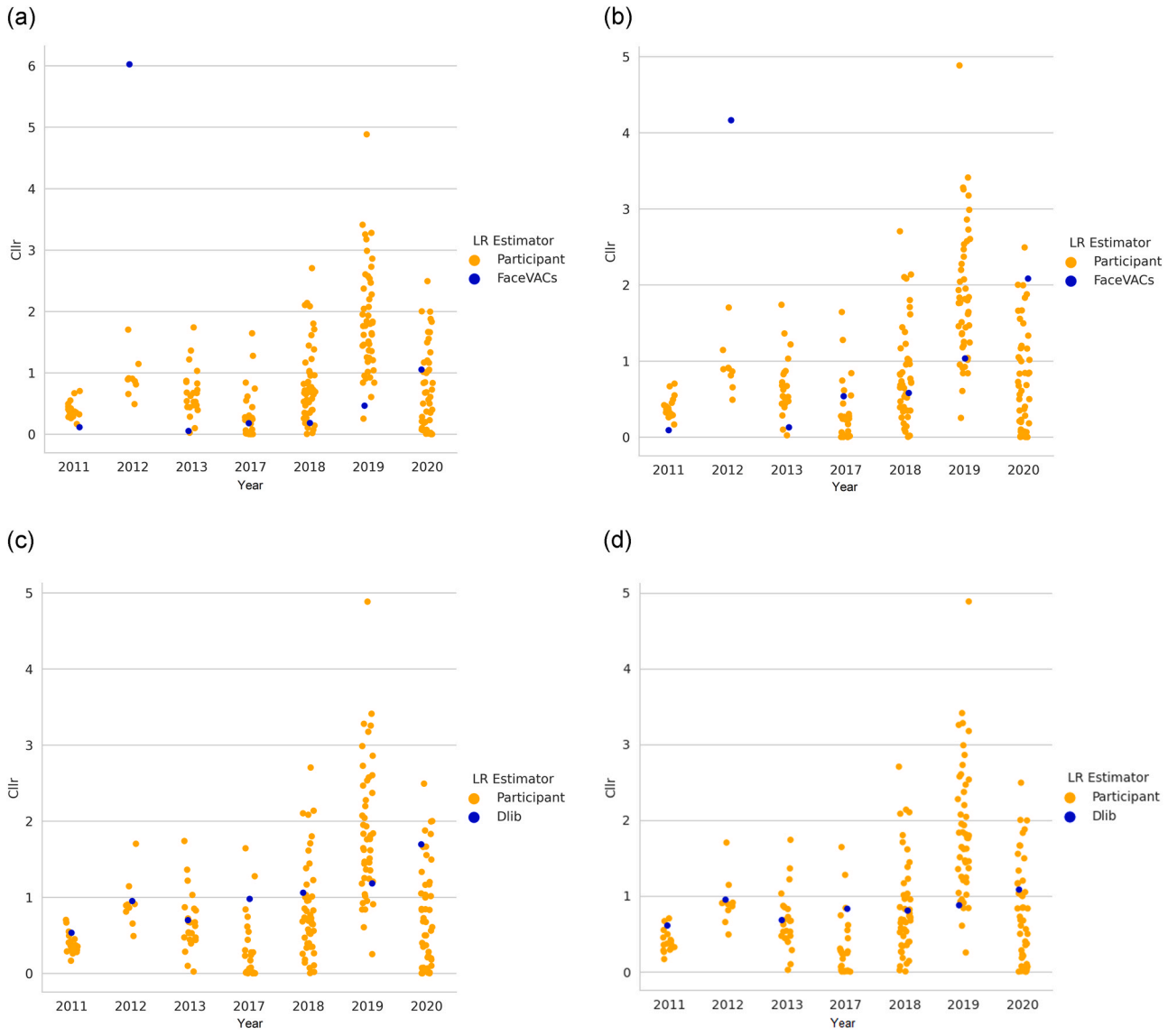

Fig. 6. Cllrs for features calibration with Dlib and FaceVACs.

calibration, as the Cllr of the automated systems was already close to zero in the naive calibration. On the other hand, there is a significant improvement in the years 2018 and 2019 using the same features calibration instead of the naive calibration. These years had as peculiarity that year 2018 had a lot of variety in the test images (age variation, pose, quality.) and 2019 had pictures of children. The year 2020 has low performance in all the cases due to photos of twin siblings being present among the test images. The automated system had difficulty differentiating these faces and gives them a high similarity score, making the calibration prone to error.

