Likelihood Ratios for Deep Neural Networks in Face Comparison

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

DOI
10.1111/1556-4029.14324
Publication date
2020
Document Version
Final published version
Published in
Journal of Forensic Sciences

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2020, Vol. 6
doi: 10.1111/1556-4029.14324
Available online at: onlinelibrary.wiley.com

PAPER

DIGITAL & MULTIMEDIA SCIENCES

Andrea Macarulla Rodriguez,[1] *M.Sc.; Zeno Geradts*,[1] *Ph.D.; and Marcel Worring*,[2] *Ph.D.*

Likelihood Ratios for Deep Neural Networks in Face Comparison*

ABSTRACT: In this study, we aim to compare the performance of systems and forensic facial comparison experts in terms of likelihood ratio computation to assess the potential of the machine to support the human expert in the courtroom. In forensics, transparency in the methods is essential. Consequently, state-of-the-art free software was preferred over commercial software. Three different open-source automated systems chosen for their availability and clarity were as follows: OpenFace, SeetaFace, and FaceNet; all three based on convolutional neural networks that return a distance (OpenFace, FaceNet) or similarity (SeetaFace). The returned distance or similarity is converted to a likelihood ratio using three different distribution fits: parametric fit Weibull distribution, nonparametric fit kernel density estimation, and isotonic regression with pool adjacent violators algorithm. The results show that with low-quality frontal images, automated systems have better performance to detect nonmatches than investigators: 100% of precision and specificity in confusion matrix against 89% and 86% obtained by investigators, but with good quality images forensic experts have better results. The rank correlation between investigators and software is around 80%. We conclude that the software can assist in reporting officers as it can do faster and more reliable comparisons with full-frontal images, which can help the forensic expert in casework.

KEYWORDS: digital forensic science, face recognition, face verification likelihood ratio, deep learning, ENFSI proficiency test

Face recognition is a powerful biometric technique to recognize a person due to its nonintrusive characteristics (1). Unlike other biometric recognition, such as for fingerprints or DNA, face recognition does not require cooperation from the suspect, making it a useful source of evidence. Digital facial evidence can appear in the form of CCTV footage, mugshots, mobile devices, or images from social media sites (2,3), which are now commonly used in court (4). An example use is a comparison between the ID image of a suspect and a face image retrieved from CCTV footage. This 1:1 comparison is known as verification or authentication. Organizations such as the European Network of Forensic Science Institutes (ENFSI) stimulate reporting the assertiveness of the statement match/nonmatch, that is, the verification stating whether it is the same person/different person or not, via a quantifiable amount (5). To that end, ENFSI enforces the use of a likelihood ratio (LR) as the mensurable method to express the confidence in the match/nonmatch decision (5,6) as also used in DNA or fingerprint comparison (7,8).

LR is based on Bayes' rule. It is defined as the ratio of the probabilities of two hypotheses: the null hypothesis, here the hypothesis of the prosecution (H_p), and the alternative hypothesis of the defense (H_d) (6). These terms are considered before certain findings, that is, the evidence E, are taken into account. Evidence in the case of face verification would come in the form of assessment if the face verification would be a match or a nonmatch. For face verification, we consider the null hypothesis to be a match and the alternative hypothesis to be a nonmatch. The LR is defined as follows:

$$\mathrm{LR}(H_p, H_d, E) = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

Would it be possible to obtain a valid LR in 1:1 face comparison suitable for forensics? For that end, we make use of the proceedings to attain an LR based on a biometric score (9,10). For face comparison, the biometric score is the value obtained from an automated system that can compute either the distance or dissimilarity between two given faces. Automated face recognition started with the eigenfaces in 1991 by M. Turk and A. Pentland (11). Since then, automated face recognition has been an active subject of research in the computer vision community. In recent years, AI and Deep Learning has allowed progress and improvement in automated face recognition systems by leaps and bounds. In 2014, DeepFace (12) reached 97.35% accuracy identifying faces in the benchmark dataset Labeled Faces in the Wild (LFW) (13) versus a human performance of 97.53%. The current state-of-the-art has boosted performance up to 99.80% (14). As a consequence of the established improvements in performance, automated systems have the potential to become assistants of judgment in court (15,16). To assess this potential, the LR obtained through the process must be validated for suitability in the forensic field (17,18).

The main contributions of this paper consist of carrying the process of a 1:1 verification end to end from an automated

[1]Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, Den Haag, The Netherlands.
[2]University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands.
Corresponding author: Andrea Macarulla Rodriguez, M.Sc.
E-mail: a.macarulla.rodriguez@nfi.nl
*Funding provided by ASGARD Project H2020-EU.3.7, Grant Agreement ID: 700381.

Received 16 Oct. 2019; and in revised form 04 Dec. 2019, 30 Jan. 2020, 27 Feb. 2020; accepted 2 Mar. 2020.

system to the final step of validation in the forensic field. We use three different open-source automated systems: OpenFace (19), SeetaFace (20), and FaceNet (21). The reason to use these three automated systems is due to their availability and transparency to the user. We obtain either a distance (OpenFace, Seeta) or a similarity (SeetaFace) that is treated as a biometric score. We transform the score through three statistical methods: Weibull distribution (22), kernel density estimation (KDE) (23), and pool adjacent violators algorithm (PAVA) (24). These methods use a set of scores to generate a probabilistic density function (Weibull, KDE) or a cumulative density function (PAVA). This process of obtaining such functions is commonly known as calibration. The set of scores is obtained from 1:1 comparison in the benchmark LFW, which is publicly available and contains a large set of unconstrained face images. After applying these steps, the LR is obtained. Once the LR is obtained, validation is performed through a comparison to the human expert. This conforms to our second contribution. The comparison with the human experts is based on the yearly ENFSI face recognition proficiency tests. These tests are performed by forensic experts giving a likelihood ratio to each pair of images analyzed, which may be of the same person or not. We will use these tests for both evaluating the performance of the automated system (match/nonmatch success through the Matthews correlation coefficient (25)) and the level of similarity to the forensic expert using rank correlation. The last contribution comes in the conclusion in the form of indications of how the automated tools can be of assistance to the expert based on the results found.

The paper is organized as follows: First, we review the related work, subdivided on the use of likelihood ratio in forensics in general, automated face recognition advances, and likelihood ratio tied to face recognition. Second, we disentangle step by step the procedure of assessing the likelihood ratio from an automated system score in Methodology. We explain each of the open-source tools, the methods, and the dataset used. In the Results section, we present the accuracy of the automated system reached with the different statistic methods, that is, when it got better or worse combinations of match/nonmatch predictions and the rank correlation with the human investigators. Finally, in the Discussion and Conclusion section, the results are analyzed and the potential of the automated system to assess forensic decisions is evaluated.

Related Work

Likelihood Ratio in Forensics

The idea of presenting evidence evaluation in court using a Bayesian probabilistic framework has been encouraged by institutions such as ENFSI in recent years as a suitable way to report evidence to justice (5,26,27) as it helps to standardize reasoning. In Europe, there have been initiatives to endorse this approach, for example, by the presentation of a guideline (18). As a result, forensic laboratories around the world use the likelihood ratio as a means to summarize their findings (6).

The use of likelihood ratio to report results has been explored in several fields of forensic research. DNA trace comparison is probably the area with the largest known use of LR in Europe and has already frequently been used in court (7,28). There has been a study in forensic speaker recognition by Ref. (29) that evaluates the performance of different methods used for forensic automatic speaker recognition. In the reference, three methods of speaker recognition (VQ, GMM, and i-vectors) are evaluated in accordance with the methodical guidelines for best practice in forensic semi-automatic and automatic speaker recognition. They conclude that in the experimental conditions of the paper, the three methods compared produce similar results. In forensic fingerprint comparison, the performance of LR for comparisons of fingerprints with fingermarks is studied in Ref. (30). They conclude that the results obtained could be used as a reference for score-based LR systems in other fields. In addition to applications in biometrics, LR computations have also been done for drug comparison (31), glass analysis (32), and gasoline analysis (33).

General guidelines for validation of the likelihood ratio approach can be found in Refs (17) and (18). The proposed process of validation takes into account two ways of obtaining likelihood ratios from a biometric comparison: score-based and feature-based. In our paper, we follow the majority of the work done in the biometric forensic field (3,29) where validation is based on scores.

Automated Face Recognition

Many methods for automated face recognition are available coming both from industry and academics (12,19,34,35). A survey carried out in Ref. (14) compares the current open-source best-performing face recognition algorithms and their accuracies in benchmarks (13). The work concludes that, since 2014, all the best-performing algorithms are based on convolutional neural networks (21,36). This state-of-the-art software outperforms human recognition in the benchmark dataset Labeled Faces in the Wild (13).

Face recognition algorithms in general consist of three steps: face detection, face normalization, and face identification or verification. Face detection aims to identify the presence of people's faces within an image (35). It is very well developed and also commonly used, for instance in autofocus in cameras. In the next step, face normalization, faces are aligned by matching landmarks. Each picture is warped so that the eyes and lips are always in the same place in the image. This will make the comparison a lot easier (37). Finally, identification tries to establish the identity of a person in an image by comparing it to a reference database, whereas in face verification, the model has to determine whether two images of a person belong to the same individual (38).

Face Recognition and Likelihood Ratio

As indicated, face recognition has been widely researched both in academia and industry, yet there has not been much research in the field of forensic face recognition (3).

There have been some attempts to compare automated systems to human performance. For instance, (16) researched groups of forensic experts (super-recognizers, i.e., people with significantly better-than-average face recognition ability, and trained facial reviewers) and untrained recognizers. In their study, they acknowledge that the best algorithms perform in the range of the best humans, that is, professional facial examiners.

