Science & Justice 64 (2024) 509–520

Contents lists available at ScienceDirect

Science & Justice

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

Improving the evidential value of low-quality face images with aggregation of deep neural network embeddings

Rafael Oliveira Ribeiro [a,b,c,*], João C. Neves [d], Arnout Ruifrok [e], Flavio de Barros Vidal [b]

[a] Aston University, Birmingham, UK
[b] Department of Computer Science, University of Brasilia, Brasília 70910-900, Brazil
[c] National Institute of Criminalistics, Brasília, 70610-902, Brazil
[d] NOVA-LINCS, University of Beira Interior, Covilhã 6201-001, Portugal
[e] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497 GB, the Netherlands

ARTICLE INFO

Keywords:
Forensic evaluation
Face comparisons
Embedding aggregation
Likelihood ratio

ABSTRACT

In forensic facial comparison, questioned-source images are usually captured in uncontrolled environments, with non-uniform lighting, and from non-cooperative subjects. The poor quality of such material usually compromises their value as evidence in legal proceedings. On the other hand, in forensic casework, multiple images of the person of interest are usually available. In this paper, we propose to aggregate deep neural network embeddings from various images of the same person to improve the performance in forensic comparison of facial images. We observe significant performance improvements, especially for low-quality images. Further improvements are obtained by aggregating embeddings of more images and by applying quality-weighted aggregation. We demonstrate the benefits of this approach in forensic evaluation settings with the development and validation of common-source likelihood ratio systems and report improvements in C_{llr} both for CCTV images and for social media images.

1. Introduction

The increasing number of indoor and outdoor surveillance cameras and the widespread availability of smartphones have raised the number of crimes in which the perpetrator's facial image is recorded. This fact has fostered the interest in using these data to uncover the perpetrator's identity [1]. Still, the uncontrolled acquisition conditions frequently result in poor quality and limited evidential value of such images for court proceedings [2,3]. This paper aims to address this issue by aggregating information from multiple facial images of the same individual under a framework suitable for forensic facial comparison.

The currently recommended method for performing forensic facial comparisons is based on the manual analysis of morphological facial features [4]. This process involves comparing morphological facial features in the questioned-source image with those in the known-source images and evaluating how similar and typical the observed features are in the relevant population [5]. This evaluation is often done subjectively by the expert, relying heavily on their experience since, to the best of our knowledge, there are no systematic databases of the relative frequencies of morphological features [6,7].

Although forensic practitioners using the current approach have demonstrated superior performance for facial comparisons relative to control groups [8,9], there has been a long-standing call for adopting more objective and quantitative methods in forensic science [10–13]. In various fields related to biometric comparisons, the research community has responded to this call by investigating the possibility of using automated systems to quantify the evidence obtained from the data by computing an LR [14–22]. Based on the evaluation of comparison scores obtained from biometric samples, this new approach is especially appealing for the face modality for two reasons. Firstly, automatic facial recognition systems have experienced an enormous improvement in performance over the last few years [23,24]. Secondly, it has been demonstrated that the combined performance of human experts and facial recognition algorithms is superior to either the human experts or the algorithms alone [25,8].

Currently, the LR paradigm is the recommended approach for evaluative reporting of source problems in forensic science [7,26]. Under this paradigm, forensic practitioners should express their evaluation using a likelihood ratio. The LR represents the degree of support of the evidence for one hypothesis relative to another mutually exclusive

[*] Corresponding author at: Aston University, Birmingham, UK.
E-mail addresses: 230254750@aston.ac.uk, rafael.ror@pf.gov.br (R.O. Ribeiro).

https://doi.org/10.1016/j.scijus.2024.07.006
Received 14 December 2023; Received in revised form 13 July 2024; Accepted 29 July 2024
Available online 7 August 2024
1355-0306/© 2024 The Authors. Published by Elsevier B.V. on behalf of The Chartered Society of Forensic Sciences. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