This study takes a step further the usability of automated facial image comparison systems in the forensic field. In the literature, such calibrations are performed normally as suspect-anchored and trace-anchored [16,39], however this type of calibration was not the use case in this study due to only having one sample of each identity in the comparison tests. This use case is given when the suspect is not yet convicted and only one image of the individual is available.

This has not impeded the automated system of reaching in most of the years the accuracy of the forensic experts. It may lead us to think that if on top of performing these calibrations with publicly available data-sets, data more relevant to the case was added (such as more images of the suspect, images of the suspect and other relevant population resembling the conditions in which the query image was taken) the results would only improve.

As future work, it would be convenient to indicate that the automated system performance (both FaceVACS and Deepface) is less reliable if there are occlusions. When the face was not detected by the automated system, it was not considered for the Cllr or ECE plot. A possible alternative to this is to add the lack of face detection as an inconclusive LR (i.e., LR = 1) which would drop the performance of the system in Cllr terms, as humans are habitually more efficient when finding faces in a picture than a automated system can be.

As indicated, an important point made by [10] is that validation of Likelihood Ratio in the forensic field should take into account not only accuracy (if it is right or wrong assessment of match-mismatch) but also its calibration, i.e., the system capacity to make strong assessments. If a system provides an LR of around 1 for a comparison corresponding to a match, the assessment (i.e., discrimination power) is right, but the calibration and functionality to help to take a decision is not very useful. On the other hand, a second system that for the same match provides an LR of 1000 is both providing a high discriminating power and good calibration. The article [10] warns that validation of LR systems should check on both characteristics. For our work, measuring with Cllr and ECE plot has this warning