The Carabinieri Forensic Investigation Department (15) in Italy carried out successful experiments on comparing commercial system performance in both the ENFSI test and 130 cases focusing on the accuracy in recognition. The results show that two of the three automatic systems performed superior compared with the mean of the forensic experts. As a next step, the authors recommend computing likelihood ratios as recommended by the

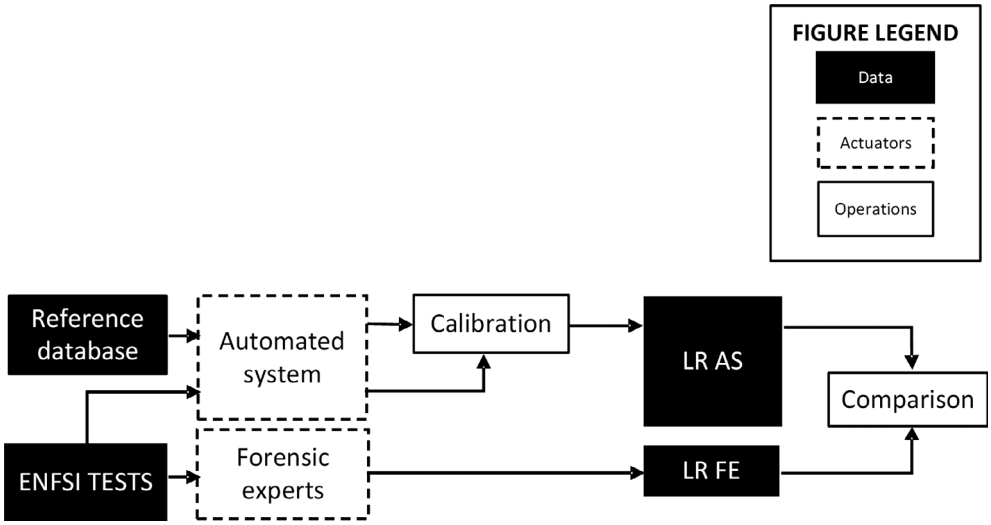

FIG. 1—*Overview of the paper. Black boxes symbolize data. LFW and ENFSI tests are image datasets, and LR AS and LR FE are the two sets of likelihood ratios obtained from the ENFSI tests from the automated system and the forensic experts, respectively. Dash line indicates actuators, such as automated system and forensic experts that receive input (both of them a pair of images to compare) and expel an output, scores, or distances in the case of automated system, and likelihood ratios in the case of forensic experts. A white box with solid black contour signifies an operation. For the automated system to output likelihood ratios, it needs to be calibrated through a reference database (in this case Labeled Faces in the Wild and the proficiency tests). The final goal and main contribution of this paper is the comparison between the LR obtained by the automated system and the forensic experts, both in accuracy and similitude.*

ENFSI guideline for evaluative reporting in forensic science. In their paper, they state a strongly optimistic view for the future use for support vector machines and convolutional neural networks.

Methodology

The objective of this work was to compare the operation of automated facial recognition systems with the way forensic experts assess their findings, and to determine whether automated systems can be helpful tools to the investigator.

The automated systems, when comparing two images, return a score or a score plus an empirically calculated threshold. The score does not directly give information whether it is the same person or not, rather the score of the system indicates the confidence of the system in the similarity of the two images under consideration. Therefore, the software output must be converted to LR values that facilitate the reporting of evidential value. To determine the usefulness of the automated systems, the results provided by the researchers must be compared with the LR values obtained from the automatic systems and the true relationship between the images. To compute a LR starting from a score, first calibration of the automated system is required. For that, we need an automatic system that provides a score and then a statistical method to convert the score into a LR. This statistical method needs a database to perform the calibration. This calibration is done using the public database Labeled Faces in the Wild. Once the LR is obtained, the performance of the automated system is evaluated through the Matthews coefficient. The Matthews coefficient condenses in a single number the quality of the classification based on the confusion matrix. The next step is to compare the LR obtained from the automated system to the LR provided by the forensic experts. This comparison between the automated system and human experts was performed with rank correlation. The overall process, and with that the structure of the paper, is illustrated in Fig. 1. In the process, the automated system and forensic experts act as actuators that receive input (both a pair of images to compare) and expel an output, scores, or distances in the case of the automated systems and likelihood ratios in the case of forensic experts. For the automated system to output likelihood ratios, it needs to be calibrated through a reference database (in this case Labeled Faces in the Wild). The final goal and main contribution of this paper is the comparison between the LR obtained by the automated system and the forensic experts, both in accuracy and similitude.

Likelihood Ratio Obtained from ENFSI Tests

ENFSI prepares every year a facial comparison test where forensic experts assess the likeliness of a match for face image pairs. Through the years, the subjects appearing in the comparisons change in nationality, quality of the picture, pose (frontal or different angle), different distances in 2011, or other challenges for face recognition such as compression of the image (2011) different ages (2012) or objects partially covering the face (2013). The characteristics of the tests evaluated can be found in Table 1.

In Table 2, the ENFSI criteria to determine the likelihood ratio associated with a certain pair of images are shown. Even

TABLE 1—*Proficiency test characteristics for years 2011, 2012, 2013, and 2017.*

	2011	2012	2013	2017
Country organizing the test	Sweden	Sweden	Sweden	Netherlands
Quality	Decent	Decent	Low (CCTV)	Good
Poses	Frontal	3 angles	Frontal	Frontal
Conditions	Distances	Similar	Similar	Similar
Other comments	Compression/ resolution	Up to 5 years in between	With glasses/ scarves. . .	–

TABLE 2—*Likelihood ratio scale that forensic experts use to assess their comparisons. Table based on Ref. (5) and ENFSI tests.*

Values of likelihood ratio	LLR value	Verbal equivalent
10,000–1,000,000	5	. . .provide *very strong* support for the first proposition rather than the alternative . . .are *far more* probable given. . . proposition. . .than proposition. . .
1000–10,000	4	. . .provide *strong* support for the first proposition rather than the alternative . . .are *much more* probable given. . .proposition. . .than proposition. . .
100–1000	3	. . .provide *moderately strong* support for the first proposition rather than the alternative . . .are *appreciably more* probable given. . . proposition. . .than proposition. . .
10–100	2	. . .provide *moderate* support for the first proposition rather than the alternative . . .are *more* probable given. . .proposition. . .than proposition. . .
2–10	1	The forensic findings provide *weak* support for the first proposition relative to the alternative. The forensic findings are *slightly more* probable given one proposition relative to the other.
0.5–2	0	The forensic findings *do not* support one proposition over the other. The forensic findings provide *no assistance* in addressing the issue.

though the true values of the likelihood ratio cover a larger range, the experts in the ENFSI tests report them on a logarithmic scale for convenience. In Table 2, the original LR value is reflected as LR, the reporter logarithmic LR as LLR, and the verbal forensic report as verbal equivalence. For LR > 1, a logarithmic scale from 0 to +5 is used (LLR). When LR < 1, the LLR will be equivalent but with a negative value (from −5 to 0).

Samples from different years are shown in Fig. 2. They are referred to as match, that is, both of the images belong to the same person, or nonmatch, which means the pictures belong to different persons. Both the investigator and the automated system must report if the comparison corresponds to match/nonmatch and the degree of certainty about it through the likelihood ratio.

Likelihood Ratio Obtained from Automated Systems

Biometric Score Obtained from Open-source Automated Systems

In forensic science, transparency and explainability are important. Three methods are chosen due to their availability to the users since no license required and the source code is available. This transparency makes OpenFace, SeetaFace, and FaceNet open-source systems suitable for forensic study, in contrast to commercial software that is not open for examination. FaceNet is used due to its high performance in the dataset used to create the LR from the scores (99.65% accuracy). OpenFace and FaceNet are both based on Ref. (21), but OpenFace has faster running time than FaceNet because of its lower number of dimensions. In principle, a higher value of dimensions provides higher accuracy, but also more computational power. Finally, SeetaFace is based on VIPL-FaceNet (20), which works with a different backbone network (the convolutional neural network that was trained to make the faces classification), and thus, the performance may be different from the other two software systems. All of them outperform human performance in the public database Labeled Faces in the Wild (14).

The three systems execute a 1:1 verification. In these automated systems (all based on a convolutional neural network), each detected face is represented as an N-dimensional vector in the space resulting from embedding the high dimensional image space to an *N*-dimensional feature space. Figure 3 shows a sketch of this procedure.

OpenFace is a Python and Torch implementation of face recognition and is based on Ref. (21). The models are trained with a combination of the two publicly available face recognition datasets: FaceScrub and CASIA-WebFace. The software used for this paper is a script that predicts a similarity score of

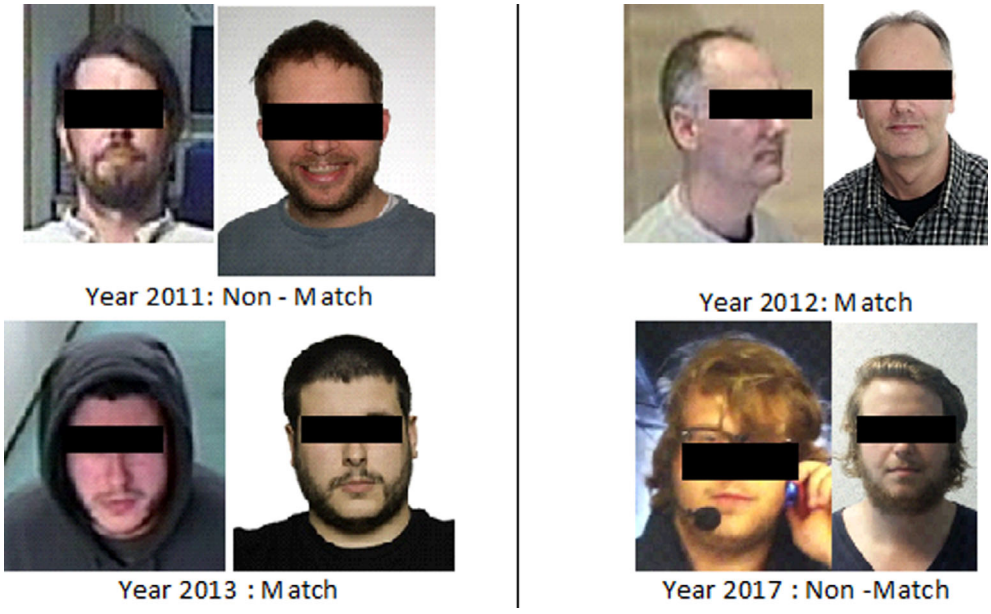

FIG. 2—*Samples of compared images. Match refers to a combination of two images that belong to the same person, and nonmatch refers to different persons. [Color figure can be viewed at wileyonlinelibrary.com]*