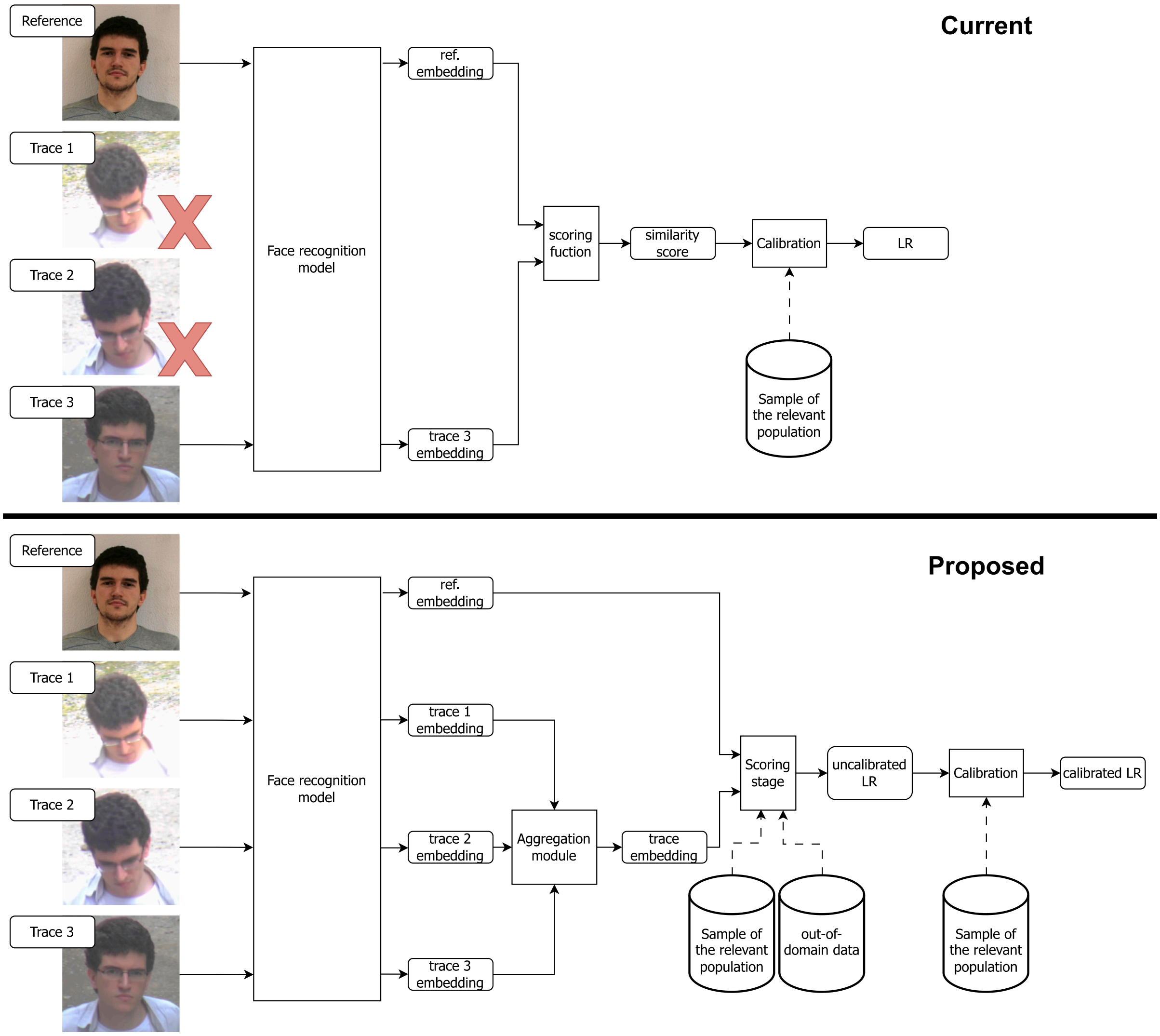


Fig. 1. Comparison between the current methods for calculating likelihood ratios for forensic facial comparisons (top) and the proposed framework (bottom). The main differences are the inclusion of the aggregation module, which improves the use of available trace material, and the scoring stage, which considers the similarity and typicality of the embeddings.

hypothesis. In this work, we consider common-source hypotheses, which, in the case of forensic facial comparison, can be defined as:

- H_p (same-source hypothesis): Both the questioned-source and the known-source images depict the face of the same person; and;
- H_d (different-source hypothesis): The questioned-source and the known-source image depict the faces of two different people from the same population[1].

The LR is computed according to

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

i.e., it is the ratio of the probabilities (P) of obtaining the evidence (E) given each hypothesis and case-relevant information (I). An example of such case-relevant information in the field of forensic facial comparison is the location (city, region or country) where the questioned-source image was captured.

Several works have proposed methods to obtain LRs by converting the similarity scores obtained from face recognition systems through the estimation of within-source and between-source distributions [20–22]. However, the existing strategies focus only on a single questioned-source image, disregarding the possibility of aggregating information from multiple images (e.g., consecutive frames from CCTV footage) to compute a single LR. Also, current proposals for calculating a score-based LR employ scores that do not consider the typicality of the facial images with respect to the relevant population. Disregarding typicality has been considered inappropriate to obtain adequate likelihood ratios [27,28]. To address both of these limitations, as depicted in Fig. 1, we introduce a novel strategy for the calculation of common-source LR in forensic facial comparison when multiple questioned-source images are available. The proposed method combines the facial embeddings[2] of each facial image. The resulting aggregated embedding is used to compute an uncalibrated likelihood ratio using Probabilistic

[1] Often referred to as *relevant population* in the forensic literature, it is the population from which an alternative suspect may have came from (e.g., young adult males from a specific region).

[2] The facial embeddings refer to the representation of a facial image obtained from a Deep Convolution Neural Network-based facial recognition system.

Linear Discriminant Analysis (PLDA) [29], which is then calibrated using a regularized logistic regression model.

The paper is organized as follows: in Section 2, we review works on using biometric systems to evaluate the comparison of facial images as forensic evidence. In Section 3, we describe the data used in this work and in 4, the proposed method is detailed. Section 5 describes the experiments performed, and Section 6 presents the results and discussion. We conclude in Section 7, presenting the limitations of this work and planned investigations on the same topic.

2. Related work

The possibility of using automated face recognition systems for quantifying forensic evidence has been studied for at least two decades [30–32,21,22].

In 2005, Gonzalez-Rodriguez et al. [31] assessed the performance of face recognition approaches for forensic applications. The authors relied on a database comprising 295 identities. They used 400 within-source comparisons and 12,250 between-source comparisons for estimating the probability density functions of the two distributions, which were subsequently used to derive the LRs from similarity scores obtained from the recognition system. The improvement in face recognition accuracy and the development of more challenging datasets fostered the proposal of novel studies in the following years.

Ali et al. [30] evaluated the log-LR obtained from within-source and between-source scores of a commercial face recognition system. Nevertheless, the authors only analyzed five identities from the Face Recognition Grand Challenge (FRGC) dataset with 35 images per identity.

Mandasari et al. [32] introduced an innovative approach based on inter-session variability modeling followed by a linear transformation of the similarity score to obtain LRs of face recognition in the Surveillance Cameras Face Database (SCface), a database comprising samples from a usual forensic scenario.

Zeinstra et al. [33] analyzed the discriminating power of facial marks in forensic scenarios. The authors proposed an innovative method based on the number of marks in each cell of an auxiliary grid superimposed over the face. The number of marks along each cell is used as the facial features that are subsequently used by the face classifier. The evaluation of the C_{llr} with respect to the number of facial marks and grid size evidenced the potential of this approach, even though the dataset considered is not particularly challenging for current face recognition systems regarding pose and occlusion.

The first publicly available study using real forensic data was carried out by Mölder et al. [34], using a national database of mugshots. However, few details were given concerning the face recognition algorithm, and the data used could not be shared.