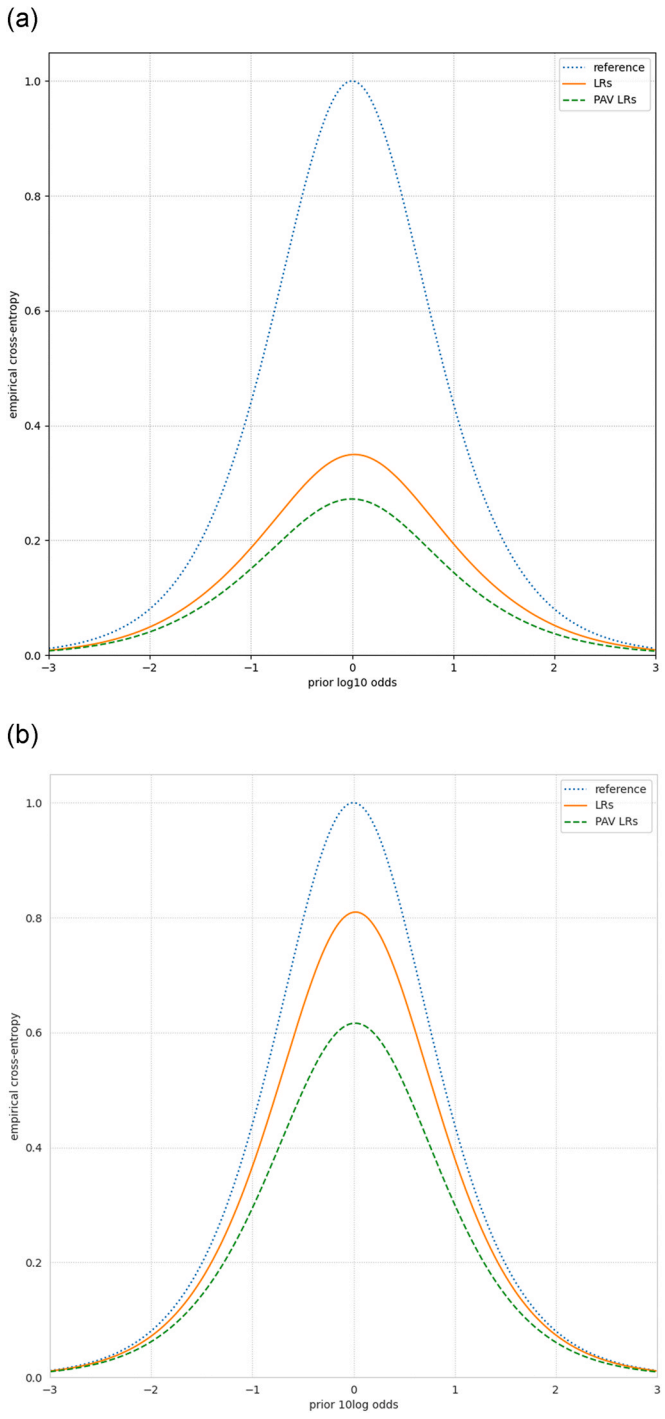

Fig. 7. ECE plot for features calibration with Dlib and FaceVACs.

covered, because looking at both Eqs. 3 and 4, the cost will increase for those systems that provide wrong assessments or low discrimination power (LR close to one).

Regarding the three calibrator methods chosen (Logistic Regression, KDE and Isotonic Regression), none of them seemed to stand out from the others. Although Isotonic Regression seemed to achieve slightly better results than the other two, future work is required to assess in which use cases one calibration is better than the other. With respect to the three calibration methods chosen, although both the confusion score and labeled filters improved the Cllr with respect to applying generic calibration, also further research is needed to help the investigator to determine which method would suit best for each use case.

Table 3
Cllr results summary for Dlib, FaceVACs (according to filters chosen) and participants.

Year	Filters	Dlib	FaceVACs	Average participants
2011	No filters	0.22	0.11	0.40
	Confusion Score	0.60	0.14	
	Yaw, Pitch	0.53	0.11	
	Glasses, Beard	0.60	0.08	
	Low Quality	0.63	0.09	
	Head Gear	0.56	0.20	
2012	No filters	1.51	4.70	0.93
	Confusion Score	0.99	5.46	
	Yaw, Pitch	0.95	6.02	
	Glasses, Beard	0.94	3.67	
	Low Quality	0.95	4.16	
	Head Gear	0.96	5.10	
2013	No filters	0.38	0.18	0.67
	Confusion Score	0.80	0.26	
	Yaw, Pitch	0.70	0.05	
	Glasses, Beard	0.64	0.12	
	Low Quality	0.70	0.13	
	Head Gear	0.73	0.35	
2017	No filters	0.99	0.76	0.35
	Confusion Score	1.10	0.87	
	Yaw, Pitch	0.98	0.18	
	Glasses, Beard	0.81	0.53	
	Low Quality	0.83	0.54	
	Head Gear	1.00	1.24	
2018	No filters	1.00	0.83	0.84
	Confusion Score	1.23	0.93	
	Yaw, Pitch	1.06	0.18	
	Glasses, Beard	0.77	0.56	
	Low Quality	0.80	0.58	
	Head Gear	1.09	1.30	
2019	No filters	1.36	1.45	1.88
	Confusion Score	1.38	1.57	
	Yaw, Pitch	1.18	0.46	
	Glasses, Beard	0.82	1.07	
	Low Quality	0.87	1.03	
	Head Gear	1.23	2.21	
2020	No filters	2.59	2.74	0.78
	Confusion Score	1.92	2.78	
	Yaw, Pitch	1.70	1.05	
	Glasses, Beard	1.00	2.01	
	Low Quality	1.05	2.08	
	Head Gear	1.71	3.74	

6. Conclusion

In conclusion, with this study it has been demonstrated that applying "filters" such as"Quality Score" and calibration with the same features as the test images improves the performance in the calibration, in terms of both Cllr and ECE. The results with open software are inferior, but they are more transparent so more research should be conducted to bring open software at par with commercial vendors. On top of performing these calibrations with publicly available data-sets, more relevant data to the case, such as more images of the suspect, images of the suspect and other relevant population resembling the conditions in which the query image was taken could be added. The results would only improve. The expert cannot be replaced by this tool, but becomes more efficient because the computer can help to reduce the amount of information to be managed by doing appropriate filtering. If facial image comparison is conducted by two experts doing the comparison independent from each other, the third might be an algorithm, and the experts can evaluate their findings as well as the findings of the algorithm to draw a conclusion.