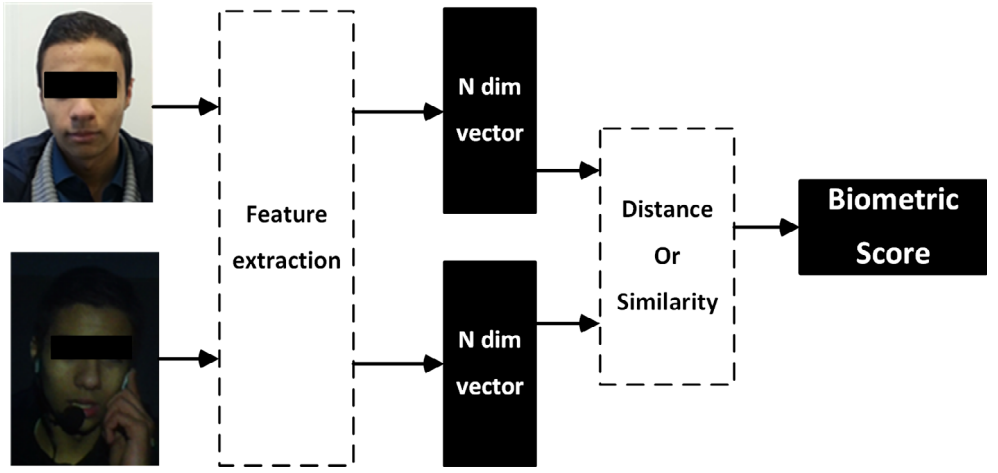

FIG. 3—*How automated systems generate a score.* N *is the number of dimensions of the embeddings for each representation. The distance measures how different the two embedded feature vectors are. [Color figure can be viewed at wileyonlinelibrary.com]*

two faces by computing the squared L2 distance between their representations, based on a normalized 128-dimensional embedding. A lower score indicates two faces are more likely of the same person. The lower the distance, the more similar the two faces are. It has accuracy on LFW of 92.92% (14). The methods in Ref. (19) also form the basis for *FaceNet* which is a TensorFlow implementation. It has been trained on VGGFace2 (34), and face alignment has been done using MTCNN (39). It does its calculations with a 512-dimensional normalized embedding and has an accuracy of 99.63% on LFW. It returns an L1 distance between 0 (same picture) and 2. Finally, *SeetaFace* is a C++ face recognition engine, which can run on a CPU with no third-party dependence. It contains three key parts, namely SeetaFace detection (40), SeetaFace alignment (41), and SeetaFace identification (20).

The image representation is a 2048-dimensional embedding, and the score provided for the comparison between two images is calculated with the cosine similarity resulting in a value between 0 (completely different) and 1 (same image). It reaches 97.1% accuracy on LFW.

From Biometric Score to Likelihood Ratio

As indicated in the introduction, the LR is obtained from two conditional probabilities namely the probability of the evidence conditional to the hypothesis of the prosecution (the two faces belong to the same person) divided by the probability of the evidence conditional to the hypothesis of the defense (the two faces belong to a different person). When we use an automatic system to calculate the similarity between the two faces to be compared, it returns a score. This score in itself has no forensic relevance and that is why we aim to convert it to an LR.

In this paper, we have chosen three methods commonly used in forensic literature (9,42) to convert biometric scores into an LR. Methods used are the Weibull model approach (22), a parametric method that approximates two probability distribution functions (PDFs), kernel density estimation (KDE) (23), a parametric method that also generates two PDFs, and the nonparametric isotonic regression that computes a cumulative distribution function (CDF) (24).

The Weibull distribution 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. Weibull 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 nonparametric test that does not make such assumptions. The third method chosen is isotonic regression commonly used machine learning model for statistical inference.

In Weibull distribution approach, if we use a sufficiently large set of scores obtained from comparisons between photographs that belong to the same person (within-source variability, WSV) and comparisons that belong to different ones (between-source variability, BSV), we can infer from these two sets two probability density functions (PDFs). Once we have these two functions, if a new comparison were made (which would be what we would consider evidence in a case), it would be enough to use the score obtained from the automated system as input and plug it in into the PDFs. Thus, we obtain two values, one for the prosecution hypothesis and another for the defense hypothesis. By dividing these two values, we obtain the likelihood ratio. A summary of this concept can be seen in Fig. 4.

The Weibull distribution is a continuous probability distribution that we fit the discrete set of scores obtained from the calibration set (LFW (13)). To approximate our set of data, we use the two-parameter Weibull, defined in Eq. 1.

$$f_w(x;\beta,\eta) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^{\beta}}. \tag{1}$$

Equation 1: Weibull function

The two-parameter Weibull distribution is commonly used in failure statistic studies and fits well with the histograms obtained with scores provided by automated systems, as seen in Fig. 5. The shape parameter (β) of the distribution changes the slope of the function, and the scale parameter (η) regulates the spread of the distribution. Their effects are illustrated in Fig. 6.

Once the calibrated data are grouped into bins on a histogram, probabilistic functions have to be fitted to the data in order to calibrate. Using both Weibull functions (prosecution generated with BSV and defense generated with WSV), LR is calculated with the following:

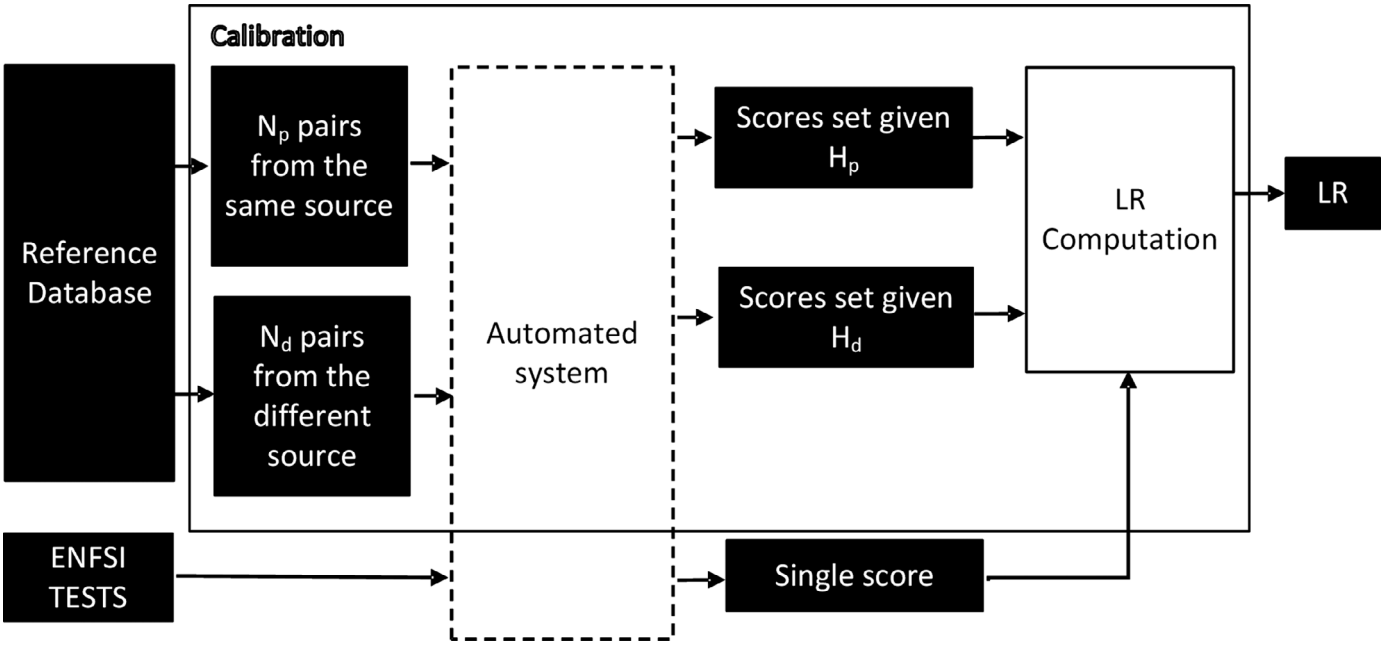

FIG. 4—*Computation of an LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen. Figure based on Ref. (9). The reference database is used to calibrate the automated system. From the calibration, two sets of scores are obtained, one for the same source pair of faces* (H_p) *and another one for different source pairs of faces* (H_d)*. For each pair of question and reference image in the ENFSI test, the automated system will provide a score. The score is transformed to an LR through the calibration methods Weibull, KDE, and isotonic regression.*

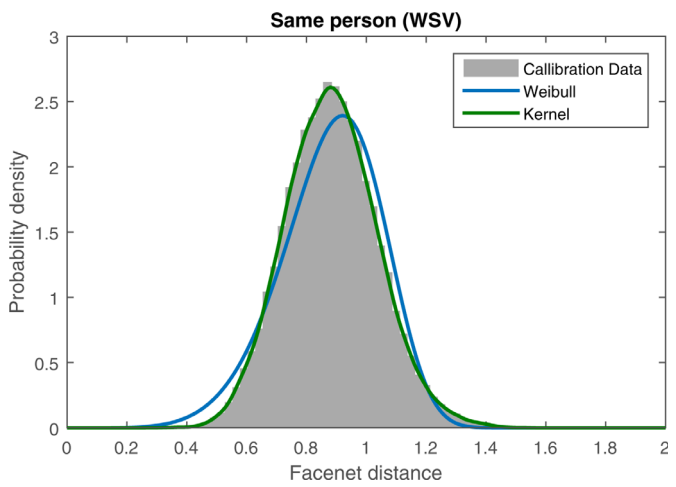

FIG. 5—*Weibull and KDE approximations to histograms generated with calibration data. [Color figure can be viewed at wileyonlinelibrary.com]*

$$\mathrm{LR}_W(s) = \frac{\Pr_w(s|H_p)}{\Pr_w(s|H_d)} = \frac{f_w^p(s;\beta_p,\eta_p)}{f_w^d(s;\beta_d,\eta_d)}$$

In kernel density estimation, A kernel distribution iss a nonparametric representation of the probability density function (PDF) of a random variable. It is used when a parametric distribution cannot properly describe the data, or when avoiding making assumptions about the distribution of the data is desired. A kernel distribution is defined by a smoothing function and a bandwidth value h, which controls the smoothness of the resulting density curve. In other words, it is a technique that lets you create a smooth curve given a set of data (23). It is given by the following equation:

$$f_k(x;h,K) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right). \quad (2)$$

Equation 2: KDE equation

where K is the kernel and h is the bandwidth. The kernel smoothing function defines the shape of the curve used to generate the probability distribution function. Similar to a histogram, the kernel distribution builds a function to represent the probability distribution using the sample data. Unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve. For this paper, we will use a Gaussian kernel for the calibrations. The bandwidth steers the smoothness of the resulting approximation. The effect of this parameter is illustrated in Fig. 7. It can be observed that small bandwidth values (0.1) can generate overfitting.