Jacquet [35] investigated the use of LR to improve the performance of face recognition systems for investigative and forensic evaluation applications and demonstrated the feasibility of using such systems for providing evidence in court, although not all face recognition systems tested offered sufficient performance.

Ruifrok et al. [19] showed that the distribution of similarity scores could be used to assess the quality of trace images, which can be subsequently exploited to optimize the score-to-LR conversion, and consequently improve the discrimination and calibration of the obtained LRs. The previously described papers use scores that consider only the similarity of the embeddings, ignoring their typicality with respect to the relevant population. Morrison and Enzinger [27] demonstrated, using simulated and real data, that scores used to calculate likelihood ratios should consider both similarity and typicality.

3. Data

We selected datasets that represent two typical scenarios in forensic casework: surveillance and social media images.

3.1. Surveillance datasets

In surveillance scenarios, subjects' images are captured without control of pose, illumination, expression, and other factors affecting facial recognition performance. Additionally, motion blur, compression artifacts, and low resolution of the face region are typical limitations present in this kind of data. On the other hand, reference images of a suspect are usually of excellent quality and captured under controlled conditions (e.g., driver's license or passport photo). Despite the multiple datasets devised to study face recognition in the wild, few datasets mimic the conditions of real surveillance scenarios [36].

Quis-Campi [37] and SCFace [38] are the most representative datasets comprising data replicating the surveillance scenarios' degradation factors while providing high-quality reference images. Still, there are important limitations to the degree to which these datasets are representative of real forensic caserwork.

In the case of SCFace, all CCTV images are captured simultaneously by different cameras. In real casework, it is more common that a sequence of images is captured by the same camera. This aspect is better represented in the Quis-Campi dataset, where the CCTV images were captured by a single camera. Even then, it is common in casework that a sequence of frames, i.e., a video containing the face of interest, is available. The CCTV images from Quis-Campi comprise a few selected frames per person.

The SCface dataset contains CCTV images of 130 subjects, captured at three different distances (*far* - 4.2 m, *medium* - 2.6 m, and *close* - 1.0 m)[3] from multiple cameras in the visible and infra-red spectrum. Additionally, it provides high-quality reference images captured in frontal pose and at varying degrees of lateral poses [38]. In our experiments, only the high-quality frontal images are considered references in the 1:1 comparisons. As for the CCTV images, which we use as traces, we only use images from the five cameras in the visible light spectrum.

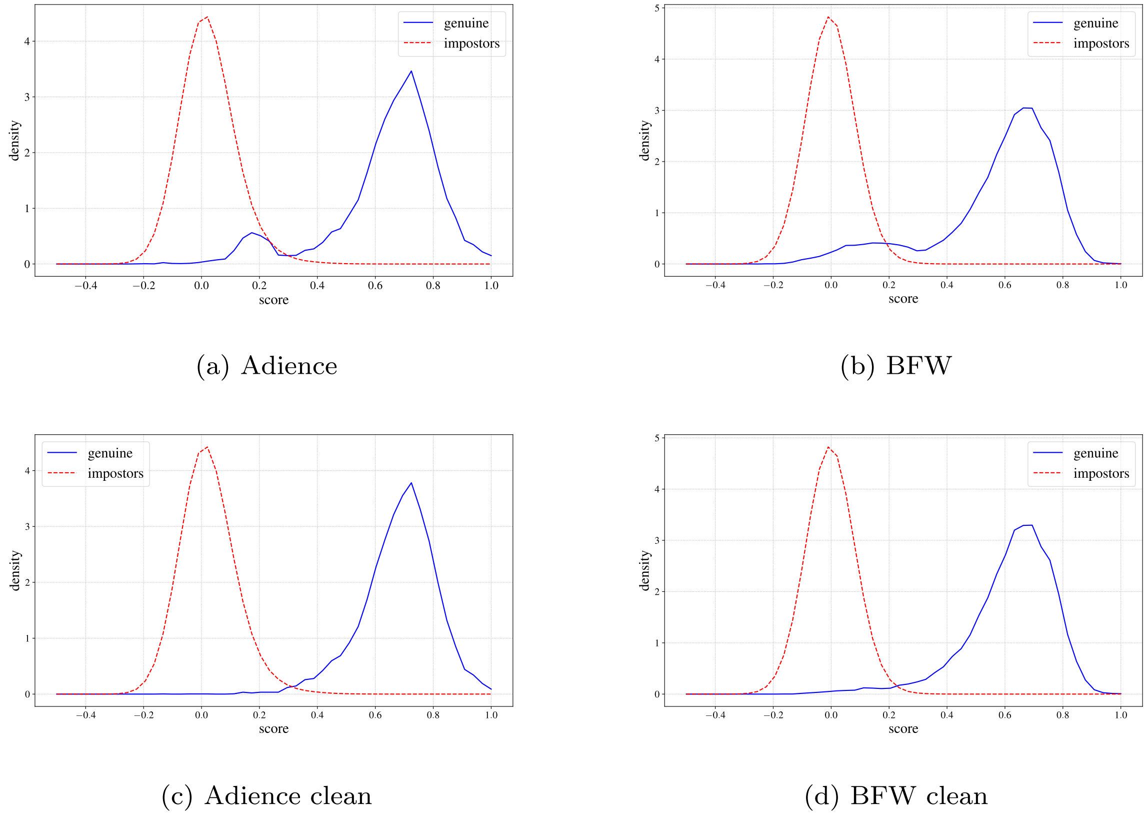
The Quis-Campi dataset contains CCTV images of 320 subjects captured in an uncontrolled outdoor environment. In addition to variations in pose and distance, also present in the SCface dataset, surveillance images from the Quis-Campi dataset have significant variations in illumination, occlusion, and facial expression. Motion blur is also present in some images. Each subject has one frontal and two lateral profile reference images, with controlled illumination and neutral expression. Only frontal images are used as references in this work. We selected a subset of the Quis-Campi dataset for our experiments with 192 identities, for which at least one reference and one CCTV image were available.