CRediT authorship contribution statement

Andrea Macarulla Rodriguez: Conceptualization, Methodology, Software, Investigation, Resources, Data curation, Writing – original draft, **Zeno Geradts**: Conceptualization, Software, Validation, Writing – review & editing, Supervision, **Marcel Worring**: Conceptualization, Validation, Writing – review & editing, Supervision.

Acknowledgements

The authors would like to express their gratitude to Rolf, Jeroen, Judith, Nivea, Simone, Jeannette, and Elina for their invaluable assistance throughout the first phases of this paper.

References

[1] Aitken, C., Berger, C.E., Buckleton, J.S., Champod, C., Curran, J., Dawid, A., Evett, I. W., Gill, P., Gonzalez-Rodriguez, J., Jackson, G. et al., 2011. Expressing evaluative opinions: A position statement. Science & Justice 51, 1–2. ⟨https://pureportal.strath.ac.uk/en/publications/expressing-evaluative-opinions-a-position-statement.

[2] C. Aitken, F. Taroni, The evaluation of evidence, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, 2005, pp. 69–118, https://doi.org/10.1002/0470011238.ch3

[3] T. Ali, Biometric Score Calibration for Forensic Face Recognition. (Ph.D. thesis), University of Twente, 2014, https://doi.org/10.3990/1.9789036536899

[4] N. Brummer, Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech. (Ph.D. thesis), University of Stellenbosch, Stellenbosch, 2010, https://doi.org/10.1016/j.csl.635 2005.08.001

[5] N. Brümmer, A. Swart, Bayesian calibration for forensic evidence reporting, stat 1050 (2014) 25.

[6] Brümmer, N., du Preez, J., 2006. Application-independent evaluation of speaker detection. Computer Speech & Language 20, 230–275. 10.1016/j.csl.2005.08.001.

[7] Cognitec, 2021. Facevacs. ⟨https://www.cognitec.com/facevacs-technology.html⟩.

[8] Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685–4694. 10.1109/CVPR.2019.00482.

[9] D. Dessimoz, C. Champod, A dedicated framework for weak biometrics in forensic science for investigation and intelligence purposes: the case of facial information, Secur. J. 29 (2016) 603–617, https://doi.org/10.1057/sj.2015.32

[10] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci. Int.: Synerg. 2 (2020), https://doi.org/10.1016/j.fsisyn.2020.08.006

[11] Norell, K., Läthén, K.B., Eklöf, F., Bergström, P., 2011. FIC test 2011 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2011. (ENFSI Internal report 2012:03).

[12] Eklöf, F., Läthén, K.B., Bergström, P., Norell, K., Leitet, E., 2012. FIC test 2012 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2012. (ENFSI Internal report 2013:05).

[13] Eklöf, F., Bergström, P., 2013. FIC test 2013 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2013. (ENFSI Internal report 2014:08).

[14] ENFSI, 2018. Best Practice Manual for Facial Image Comparison. European Network of Forensic Science Institutes (ENFSI).

[15] M. Grgic, K. Delac, S. Grgic, Scface–surveillance cameras face database, Multimed. Tools Appl. 51 (2011) 863–879.

[16] M. Jacquet, C. Champod, Automated face recognition in forensic science: Review and perspectives, Forensic Sci. Int. 307 (2020) 110–124.

[17] S. Jilani, S.C. Martínez, A. Ruifrok, F. Eklöf, L. Neale, FIC test 2020 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons, ENFSI Intern report (2020).

[18] Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of gans for improved quality, stability, and variation. In: Proceedings of the International Conference on Learning Representations.

[19] J. Kim, C.D. Scott, Robust kernel density estimation, J. Mach. Learn. Res. 13 (2012) 2529–2565.

[20] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic Regression, Springer, 2002.

[21] Gary B. Huang, Erik Learned-Miller, Labeled Faces in the Wild: Updates and New Reporting Procedures. Technical Report, University of Massachusetts, Amherst, 2014.

[22] J. de Leeuw, K. Hornik, P. Mair, Isotone optimization inR: pool-adjacent-violators algorithm (PAVA) and active set methods, J. Stat. Softw. 32 (2009), https://doi.org/10.18637/jss.v032.i05

[23] van Leeuwen, D., Brümmer, N., 2013. The distribution of calibrated likelihood-ratios in speaker recognition. In: Proceedings of the Biometric Technologies in Forensic Science, pp. 24–29.