Using both kernel functions (prosecution generated with BSV and defense generated with WSV), LR is calculated with the following:

$$\mathrm{LR}_k(s) = \frac{\Pr_k(s|H_p)}{\Pr_k(s|H_d)} = \frac{f_k^p(s;h,K)}{f_k^d(s;h,K)}$$

Isotonic regression (pool adjacent violators algorithm) can be understood as approximating given series of 1-dimensional observations with a nondecreasing function which has to lie as close to the observations as possible. Isotonic regression is given by the following formula (43):

$$\min_{g\in A}\sum_{i=1}^{n} w_i(g(x_i)-f(x_i))^2. \quad (3)$$

Equation 3: Isotonic regression formula

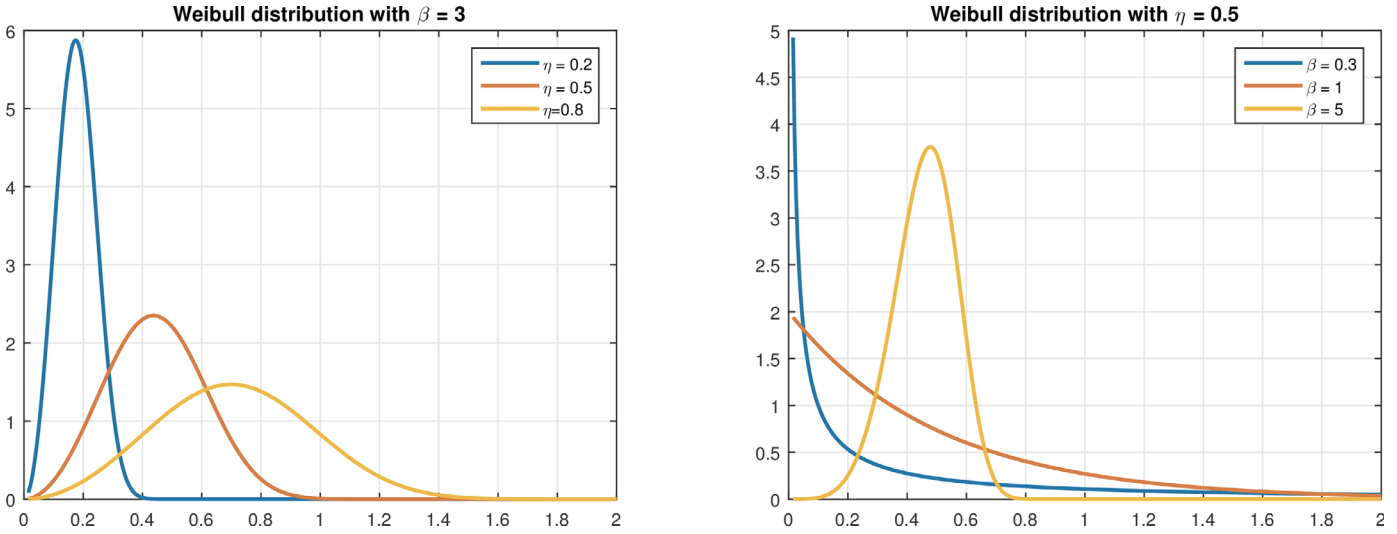

FIG. 6—*Different shape parameters (left figure) and scale parameters (right figure) in the Weibull distribution. The shape parameter (β) of the distribution changes the slope of the function, and the scale parameter (η) regulates the spread of the distribution. [Color figure can be viewed at wileyonlinelibrary.com]*

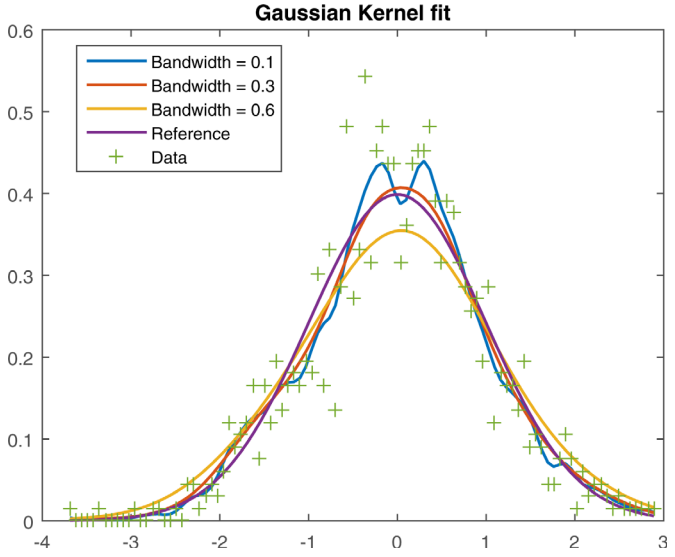

FIG. 7—*KDE with different bandwidth values (*h*). The bandwidth steers the smoothness of the resulting approximation. Higher values of* h *smooth the curve, whereas the low values make the curve fit the samples better. However, this can cause overfitting. [Color figure can be viewed at wileyon linelibrary.com]*

where A is the set of all piecewise linear, nondecreasing, continuous functions and f is a known function.

To apply the linear isotonic regression method, we use the pool adjacent violators algorithm (PAVA). Applying PAVA, an increasing function from the scores of a distance (OpenFace and FaceNet) or similarity (SeetaFace) is built. The input to feed the function is calibration scores from both WSV and BSV. In OpenFace and FaceNet, WSV corresponds to low score values (WSV corresponds to a comparison of the same person) and BSV corresponds to high values (comparisons of different persons). The larger the distance value, the higher the probability of the input being different persons. The relationships are completely the opposite of SeetaFace.

Each score obtained from the automated system is assigned a point in the xy plane. In this plane, x is the value of the obtained distance (in OpenFace and FaceNet) or similarity (in SeetaFace). The variable y will be assigned a value of 0 if it belongs to WSV and a value of 1 if it belongs to BSV (OpenFace, FaceNet), and the opposite for SeetaFace. Figure 8 left shows a scatter of this value allocation. To achieve isotonic regression, the requirements $y_{i+1} \geq y_i$ for every $x_{i+1} > x_i$ must be satisfied. As seen in Fig. 8, the distance values obtained are discrete, they do not satisfy $y_{i+1} \geq y_{i.}$ To satisfy this term, PAVA is applied. The outcome of PAVA is a nondecreasing function with $y_{i+1} \geq y_i$.

There are points with x values that are equal (i.e., $x_{i+1} > x_i$ is not satisfied). All the points with the same x value are substituted by one that has the y value of the average. Also, that point is assigned a weight equal to the number of original points for that x value. With this step, a point cloud with different weights is obtained, but this time $x_{i+1} > x_i$ is satisfied for every i, as shown in Fig. 9. The next step is applying the pool adjacent violators algorithm (PAVA) making sure the requirement $y_{i+1} \geq y_i$ is satisfied. Going from the smallest x value in increasing order, if a violation of this requirement is encountered, the value of the point y_{i+1} (the violator) and the **left** adjacent points with the same y value are changed to the average of all of them, considering the assigned weights. With that, the decrease in the function is avoided at this point, augmenting the value of the violator and decreasing the value of the adjacent left points. However, after this step, it is possible that a new violator to the left of x_i has been created. It is for that reason that after a change in the value it is required to start from the smallest value of x again. The algorithm ends when all the violators are eliminated, that is, the obtained points define a nondecreasing function as shown in Fig. 8 right.

The resulting function can be considered an estimation of the probability of the comparison being two different persons, conditioned on a distance value or evidence. Also defined as following:

$$y(x) = P(\mathrm{BSV}|x).$$

Hence, its complementary value to 1 corresponds to the probability that the two people in the comparison are the same person conditioned to a distance value (or evidence):

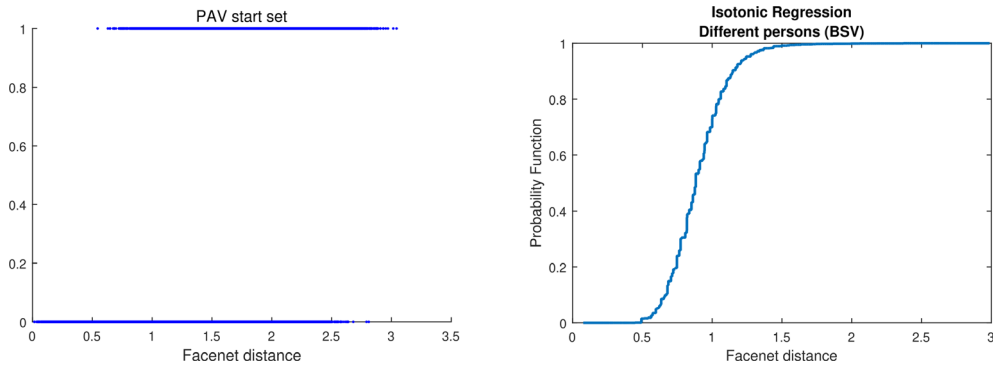

FIG. 8—*Left figure: points* (x_i, a_i)*, where* a_i *= 0 or 1, depending on the scores obtained when the person is the same (0) or different (1). Right figure: outcome of PAVA. This curve is the nondecreasing curve which best fits the set of scores in the left figure. [Color figure can be viewed at wileyonlinelibrary.com]*