3.1.1. Novel verification protocol for the quis-campi dataset

To evaluate the proposed method in a more forensically realistic scenario, we present a new verification protocol[4] for the Quis-Campi dataset, based on the concept of *encounters*. In this protocol, the surveillance images of each identity are grouped into sets of images captured during an encounter of the person of interest with the camera. For this purpose, we selected a threshold of two minutes as the criteria for separating the encounters of each person in the dataset. Each group of trace images of an encounter is compared to the corresponding reference image, according to the strategies described in Section 4. This protocol is representative of cases where images of a perpetrator are registered in the video, and no other surveillance images that can be safely attributed to the same perpetrator are available. Results using this protocol are referred to as *Quis-campi encounters*.

[3] We refer to the experiments with the three different subsets of the surveillance images of the SCface dataset as SCface 1 for images captured at 4.2 m, SCface 2 for images captured at 2.6 m and SCface 3 for images captured at 1 m.

[4] Specification of this protocol is available at https://github.com/rafribeiro/embedding_aggregation.



(a) Adience

(b) BFW

(c) Adience clean

(d) BFW clean

Fig. 2. Bi-modal behavior of genuine scores distributions for the Adience (a) and BFW (b) datasets, suggestive of identity labeling errors. After cleaning, the genuine distributions no longer exhibit this bi-modal behavior (c, d).

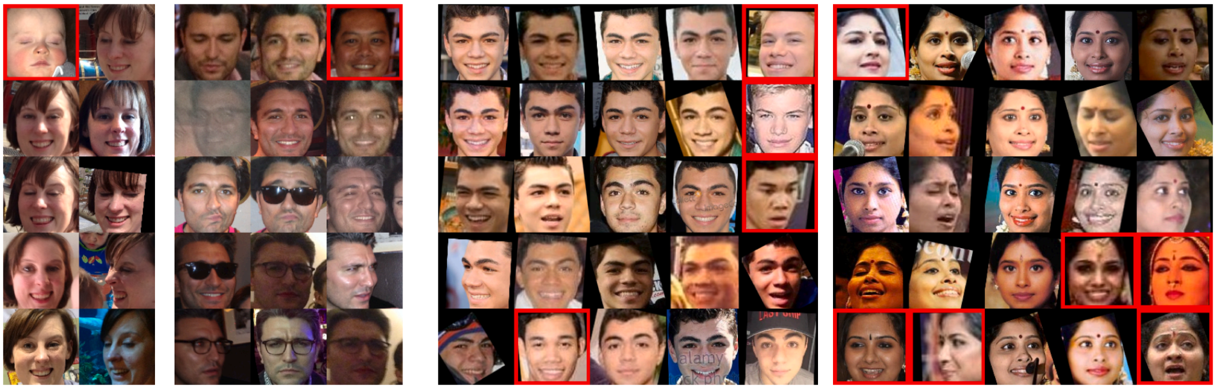
3.2. Social media datasets

Images obtained from social media platforms are usually of better quality than those obtained in surveillance scenarios. Nevertheless, these data still exhibit large variations in pose, illumination, facial expression, and resolution. Moreover, traditional and digital makeup effects are also frequently present in these images. Two datasets were selected to evaluate our approach in this scenario: Adience [39] and Balanced Faces in the Wild (BFW) [40]. We note that it may not be trivial to guarantee that multiple questioned-source images collected from social media platforms could be assigned to the same person, which would undermine the applicability of the proposed approach for this kind of questioned-source images. Nevertheless, the proposed approach can be applied to cases where multiple reference, known-source images are obtained from social media platforms.

The Adience dataset was created to study age and gender recognition in data obtained in real-world imaging conditions. For this, 26,580 photos of 2,284 subjects were obtained from online image repositories. Images were acquired using smartphones and other mobile devices and presented significant variations in pose, lighting condition, facial expression, and image quality.

Considering that the number of images per identity is heavily imbalanced, we selected a subset of the Adience dataset, including identities with at least 11 images - one for reference and at least ten as traces. This selection resulted in a set of 14,143 images from 373 identities.

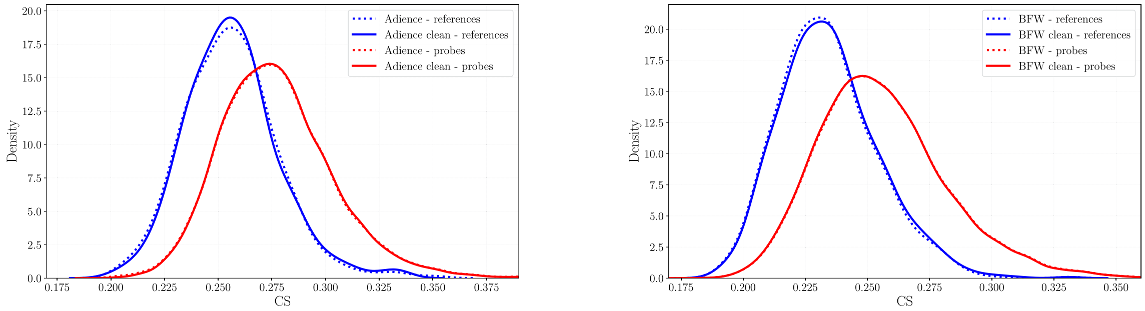
The BFW dataset contains 20,000 images of 800 individuals labeled for gender (female, male) and ethnicity (Asian, Black, Indian, White).