[24] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017. Sphereface: deep hypersphere embedding for face recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6738–6746.

[25] Martínez, S.C., Eklöf, F., Ruifrok, A. Moreton, A., 2018. FIC test 2018 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2018. ENFSI Internal report.

[26] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, https://doi.org/10.1016/j.forsciint.2016.03.048

[27] Michalski, D., Snyder, G., Martínez, S.C., Ruifrok, A., Eklöf, F. Neale, L. 2019. FIC test 2019 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2019. ENFSI Internal report

[28] G.S. Morrison, The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings, Forensic Sci. Int. 283 (2017) e1–e7, https://doi.org/10.1016/j.forsciint.2017.12.024

[29] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality, Sci. Justice 58 (2018) 47–58, https://doi.org/10.1016/j.scijus.2017.06.005

[30] Morrison, G.S., Ochoa, F., Thiruvaran, T., 2012. Database selection for forensic voice comparison, In: Proceedings of the Odyssey 2012-The Speaker and Language Recognition Workshop.

[31] Nandwana, M.K., Ferrer, L., McLaren, M., Castan, D., Lawson, A., 2019. Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems. In: Proceedings of the Interspeech 2019, pp. 4325–4329. 10.21437/Interspeech.2019–1808.

[32] Netherlands Forensic Institute. (2021). LIR Python Likelihood Ratio Library. Github. https://github.com/NetherlandsForensicInstitute/lir.

[33] R.A. Nichols, Interpreting dna evidence: statistical genetics for forensic scientists, Heredity 82 (1999) 585–586.

[34] Peng, Y., 2019. Face recognition at a distance: low-resolution and alignment problems. Ph.D. thesis. UT. Netherlands. 10.3990/1.9789036547116.

[35] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (2018) 208.

[36] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (2013) 156–169, https://doi.org/10.1016/j.forsciint.2013.04.014

[37] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci. 58 (2013) 1503–1518.

[38] D. Ramos, R. Haraksim, D. Meuwly, Likelihood ratio data to report the validation of a forensic fingerprint evaluation method, Data Brief 10 (2017) 75–92.

[39] D. Ramos, R.P. Krish, J. Fierrez, D. Meuwly, From biometric scores to forensic likelihood ratios, Handbook of Biometrics for Forensic Science, Springer, 2017, pp. 305–327.

[40] A.M. Rodriguez, Z. Geradts, M. Worring, Likelihood ratios for deep neural networks in face comparison, J. Forensic Sci. 65 (2020) 1169–1183, https://doi.org/10.1111/1556-4029.14324

[41] Ruifrok, A., Eklöf, F., Moreton, R., 2017. FIC test 2017 -Planning and implementation of the ENFSI-DIWG proficiency test concerning facial image comparisons 2017. ENFSI Internal report

[42] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (2005) 892–895.

[43] Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. 10.1109/CVPR.2015.7298682.

[44] Serengil, S.I., Ozpinar, A., 2020. Lightface: A hybrid deep face recognition framework. In: Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE. pp. 23–27. 10.1109/ASYU50717.2020.9259802.

[45] M. Tistarelli, C. Champod (Eds.), Handbook of Biometrics for Forensic Science, Springer International Publishing, 2017 https://doi.org/10.1007/978-3-319-50673-9

[46] Villalba, J., Brümmer, N., 2011. Towards fully bayesian speaker recognition: integrating out the between-speaker covariance. In: Proceedings of the Interspeech 2011, ISCA. 10.21437/interspeech.2011–142.

[47] G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, John Wiley & Sons, 2013.

[48] G. Zadora, D. Ramos, Evaluation of glass samples for forensic purposes-an application of likelihood ratios and an information–theoretical approach, Chemom. Intell. Lab. Syst. 102 (2010) 63–83.

[49] C.G. Zeinstra, D. Meuwly, A.C. Ruifrok, R.N. Veldhuis, L. Spreeuwers, Forensic face recognition as a means to determine strength of evidence: a survey, Forensic Sci. Rev. 30 (2018) 21–32.

[50] C.G. Zeinstra, R.N. Veldhuis, L.J. Spreeuwers, A.C. Ruifrok, D. Meuwly, Forenface: a unique annotated forensic facial image dataset and toolset, IET Biom. 6 (2017) 487–494.