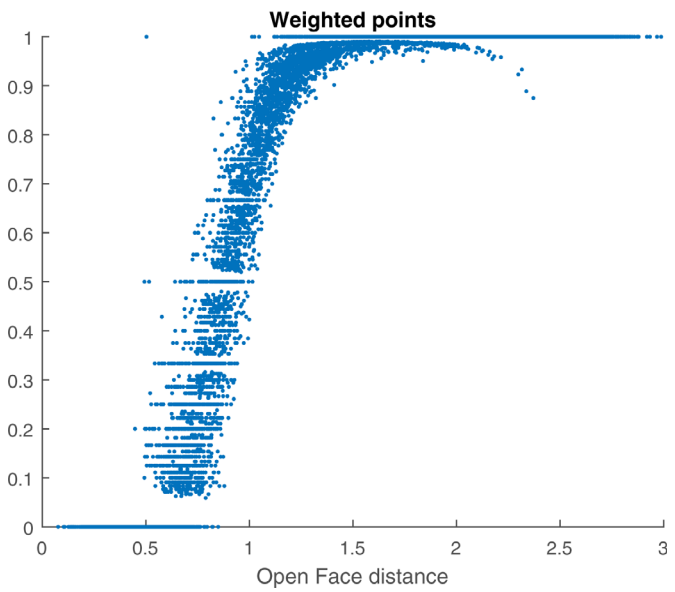

FIG. 9—*Point cloud with* $x_{i+1} > x_i$ *satisfied. Data points indicated in Fig. 8 left are not suitable for the PAVA. Points with the same value in the x-axis are substituted by a single weighted point. The result is a cloud of points all of them with different* x *values. [Color figure can be viewed at wileyonlinelibrary.com]*

$$1 - y(x) = P(\mathrm{WSV}|x).$$

The division of the two returns the LR:

$$\mathrm{LR}_{\mathrm{ISO}}(s) = \frac{1 - y(s)}{y(s)}$$

Dataset for Calibration Data

To perform the actual calibration, a large dataset is needed from which we can learn the required probability functions. We do so by employing the Labeled Faces in the Wild database (13). This is a database of face photographs designed for studying the problem of unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the Web. Each face has been labeled with the identity of the person pictured. 1680 of the people pictured have two or more distinct photographs in the dataset (13). It is widely used as a benchmark for face recognition performance. With this dataset, two sets of image pairs are generated: pairs of the same person (WSV) and a different person (BSV). Around 137,000 comparisons were performed in this dataset to achieve the calibration test.

Comparing ENFSI Investigators and Automated Systems

Correlation Between Automated Systems and Investigators

We now move to the comparison of the automated system and the human expert. This comparison is done with the Spearman correlation coefficient (referred to as rank correlation from now on). A graphical description of this comparison can be seen in Fig. 10.

The correlation between the n-dimensional vector LLR (logarithmic likelihood ratio) given by an investigator (x) and the vector LLR computed by the software (y) is as follows:

$$\rho_{xy} = 1 - \frac{6\sum d^2}{\sqrt{n\sum(n^2 - 1)}}$$

where d is the difference between the ranks of the two vectors, and n is the length of each vector. The possible values of this coefficient go from −1 (opposing criteria between the investigator and the automated system) to +1, which expresses perfect concordance of criteria. A value of 0 means no relation between them or randomness. We use the LLR due to the nature of the ENFSI tests, in which the investigators provide LLR instead of LR. For automated systems, the LLR is computed using the values in Table 2.

Confusion Matrix

To get insight into the performance of a set of results, being it from an investigator or an automatic system, we use a confusion matrix. The following terms play a role here:

TP: true positives—the number of cases where both images are considered belonging to the same person and it was a match.

FP: false positives—the number of cases where both images are considered belonging to the same person and it was not a match.

TN: true negatives—the number of cases where both images are considered belonging to different persons and it was not a match.

FN: false negatives—the number of cases where both images are considered belonging to different persons and it was a match.

Confusion Matrix		Actual	
		WSV	BSV
Prediction	Same person	TP	FP
	Different person	FN	TN

From these values, a set of other metrics can be calculated namely:

Precision: TP/(TP + FP)
NPV: negative predicted value = TN/(TN + FN)
Sensitivity: TP/(TP + FN)
Specificity: TN/(TN + FP)

These values are expressed as percentages, and the classification is better when they are near to 100%.

Matthews Correlation Coefficient

Based on the confusion matrix, we can compute another measure of classification namely the Matthews correlation coefficient (MCC) given by

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) * (\mathrm{TP}+\mathrm{FN}) * (\mathrm{TN}+\mathrm{FP}) * (\mathrm{TN}+\mathrm{FN})}}$$

This coefficient condenses in only one value the quality of the binary classification. The absolute value of this coefficient is less or equal to 1. The higher the value, the better the classification is. A value of zero means that the classification is as good as a random one.

Log-Likelihood Ratio Cost (C_{llr})

A final measure we consider is the log-likelihood ratio cost which is based on LR values directly (10):

$$C_{\mathrm{llr}} = \frac{1}{2 \cdot N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{\mathrm{LR}_{i_p}}\right) + \frac{1}{2 \cdot N_d} \sum_{j_d} \log_2 (1 + \mathrm{LR}_{j_d}).$$

where N_p and N_d are the number of cases, H_p and H_d are true, respectively, and LR_p and LR_d are the likelihood ratios for these cases. This coefficient is always positive, and the lower the value, the better the performance of LR values is. In this paper, C_{llr} is only used to compare calibration methods, not to compare them to forensic investigators.

Results

To present comparisons between the automated system and forensic investigators, correlation graphics and boxplots will be used. Although ROC and FAR/FRR are commonly used in literature, they do not apply to this paper because they can only be obtained from calibration data. The data obtained from investigators are not enough for this kind of graph. We show for representation the correlation and results from ENFSI test 2011 in Fig. 11, and the rest of the years (2012, 2013, and 2017) are available in the annex.

ENFSI Test 2011

Figure 11, Figures S1, S3, and S5 (in the annex) show the rank correlation between each of the three scores to LR methods (Weibull, KDE, and IR) and every one of the investigators with the three types of automated system described before. They present the investigators ordered by their correlations concerning the three methods (Weibull, KDE, and IR).

Figure 12 (left figures, Figures S2, S4, S6) show the right (TP + TN) and wrong (FP + FN) answers of investigators (blue *x*) and automated systems (red triangles) and (right figures) the

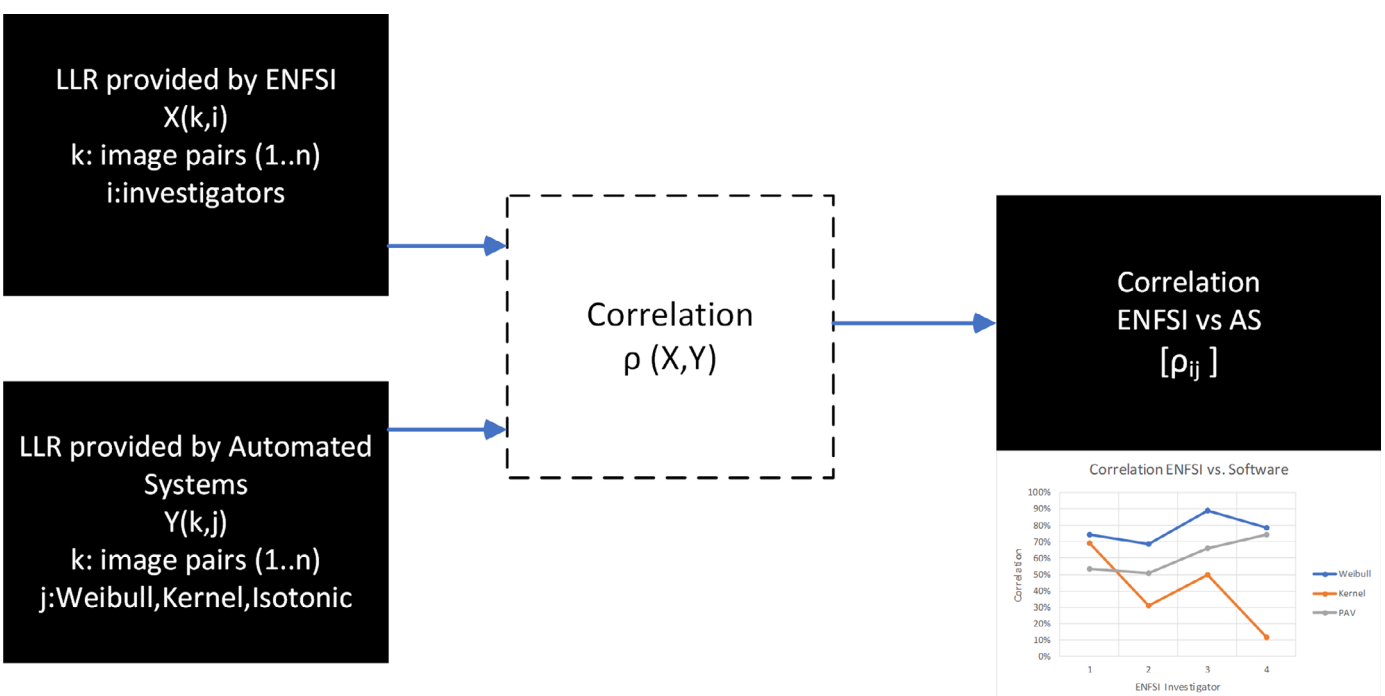

FIG. 10—*Correlation among ENFSI investigators and automated systems (AS). There are two logarithmic likelihood ratios (LLRs) obtained. First one from the forensic experts and the second one from automated systems. They are compared through a correlation, and a matrix is obtained and is represented in graphs. [Color figure can be viewed at wileyonlinelibrary.com]*