(a) Adience

(b) BFW

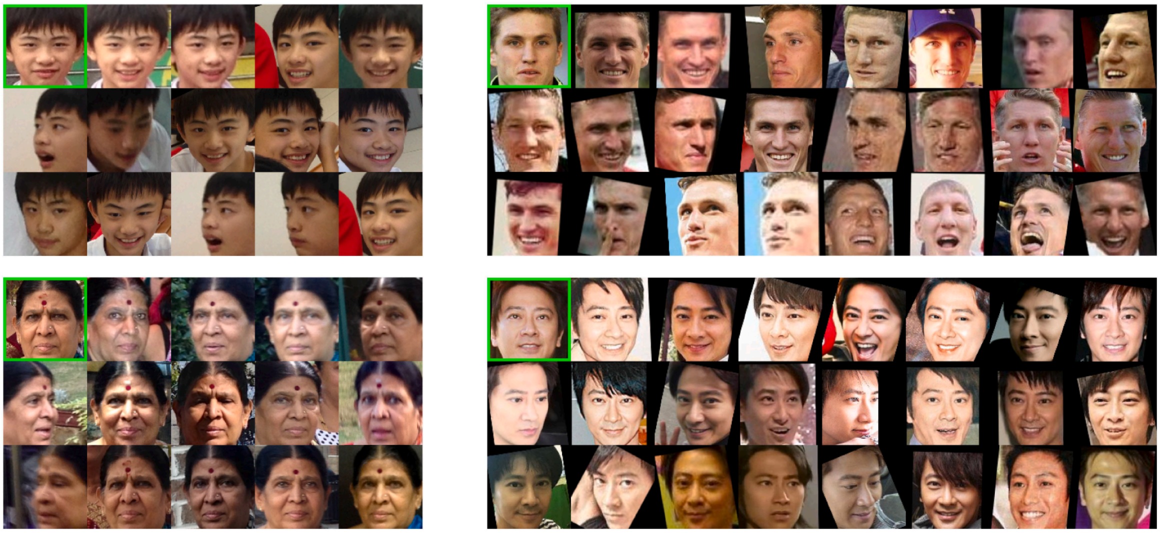
Fig. 3. Examples of identity labeling errors (red boxes) in the Adience and BFW datasets.



(a) Adience

(b) BFW

Fig. 4. Distributions of Confusion Scores for the references and probes from the BFW and Adience datasets, before and after cleaning.



(a) Adience clean

(b) BFW clean

Fig. 5. Examples of references selected for the Adience clean and BFW clean datasets. For each identity, the face at the top left (in green) is selected as a reference, and the others are used as traces.

The dataset is balanced, with 25 images per subject and 100 subjects in each demographic group.

3.2.1. *Identity errors in adience and BFW Datasets*

During our preliminary experiments on the Adience and BFW datasets, we observed an atypical bi-modal distribution of the genuine scores (Fig. 2a and 2b). This unexpected behavior raised suspicion that errors in the identity labels might be present in these two datasets.

A manual review of the images most frequently involved in low genuine scores confirmed that many identity labels were incorrect in both datasets. Fig. 3 shows some examples of these errors.

We adopt a strategy to clean the datasets automatically to mitigate the effects of such errors. We rely on the approach proposed in [41] that allows the re-assignment of the identity label for images initially deemed incorrectly labeled, minimizing the number of images discarded from the original datasets. Additionally, we manually identified and removed 841 duplicated (same cryptographic hash) images in the Adience dataset. The cleaned versions of the Adience and BFW datasets, hereafter referred to as *Adience clean* and *BFW clean*, are composed of 13,160 images from 355 identities, 19,131 images from 800 identities, respectively.[5]

To assess the effectiveness of the cleaning process, we observe the differences between the distribution of the genuine and impostors scores before and after cleaning the datasets. The distributions of genuine scores of both cleaned datasets present a typical uni-modal distribution (Figs. 2c and 2d), indicating that the automated cleaning process succeeded in determining the mislabeled images.

To evaluate if the cleaning procedure had changed the difficulty for face recognition of the datasets, we investigated the differences in the distribution of Confusion Scores of the reference and probe images before and after cleaning. As depicted in Fig. 4, the distributions of the quality scores (CS) before and after the cleaning process are highly similar, suggesting that the cleaning procedure did not change the intrinsic difficulty of the datasets.

3.2.2. *Definition of References for Adience clean and BFW clean Datasets*

The concept of the reference image is absent in the social media datasets. Based on the assumption that in forensic scenarios, the reference images are typically acquired in more controlled scenarios, we select the image with the highest face quality as the reference image from each identity.

[5] The list of files corresponding to the cleaned versions of the Adience and BFW datasets is available at https://github.com/rafribeiro/embedding_aggregation.

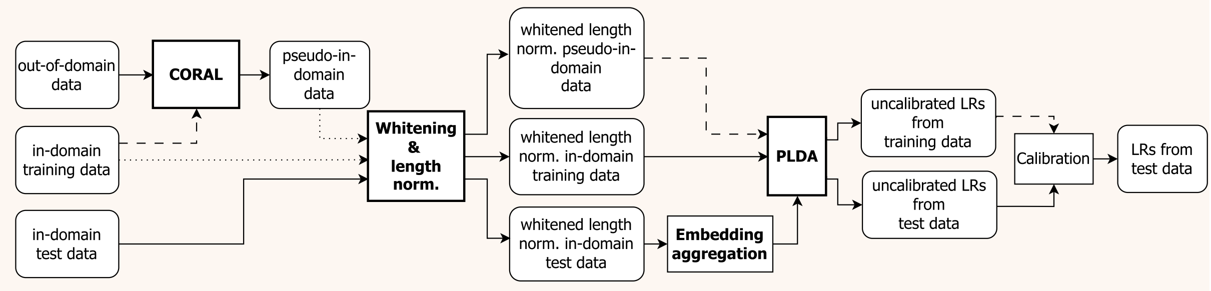

out-of-domain data
CORAL
pseudo-in-domain data
in-domain training data
in-domain test data
Whitening & length norm.
whitened length norm. pseudo-in-domain data
whitened length norm. in-domain training data
whitened length norm. in-domain test data
Embedding aggregation
PLDA
uncalibrated LRs from training data
uncalibrated LRs from test data
Calibration
LRs from test data

Fig. 6. Diagram of stages for processing the embeddings. In this diagram, the rectangles with rounded corners represent data, the rectangles with squared corners represent processes or models, the dashed-line arrows represent training data to the corresponding model, dotted-line arrows represent data that is being used for training the corresponding model and that will also be transformed by the trained model, and solid-line arrows represent data that will be transformed by a trained model. The models in bold compose the proposed scoring stage.