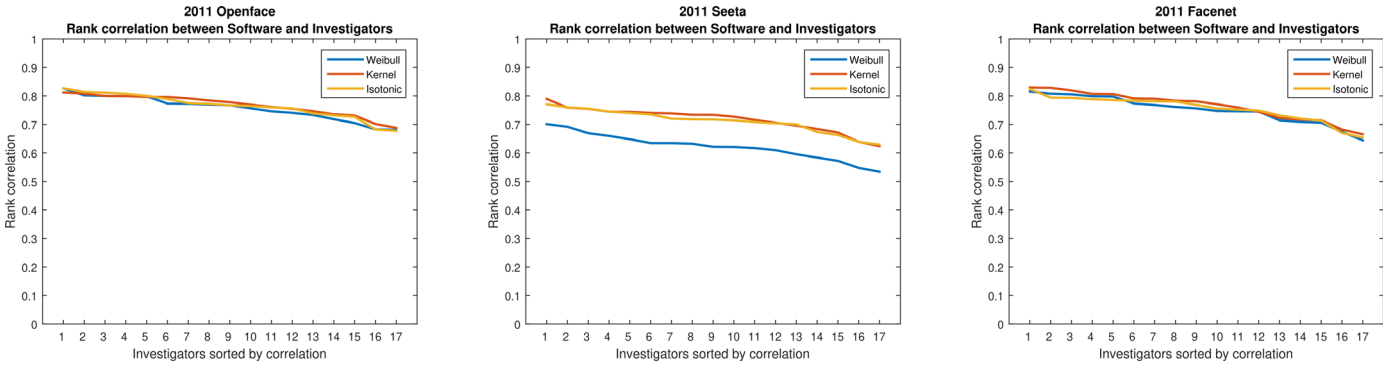

FIG. 11—*Correlation ENFSI vs automated systems, year 2011. These graphs show the correlation between each of the three scores to LR methods (Weibull, KDE, and IR) and every one of the investigators with the three types of automated system (OpenFace, SeetaFace, and FaceNet). Each figure represents one automated system: on the left, OpenFace; on the center, SeetaFace; and on the right, FaceNet. Higher values indicate higher concordance between the forensic expert and the automated software. The forensic experts are ordered from left to right according to the highest to the lowest correlation. [Color figure can be viewed at wileyonlinelibrary.com]*

individual values of confusion matrix with investigators results (boxplot) and automated systems (red triangles).

For the experiments realized in the year 2011, one can see that out of the three software programs, the highest correlation is presented by FaceNet, closely followed by OpenFace. The three calibration methods have very similar results, except for Seeta, for which Weibull has less correlation than the other methods. Seeta has a higher number of wrong answers for an equivalent number of right answers to OpenFace. In OpenFace case, the most accurate method is the isotonic regression. In FaceNet, the number of correct answers significantly higher resembles the investigators. The best procedures are Weibull and KDE.

OpenFace has several right answers similar to the researchers, but more failures. The true positives of the three methods are equal to the researchers, and the true negatives are somewhat inferior. But OpenFace has more false negatives and false positives than researchers. Seeta hits all true negatives; however, it is below in the true positives. It has 0 false positives and high false negatives. FaceNet such as Seeta hits all the negatives but has fewer false negatives. Weibull and KDE are as good as OpenFace to hit the true positives, and they also have the 3 methods 0 false positives.

ENFSI Test 2012

From Figure S1, one can see that the correlation between the methods and the investigators reaches negative numbers for OpenFace which indicates opposite criteria to forensic experts. With Seeta, correlation values stay positive but low. In two out of the three methods, Weibull performs better than the other two methods.

Looking at Figure S2 (left), it can be observed that OpenFace did not detect all of the faces and consequently returned few outputs (13 out of 30). The number of right answers is similar to the number of wrong answers. This software has poor quality results with images taken in different poses. Seeta performs a good number of true negatives, but it also has a high number of false negatives and low true positives. Nevertheless, investigators had a higher number of false positives. FaceNet behaves very similarly to Seeta.

In year 2012 experiments, the researchers have a great dispersion with the true negatives (Figure S2 right). Seeta and FaceNet have surpassed the researchers in the true negatives, and the three types of software have had a terrible rating in true positives, well below humans. Seeta and OpenFace have no false positives; however, they have many false negatives.

ENFSI Test 2013

With Figure S3, it can be noted that the correlations with FaceNet given by the three methods are very similar. However, with Seeta, Weibull calibration stands out among the other two. Correlations are higher in FaceNet than the others and in Seeta-Weibull higher than in OpenFace.

The number of right and wrong answers (Figure S4 left) with OpenFace is the same for the three density estimation methods, and similar to the ones Seeta has. For Seeta, the best density function model for calibration is Weibull. Seeta has less true positives and more false negatives compared with investigators. Nevertheless, its performance is better than investigators concerning true negatives and false positives. For FaceNet, isotonic regression results in a good number of true negatives, but a bad number of false negatives. Weibull and KDE behave similarly with a good number of false positives and negatives, and moderate numbers of true positives and negatives.

OpenFace has the highest rating in true positives, better than humans, and Seeta is the best with true negatives, also surpassing humans. OpenFace has many false positives; however, Seeta and FaceNet are at the same level as humans (Figure S4 right).

ENFSI Test 2017

For the year 2017 (Figure S5), Seeta calibration presents higher correlation values than other years, but FaceNet is the automated system with the best results in terms of correlation with investigators and KDE seems to be the best approximation. OpenFace has the worst results and isotonic performs better than Weibull and kernel.

The quality of results (right and wrong answers in Figure S6 left) is much better in Seeta than OpenFace with any of the three methods. The three density function estimation methods behave similarly in both Seeta and OpenFace. In FaceNet, the right answers and wrong answers are similar to Seeta with Weibull being the best option.

FaceNet using the Weibull and KDE methods is the one method with the highest number of true positives, equal to the majority of the researchers (median). However, the true negatives have been detected by Seeta very well and OpenFace very badly. While Seeta does not have any false positives, FaceNet and above all, OpenFace has many more than researchers as can be seen in Figure S6 right.

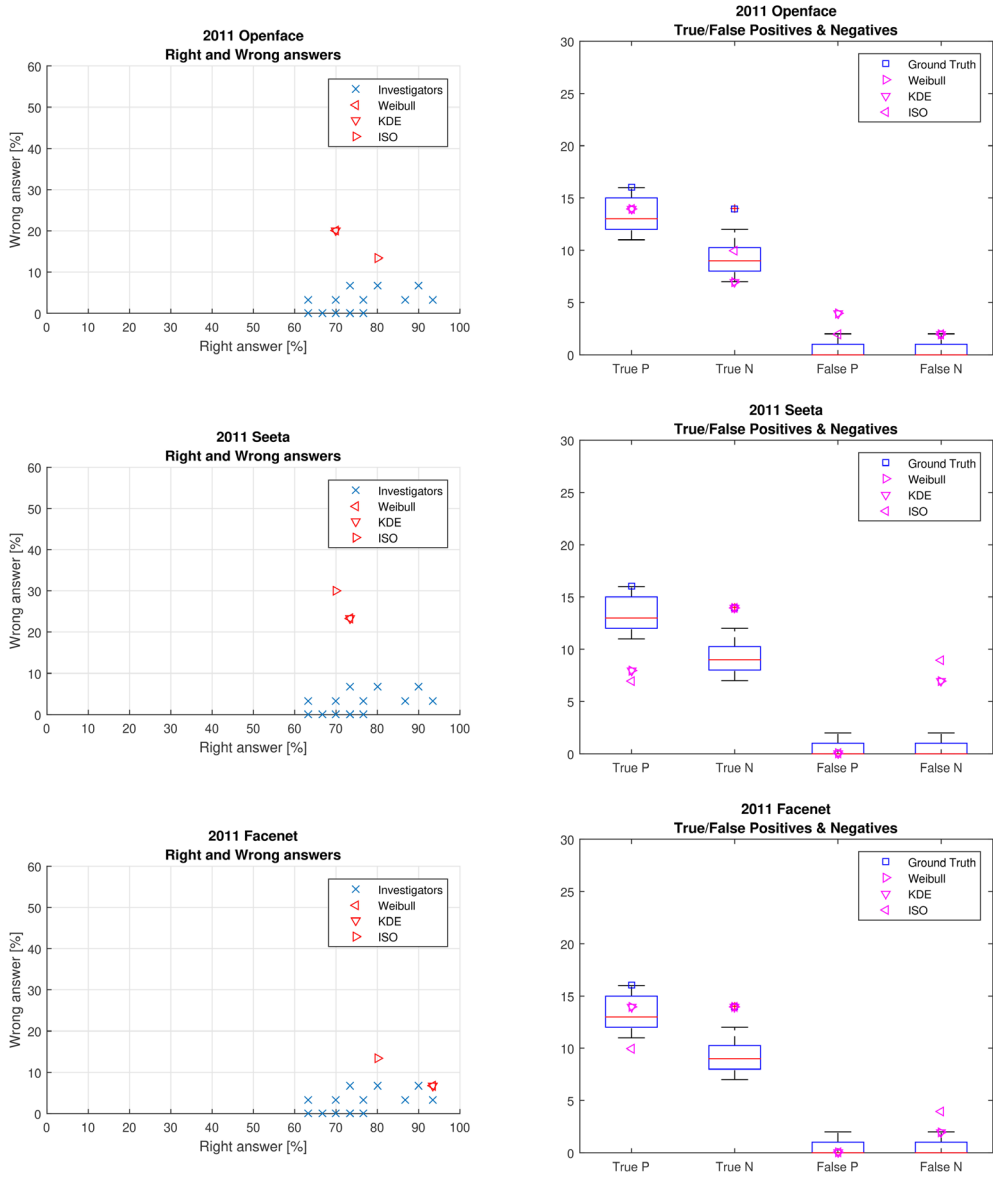

FIG. 12—*Right and wrong answers. Binary classification results. Year 2011. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts, and triangles, automated systems. On the right, a boxplot of the false positives, false negatives, true positives, and true negatives is shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE, and isotonic regression) is superimposed in the same graph. [Color figure can be viewed at wileyonlinelibrary.com]*

In conclusion, in all the tested years (2011, 2012, 2013, and 2017), the method that performs better is not always the same and it depends on the quality and poses of the images.