In particular, we rank the images according to their *Ser-Fiq* and Confusion Scores and select the image with the best combined ranking. Fig. 5 depicts the selected references for two identities of each of the Adience and BFW datasets, illustrating that this strategy resulted in the selection of good-quality reference images.

4. Proposed method

Our proposal is focused on obtaining embeddings that better represents the identity corresponding to the trace images. To accomplish this, we aggregate the information available in multiple facial images in the embedding space. We investigate multiple aggregation strategies, described in Section 4.2. We then use the aggregated embedding from the trace images to calculate a likelihood ratio for the comparison between the set of trace images and the reference image.

To obtain well-calibrated likelihood ratios, we adopt a pipeline inspired by forensic voice comparison systems. The stages of this pipeline are described in Section 4.3.

We note that this pipeline involves the calculation of a score obtained from a PLDA model. This score is an attempt to calculate a likelihood ratio and considers how similar the embeddings are and how typical they are in the relevant population. Because there is generally a small amount of data to train the PLDA model relative to the number of parameters that must be estimated, its output is frequently ill-calibrated. A calibration model is then applied to the uncalibrated likelihood ratios to obtain calibrated likelihood ratios.

4.1. Embedding extraction

We conduct all the experiments using a single face recognition model to extract embeddings from facial images. Before extracting the embeddings, we used a face detection model based on SCRFD [42] to locate the facial region in the image, which was then cropped and aligned using an affine transformation, resulting in an aligned facial image with 112x112 pixels. For embedding extraction, a ResNet-101 was trained on the MS1MV2 dataset [43], using the Arcface loss [43], achieving an accuracy of 99.83% on the LFW dataset [44], which is on par with performance reported by state-of-the-art face recognition models. The detection, alignment and embedding extraction procedures were implemented using the open-source InsightFace library[6].

4.2. Embedding aggregation

When multiple images of the same person are available, it is possible to perform multiple comparisons, raising the question of which comparison to derive the LR from or how to combine the multiple comparisons to obtain a single LR for the case. Multiplication of the LRs is not reasonable because the independence of the facial images of the same person could not be justifiably assumed. To address these issues, we propose to perform a single comparison in each simulated case. For this comparison, we use an aggregated embedding obtained from a linear combination of the embeddings from the trace images that can be considered to be of the same individual, as described in Eq. 2:

$$\mathbf{v}^t = \sum_{i=1}^{N} w_i \mathbf{v}^{t_i}, \tag{2}$$

where w_i is the weight assigned to image i in the aggregation process. Different strategies to assign this weight are described in the next sections.

We also consider the strategy of obtaining a single score by averaging the scores of all possible comparisons between a reference image and each trace image of the same individual.

4.2.1. Average Pooling

This embedding aggregation strategy involves assigning the same weight w_i for each trace image. This is essentially an unweighted average of each component of the embeddings.

The following strategies are based on the assumption that the facial image quality should be related to the weight assigned to each image. We use two quality estimation approaches as proxies for facial image quality[7].

4.2.2. Ser-Fiq Pooling

Ser-Fiq is a facial quality estimation method proposed in [45]. It is based on the robustness of the embeddings obtained from multiple configurations of the same embedding extractor network. Images that result in embeddings with more variation given the multiple configurations of the embedding extractor are considered to be of lower quality than images that result in embedding with less variations.

We used the normalized *Ser-Fiq* quality score s_i of each trace image as w_i:

$$w_i = \frac{s_i}{\sum_{j=1}^{N} s_j}. \tag{3}$$

4.2.3. CS Pooling

We also considered the recently proposed face quality estimator *Confusion Score* (CS) [19] as a weighting mechanism for aggregation. This method is based on the observation that the distribution of impostor

[6] Available at https://github.com/deepinsight/insightface.

[7] We refer to quality as a measure of the utility of the facial image for its use in face recognition systems, and not as a synonym to the human-based perception of image quality.

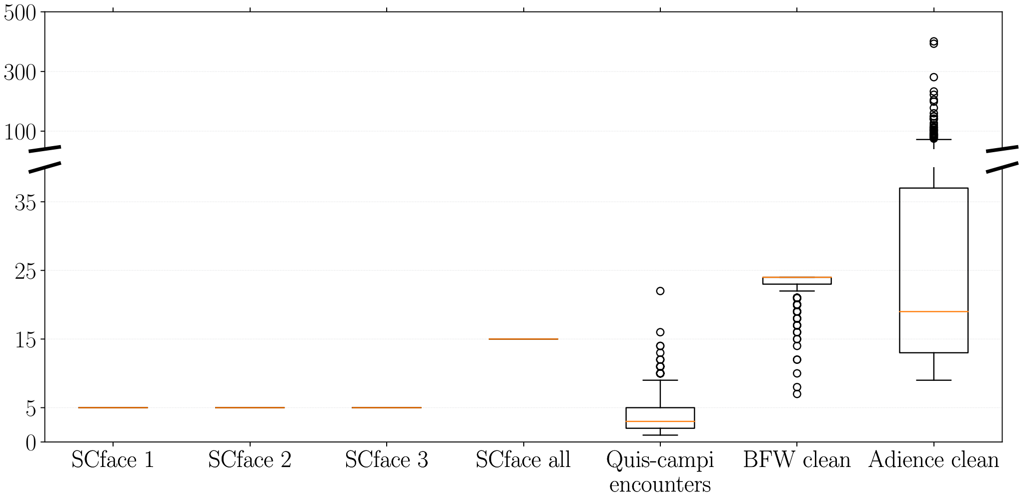


Fig. 7. Distribution of the number of embeddings aggregated per identity in each dataset.

scores from lower quality images extend to higher values than the distribution of impostor scores of higher quality images. The CS values are constrained between −1 and 1, with higher values indicating images with lower quality and lower values indicating images with better quality. To account for this inverse relation between the numerical value of CS and image quality, in this strategy, the weight w_i of a trace image with Confusion Score cs_i is computed according to Eq. 4:

$$w_i = \frac{1 - cs_i}{\sum_{j=1}^{N}(1 - cs_j)} \tag{4}$$

4.3. Similarity and typicality scores

We adopt scores that consider not only the similarity but also the typicality of the embeddings with respect to the relevant population, as recommended in [27]. To properly account for typicality, we computed scores using PLDA[8], following a similar data processing pipeline for the scoring stage to what is frequently used in forensic voice comparison systems.

The following subsections briefly describe each component of the scoring stage, visually depicted in Fig. 6.

4.3.1. CORAL

Correlation Alignment is an unsupervised domain adaptation technique proposed in [46]. It is frequently used in automatic speaker recognition systems to enable the use of larger amounts of data from different domain(s). CORAL is used specifically to align the total covariance matrix of source (out-of-domain) data to the total covariance matrix of target (in-domain) data. In our experiments, we used CORAL in the following manner: for the experiments with each of the two datasets in each scenario (see Section 3), we used the other dataset of the same scenario as out-of-domain data.

4.3.2. Whitening and length-normalization

Whitening and length-normalization of the embeddings are performed as the last step before the PLDA model. These two transformations are applied to better condition the embeddings to the assumption of gaussian distribution of the data for PLDA [47]. The whitening transformation is trained using the CORAL-transformed out-of-domain data and the training portion of the in-domain data for each experiment.

4.3.3. PLDA

The Probabilistic Linear Discriminant Analysis model used in these experiments was proposed in [29]. This model, as well as other variations of PLDA, have traditionally been used as a scoring method in automatic speaker recognition [48]. It is a generative model that can be used to compute common-source likelihood ratios. Because the amount of parameters to be estimated is large relative to the amount of data usually available to train the model, the likelihood ratios output by PLDA are generally not well calibrated. Therefore, it is common practice in the forensic voice comparison community to treat the output of PLDA as a score, and then apply a calibration procedure to obtain a well-calibrated likelihood ratio [48]. We follow the same approach in this paper.

4.3.4. Regularized logistic regression calibration

We use a regularized logistic regression model as the last stage to obtain a calibrated LR. The model is implemented in the open-source LIR Python package[9].

Logistic regression has traditionally been used as a calibration model in forensic speaker comparison [49–51]. In contrast with other common calibration models, it does not assume a specific distribution of the training scores, it is less susceptible to sampling variability than other calibration methods [20,22,52], and it also guarantees a monotonic conversion of scores into LRs.

Regularization is used to induce shrinkage of the LR values. Shrinking the likelihood ratio values is a conservative approach to avoid overstating the strength of evidence when only a small amount of data is available to train the system [53]. Since having small amounts of data in forensically relevant conditions is common in casework, we opted to include shrinkage[10] in our experiments so it better reflects the scenarios expected to be encountered in casework.

5. Experiments

We focus our experiments on 1:1 comparisons between a reference image and a set of trace images. In forensic settings, these sets of trace images may originate from a surveillance video, with multiple frames depicting the person of interest, or from a set of images collected from

[8] The implementation used in our experiments is based on the PLDA model proposed by S. Ioffe [29] which is available at https://github.com/RaviSoji/plda.

[9] Available in https://github.com/netherlandsforensicinstitute/lir

[10] We used the same regulation parameter of 1 in all experiments.

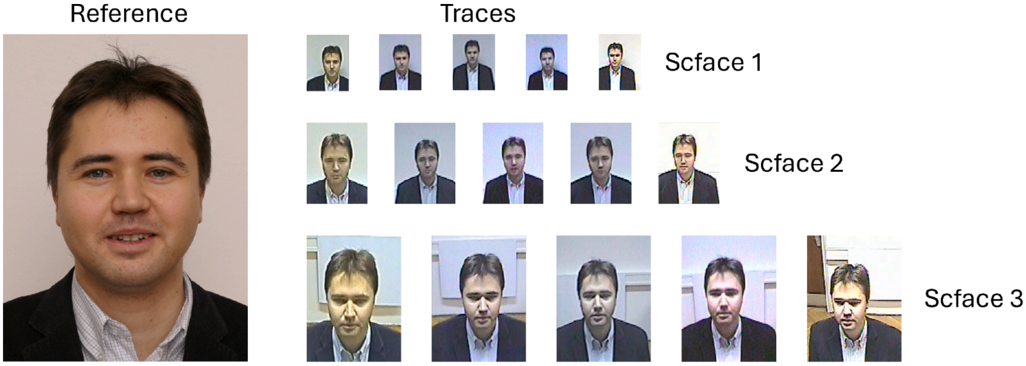


Fig. 8. Images used in the worked example. The scale of the trace images is correct relative to each distance (1, 2 and 3), but the reference image is of proportionally higher resolution than shown in the figure.

social media profiles.

As a baseline, we rely on the comparisons between each reference image and each trace image, without any form of aggregation. This baseline is representative of common practices in forensic laboratories, where a single trace image is evaluated independently against the reference without any form of aggregation.

We evaluate different strategies to integrate the information available in multiple images of the trace sets, as described in Section 4. The distribution of the number of embeddings that are aggregated per identity in each dataset is shown in Fig. 7.

For each aggregation strategy, we calibrate the scores into LRs using a regularized logistic regression model, described in Section 4.3.4. To avoid training and testing on the same data, and to avoid data leakage between training of the PLDA model and the calibration model, we adopted a nested 10-fold cross validation strategy to train these two models and to obtain validation LRs for each experiment.