Confusion Matrix and MCC Results

A summary of the findings can be seen in the following Tables 3–5. From them, we can see that the quality of classification by the investigators is better than the one by the automated systems.

Discussion

When we compare images taken in frontal poses and lateral poses, the best results with all the automated systems are obtained when poses are frontal. The three automated system softwares give more incorrect answers when pose is lateral (45 Yaw, with a slight pitch ("from above") or with the time difference (age) between reference and questioned images). When the pose is 90° yaw, the software is unable to detect the face and returns an empty answer. To detect the face, the currently used software looks for two eyes, and this is not possible with a profile image.

With lateral poses, the correlation between software and human detection is random, it contains positive and negative values, and the software returns about 50% of wrong responses, being isotonic regression the method with best results. Forensic experts provide better results in these cases but they also present low correlation among them, values about 0.4 which means that they present difficulties to take decisions and their criteria are different.

When the comparison is made only with frontal poses, the correlation between forensic experts and software is better.

TABLE 3—*Confusion matrix values for OpenFace. [Color table can be viewed at wileyonlinelibrary.com]*

OpenFace		Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	68	71	79	96
	Negative predicted value	86	78	73	96
	Sensitivity	93	86	79	98
	Specificity	50	58	73	93
	Matthews correlation coefficient	48	46	51	91
	Cllr	99	97	82	
2013	Precision	82	85	83	89
	Negative predicted value	91	83	82	100
	Sensitivity	95	89	83	100
	Specificity	71	77	82	86
	Matthews correlation coefficient	69	67	66	87
	Cllr	80	82	138	
2012	Precision	67	100	100	84
	Negative predicted value	50	56	50	81
	Sensitivity	33	33	17	83
	Specificity	80	100	100	82
	Matthews correlation coefficient	15	43	29	65
	Cllr	138	123	175	
2011	Precision	100	100	100	97
	Negative predicted value	86	87	70	95
	Sensitivity	83	83	57	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	85	85	63	92
	Cllr	53	51	66	

TABLE 4—*Confusion matrix values for Seeta. [Color table can be viewed at wileyonlinelibrary.com]*

Seeta		Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	100	100	100	96
	Negative predicted value	75	71	70	96
	Sensitivity	79	75	70	98
	Specificity	100	100	100	93
	Matthews correlation coefficient	77	73	70	91
	Cllr	59	61	84	
2013	Precision	93	93	100	89
	Negative predicted value	77	71	69	93
	Sensitivity	74	65	53	95
	Specificity	94	94	100	86
	Matthews correlation coefficient	69	62	61	82
	Cllr	65	75	109	
2012	Precision	100	100	100	84
	Negative predicted value	61	59	56	81
	Sensitivity	25	18	15	83
	Specificity	100	100	100	82
	Matthews correlation coefficient	39	33	29	65
	Cllr	247	143	211	
2011	Precision	100	100	100	97
	Negative predicted value	67	67	61	95
	Sensitivity	53	53	44	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	60	60	52	92
	Cllr	167	98	154	

When the quality of questioned images is high, forensic experts have much better results (correct answers) and high values of correlation among them (greater than 90% in many cases). In this case, software methods give as many right answers as to when then image quality is low or decent. The method with best results and correlation is Weibull but with no significant difference with respect to the others. So, for frontal poses and low-quality images, the software systems are at the same level as forensic experts, but when the quality of images is good, the experts obtain better results. We conjecture that automatic systems are not able to take advantage of little details such as scars and freckles but, at the same time, are not sensitive to occlusions of the face by glasses, hats, or microphones.

TABLE 5—*Confusion matrix values for FaceNet. [Color table can be viewed at wileyonlinelibrary.com]*

FaceNet		Weibull (%)	Kernel (%)	ISO (%)	ENFSI (%)
2017	Precision	83	83	100	96
	Negative predicted value	100	89	86	96
	Sensitivity	100	95	87	98
	Specificity	67	67	100	93
	Matthews correlation coefficient	75	67	86	91
	Cllr	59	58	50	
2013	Precision	88	88	100	89
	Negative predicted value	92	86	72	93
	Sensitivity	94	88	53	95
	Specificity	85	86	100	86
	Matthews correlation coefficient	79	74	62	82
	Cllr	60	57	76	
2012	Precision	83	83	100	84
	Negative predicted value	60	62	54	81
	Sensitivity	38	38	7	83
	Specificity	92	93	100	82
	Matthews correlation coefficient	37	38	20	65
	Cllr	164	138	163	
2011	Precision	100	100	100	97
	Negative predicted value	88	88	78	95
	Sensitivity	88	88	71	96
	Specificity	100	100	100	96
	Matthews correlation coefficient	88	88	75	92
	Cllr	58	43	60	

To perform the calibration, the LFW database was used, which is unrelated to the ENFSI tests. LFW is large but may be biased due to most of the images being high quality and a lot of them frontal images. That gives room to better results in the LR obtained computed with scores in the case of fully frontal comparisons. Another public dataset, SC Faces was tested but offered similar results as LFW. To check that a large unrelated database provides better results than a small biased one, another experiment was made. The ENFSI tests were not used only as a test, but also as the mean to transform scores to LR. The number of comparisons was significantly reduced due to the number of pictures available (from ~ 130,000 comparisons in LFW to ~50 in ENFSI tests) resulting in score sets that are difficult to fit with a function. Hence, the LR computed using the ENFSI report as a data generator provides worse results than using a big, unrelated database. We could conclude that it is better to use a large unrelated dataset to the case material than to calibrate the system in data that are closer to the case material but biased. As proven by the results, the machine behaves more similarly to the forensic expert if the calibration dataset is large and unrelated to the test data than if it is of the same characteristics of the test data but a small number of images to calibrate. This can be seen in Fig. 13. The left graph corresponds to the results of a calibration

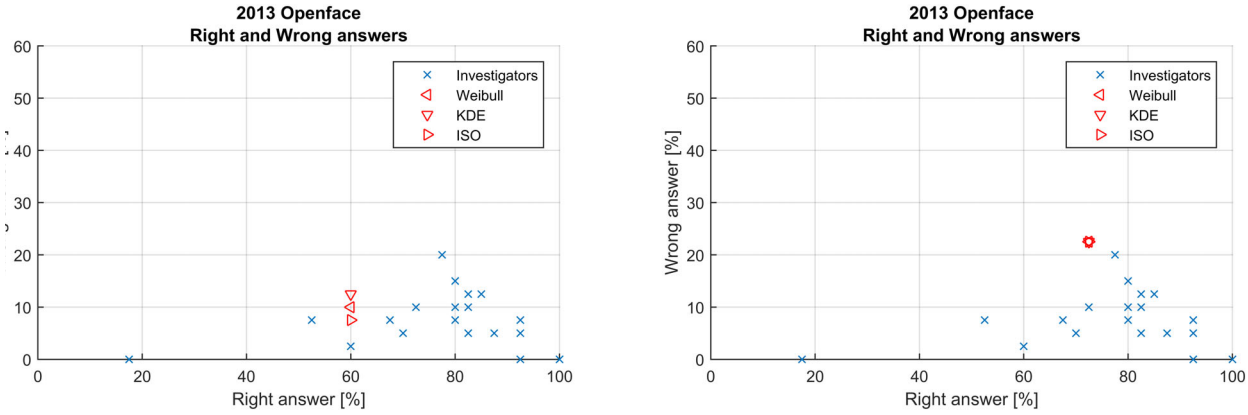

FIG. 13—*Difference in results with different calibration datasets. The left graph corresponds to the results of a calibration computed with the ENFSI tests themselves for the year 2013 (few samples for both WSV and BSV), whereas in the right there are the results for calibration made with the LFW dataset. [Color figure can be viewed at wileyonlinelibrary.com]*

computed with the ENFSI tests themselves for the year 2013 (few samples for both WSV and BSV), whereas in the right there are the results for calibration made with the LFW dataset. The difference is over 10 % of more right answers in the right graph than on the left.

Conclusion

Observing the results obtained after comparing proficiency tests and likelihood ratios calculated from the scores provided by OpenFace, Seeta, and FaceNet, one can say that the software can assist in reporting officers as it can do faster and more reliable comparisons with full-frontal images. Although the software presents limitations, these should not dictate what is feasible in terms of interpretation. It is expected that algorithms will evolve to adapt to all kinds of profiles and increase their performance. We have to think about it as a tool, never as a constraint to limit its usage. The expert cannot be replaced by this tool but becomes more efficient because the computer can help to reduce the amount of info to be managed doing appropriate filtering. If face comparison is conducted by two independent 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. Due to the high accuracy of the automated systems in the full-frontal images, it makes this kind of open-source system especially adequate to full-frontal images comparison, such as an ID picture to a mugshot, which can be useful to forensic experts.

Acknowledgments

We want to thank Arnout Ruifrok for his fruitful talks and comments.

References

1. Celentino JC. Face-to-face with facial recognition evidence: admissibility under the post-Crawford confrontation clause. Mich L Rev 2016;114 (7):1317–53.
2. Meuwly D, Ruifrok AC, Veldhuis RN, Spreeuwers LJ, Zeinstra CG. Forensic face recognition as a means to determine strength of evidence: a survey. Forensic Sci Rev 2018;30(1):21–32.
3. Ali T, Veldhuis R, Spreeuwers L. Forensic face recognition: a survey. In: Quaglia A, Epifano CM, editors. Face recognition: methods, applications and technology. Enschede, Netherlands: NOVA Publishers, 2012;9–15.
4. Spaun NA. Face recognition in forensic science. In: Li SZ, Jain AK, editors. Handbook of face recognition. London, UK: Springer, 2011;655–70.
5. McKenna L, McDermott S, O'Donell G, Barrett A, Rasmusson B, Nordgaard A, et al. ENFSI Guideline for Evaluative Reporting in Forensic Science: Strengthening the evaluation of forensic results across Europe (STEOFRAE). Wiesbaden, Germany: European Network of Forensic Science Institutes, 2015;30–41.
6. Lund SP, Iyer H. Likelihood ratio as weight of forensic evidence: a closer look. J Res Nat Inst Stand Technol 2017;122(27):1–32.
7. Collins A, Morton NE. Likelihood ratios for DNA identification. Proc Natl Acad Sci USA 1994;91(13):6007–11.
8. Neumann C, Champod C, Puch-Solis R, Egli N, Anthonioz A, Meuwly D, et al. Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. J Forensic Sci 2006;51(6):1255–66.
9. Ali T. Biometric score calibration for forensic face recognition [PhD thesis]. Enschede, Netherlands: University Library/University of Twente, 2017;11–26.
10. Ramos D, Krish RP, Fierrez J, Meuwly D. From biometric scores to forensic likelihood ratios. In: Tistarelli M, Champod C, editors. Handbook of biometrics for forensic science. Cham, Switzerland: Springer International Publishing, 2017;305–27.
11. Turk MA, Pentland AP. Face recognition using eigenfaces. In: Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 1991 June 3-6; Maui, HI. Piscataway, NJ: IEEE, 1991;586–91.
12. Parkhi OM, Vedaldi A, Zisserman A.Deep face recognition. In: Xie X, Jones MW, Tam GKL, editors. Proceedings of the 26th British Machine Vision Conference; 2015 Sept 7–10; Swansea, U.K. Durham, UK: British Machine Vision Association, 2015;41.1–41.12.
13. Huang GB, Mattar M, Berg T, Learned-Miller E.Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Amherst, MA: University of Massachusetts, 2007; Report No.: 7–49.
14. Masi I, Wu Y, Hassner T, Natarajan P. Deep face recognition: a survey. In: Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI); 2018 Oct 29 – Nov 1; Foz Do Iguacu, Brazil. Piscataway, NJ: IEEE, 2018;471–8.
15. Salici A, Ciampini C. Automatic face recognition and identification tools in the forensic science domain. In: Piva A, Tinnirello I, Morosi S, editors. Digital communication. Towards a smart and secure future Internet. Cham, Switzerland: Springer International Publishing, 2017;8–17.
16. Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proc Natl Acad Sci USA 2018;115(24):6171–6.
17. Ramos D, Haraksim R, Meuwly D. Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. Data Brief 2017;10:75–92.
18. Haraksim R, Ramos D, Meuwly D. Validation of likelihood ratio methods for forensic evidence evaluation handling multimodal score distributions. IET Biom 2017;6(2):61–9.
19. Amos B, Ludwiczuk B, Satyanarayanan M.OpenFace: a general-purpose face recognition library with mobile applications. Pittsburg, PA: CMU School of Computer Science, 2016; Technical Report CMU-CS-16-118.

20. Meina K, Wanglong W, Shiguang S, Xin CXL. VIPLFaceNet: an open source deep face recognition SDK. Front Comput Sci 2017;11(2):208–18.
21. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the 28th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2015); 2015 June 7–12; Boston, MA. Piscataway, NJ: IEEE, 2015;815–23.
22. Lai CD, Murthy DN, Xie M. Weibull distributions and their applications. In: Pham H, editor. Springer handbook of engineering statistics. London, UK: Springer, 2006;63–78.
23. Chen YC. A tutorial on kernel density estimation and recent advances. Biostat Epidemiol 2017;1(1):161–87.
24. Leeuw JD, Kurt H, Mair P. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. J Stat Softw 2009;32 (5):1–24.
25. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta Proteins Proteom 1975;405(2):442–51.
26. Aitken CCC, Taroni F. Statistics and the evaluation of evidence for forensic scientists, 2nd edn. Chichester, England: John Wiley & Sons Ltd, 2004;69–118.
27. Biedermann A, Champod C, Willis S. Development of European standards for evaluative reporting in forensic science: the gap between intentions and perceptions. Int J Evid Proof 2016;21(1–2):14–29.
28. Perlin MW. Explaining the likelihood ratio in DNA mixture interpretation. In: Proceedings of Promega's Twenty First International Symposium on Human Identification; 2010 Oct 11-14; San Antonio, TX. Fitchburg, WI: Promega Corporation, 2010;1–32.
29. Haraksim R, Drygajlo A. Measuring performance in forensic automatic speaker recognition: VQ, GMM-UBM, i-vectors. In: Brömme A, Busch C, Rathgeb C, Uhl A, editors. Proceedings of the 15th International Conference of the Biometrics Special Interest Group (BIOSIG 2016); 2016 Sep 21–23; Darmstadt, Germany. Bonn, Germany: Gesellschaft für Informatik e.V., 2016;225–32.
30. Leegwater AJ, Meuwly D, Sjerps M, Vergeer P, Alberink I. Performance study of a score-based likelihood ratio system for forensic fingermark comparison. J Forensic Sci 2017;62(3):626–40.
31. Bolck A, Ni H, Lopatka M. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. Law Probab Risk 2015;14(3):243–66.
32. Es A, Wiarda W, Hordijk M, Alberink I, Vergeer P. Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis. Sci Justice 2017;57(3):181–92.
33. Vergeer P, Bolck A, Peschier LJC, Berger CEH, Hendrikse JN. Likelihood ratio methods for forensic comparison of evaporated gasoline residues. Sci Justice 2014;54(6):401–11.
34. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. VGGFace2: a dataset for recognising faces across pose and age. In: Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15–19; Xi'an, China. Piscataway, NJ: IEEE, 2018;67–74.
35. Kwolek B. Face detection using convolutional neural networks and Gabor filters. In: Duch W, Kacprzyk J, Oja E, Zadrożny S, editors. Proceedings of the 15th International Conference on Artificial Neural Networks: Biological Inspirations (ICANN 2005); 2005 Sept 11–15; Warsaw, Poland. Berlin, Germany: Springer, Berlin Heidelberg, 2005;551–6.
36. Taigman Y, Yang M, Ranzato M, Wolf L. DeepFace: closing the gap to human-level performance in face verification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 June 23–28; Columbus, OH. Piscataway, NJ: IEEE, 2014;1701–8.
37. Sharif M, Mohsin S, Jamal MJ, Raza M. Illumination normalization preprocessing for face recognition. In: Proceedings of the 2nd Conference on Environmental Science and Information Application Technology 2010 July 17-18; Wuhan, China. Piscataway, NJ: IEEE, 2010;44–7.
38. Kumar R, Kumar P, Gupta A. A review paper on face recognition techniques. Int J Eng Res Appl 2019;7(5):223–9.
39. Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 2016;23(10):1499–503.
40. Wu S, Kan M, He Z, Shan S, Chen X. Funnel-structured cascade for multi-view face detection with alignment-awareness. Neurocomputing 2017;221:138–45.
41. Zhang J, Shan S, Kan M, Chen X. Coarse-to-Fine Auto-Encoder Networks (CFAN) for real-time face alignment. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Proceedings of the 13th European on Computer Vision (ECCV 2014); 2014 Sept 6–12; Zurich, Switzerland. Cham, Switzerland: Springer International Publishing, 2014;1–16.
42. Ali T, Spreeuwers L, Veldhuis R. A review of calibration methods for biometric systems in forensic applications. In: Veldhuis RNJ, Spreeuwers LJ, Goesling J, Shao X, editors. Proceedings of the 33rd WIC Symposium on Information Theory in the Benelux and the 2nd Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux; 2012 May 24-25; Boekelo, the Netherlands. Enschede, the Netherlands: Werkgemeenschap voor Informatie- en Communicatietheorie (WIC), 2012;126–33.
43. Shively TS, Sager TW, Walker SG. A Bayesian approach to non-parametric monotone function estimation. J R Stat Soc Series B Stat Methodol 2009;71(1):159–75.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Correlation ENFSI vs Automated systems, year 2012. These graphs show the correlation between each of the three scores to LR methods (Weibull, KDE and IR) and every one of the investigators with the three types of Automated system (OpenFace, SeetaFace and FaceNet). Each figure represents one automated system. On the left, OpenFace, on the center, SeetaFace and on the right FaceNet. Higher values indicate higher concordance between the forensic expert and the automated software. The forensic experts are ordered from left to right according to the highest to the lowest correlation.

Figure S2. Right and wrong answers. Binary Classification results. Year 2012. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts and triangles automated systems. On the right, a boxplot of the False positives, false negatives, True positives and true negatives are shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE and Isotonic regression) are superimposed in the same graph.

Figure S3. Correlation ENFSI vs Automated systems, year 2013. These graphs show the correlation between each of the three scores to LR methods (Weibull, KDE and IR) and every one of the investigators with the three types of Automated system (OpenFace, SeetaFace and FaceNet). Each figure represents one automated system. On the left, OpenFace, on the center, SeetaFace and on the right FaceNet. Higher values indicate higher concordance between the forensic expert and the automated software. The forensic experts are ordered from left to right according to the highest to the lowest correlation.

Figure S4. Right and wrong answers. Binary Classification results. Year 2013. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts and triangles automated systems. On the right, a boxplot of the False positives, false negatives, True positives and true negatives are shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE and Isotonic regression) are superimposed in the same graph.

Figure S5. Correlation ENFSI vs Automated systems, year 2017. These graphs show the correlation between each of the three scores to LR methods (Weibull, KDE and IR) and every one of the investigators with the three types of Automated system (OpenFace, SeetaFace and FaceNet). Each figure represents

one automated system. On the left, OpenFace, on the center, SeetaFace and on the right FaceNet. Higher values indicate higher concordance between the forensic expert and the automated software. The forensic experts are ordered from left to right according to the highest to the lowest correlation.

Figure S6. Right and wrong answers. Binary Classification results. Year 2017. In the figure, the graphs are deployed as follows: Figures on the left correspond to right and wrong answers from the automated systems and the forensic experts. Crosses represent experts and triangles automated systems. On the right, a boxplot of the False positives, false negatives, True positives and true negatives are shown. Boxplots are obtained from the forensic experts' data. The outcome from the three methods (Weibull, KDE and Isotonic regression) are superimposed in the same graph.