We assess the performance of the resulting forensic evaluation systems using the log-likelihood ratio cost (C_{llr}) [16] and Tippett plots [54]. These criteria are detailed in the following section.

5.1. Performance assessment

We adopt the performance assessment methods described in a recently published consensus on validation of methods for forensic voice comparison [18]. These methods consist of a metric of accuracy of forensic evaluation systems - the log-likelihod ratio cost (C_{llr}) - and a graphical representation of the empirical performance of the system - Tippett plots. Both are described in the following sections.

5.1.1. Log-likelihood ratio cost - C_{llr}

The log-likelihood ratio cost is an accuracy metric considered appropriate for assessing the performance of forensic evaluation systems that output likelihood ratios [16–18]. It can be computed by the following formula:

$$C_{\text{llr}} = \frac{1}{2}\left[\frac{1}{N_{ss}}\sum_{i} log_2\left(1+\frac{1}{LR_i}\right) + \frac{1}{N_{ds}}\sum_{j} log_2(1+LR_j)\right], \tag{5}$$

where the first sum is over the N_{ss} individual test likelihood-ratios (LR_i) for which the ground truth is same-source, and the second sum is over the N_{ds} individual test likelihood ratios (LR_j) for which the ground truth is different-sources.

Contrary to common performance metrics for biometric systems like false acceptance rate, false rejection rate or equal error rate, C_{llr} does not depend on the application of a threshold, which is more aligned with how likelihood ratios should be interpreted. It also gives different penalties according to the magnitude of the log-likelihood ratio: for a validation pair for which the ground truth is different-sources, an LR much larger than 1 will be more heavily penalized than an LR not much larger than 1. For a validation pair for which the ground truth is same-source, an LR much smaller than 1 will be more heavily penalized than an LR not much smaller than 1.

Forensic evaluation systems with good performance will exhibit C_{llr} closer to zero, while a forensic evaluation system that gives no useful information (e.g., a system that always output an LR equal to 1) will have a C_{llr} value of 1. Well-calibrated systems are considered useful for forensic casework if they have C_{llr} lower than 1 [18].

5.1.2. Tippett plots

Tippett plots allow one to visually assess the level of calibration of the system, its discriminating power, as well as the range of LRs that the system can output [17,18]. It displays empirical cumulative distributions of the validation LRs. Two curves are displayed in the same figure, one corresponding to validation LRs where the ground truth is different-sources, and another corresponding to validation LRs for which the ground truth is same-source. The horizontal axis exhibits the range of LRs output by the system for the validation pairs, in base-10 logarithmic scale. The vertical axis displays the proportion of LRs that are greater than or equal to the value in the x-axis, for validation LRs obtained from different-sources pairs, and the proportion of LRs that are smaller than or equal to the value in the x-axis for same-source validation pairs. Ideally, the empirical cumulative distributions are symmetrically distributed around the neutral value of $log_{10}LR = 0$, and are far from each other in the horizontal direction. Appendix C of [18] shows examples of Tippett plots for systems with various degrees of calibration and discriminating power.

5.2. Worked example

To illustrate our experiments, a worked example is shown in this section, with images from the SCface dataset. We select images from the identity labeled as 001. There is one reference image, with high quality and captured under ideal conditions, and 15 images of the same person captured from CCTV cameras from 3 different distances, resulting in images of three distinct levels of quality, as shown in Fig. 8.

Our baseline involves comparing the reference image to each CCTV image independently, and calculate a likelihood ratio for each comparison. For the AvgPool experiment, we compute the component-wise average of the embeddings for each group of 5 images from each distance. We obtain one embedding representing the 5 images from the SCface 1, another embedding representing the 5 images from the SCface 2, and a third embedding representing the 5 images from SCface 3. We then compare the reference image to each of the three aggregated embeddings, obtaining 3 LR values. The results are shown in Table 1, which shows higher LR values for the aggregated cases than the individual comparisons. We also note that most of the individual comparisons resulted in LRs that support the wrong hypothesis, while all aggregated results resulted in LRs that support the correct hypothesis.

Table 1
ln(LR) values for individual and aggregated comparisons for the identity 001 of the SCface dataset. Since this is a same-source comparison, higher values are better.

	1	2	3	4	5	AvgPool
SCface 1	−1.9	−4.7	−3.1	−3.0	−1.8	**1.1**
Scface 2	1.5	−4.9	−3.1	−2.1	−3.7	**4.3**
Scface 3	1.1	−3.7	−0.6	2.1	−1.8	**7.5**

Table 2
C_{llr} for the surveillance scenario.

	SCface 1	SCface 2	SCface 3	SCface	Quis-Campi	Quis-Campi encounters
Mandasari et al. [32] Raw scores	0.659	0.313	0.378	0.503	-	-
Mandasari et al. [32] ZT-norm scores	0.664	0.243	0.287	0.419	-	-
Baseline	0.545	0.143	0.043	0.217	0.237	0.237
AvgScore	0.328	0.081	0.093	0.002	**0.107**	0.195
AvgPool	**0.327**	0.069	0.081	**0.0005**	0.138	0.184
CSPool	0.576	0.121	0.083	0.009	0.124	0.182
Ser-FiqPool	0.357	**0.056**	**0.37**	0.001	0.124	**0.178**

6. Results and discussion

Results for the datasets of the surveillance scenario are shown in Table 2 and Fig. 9.

We first observe the improvements in C_{llr} compared to Mandasari et al. [32] on the SCface dataset. This improvement is mainly attributed to the discriminating power of the facial recognition module since even our baseline approach offered substantially better results. We also observe that both embedding aggregation and score averaging were effective approaches to improve the performance of the forensic evaluation system with respect to our baseline. We also observe that larger improvements in C_{llr} occurred for the surveillance scenario datasets with more embeddings to be aggregated (SCface - 15 embeddings), and with images of lower quality (SCface 1 and SCface 2).

An unexpected result was observed when comparing the C_{llr} values obtained for SCface and SCface 3. Because SCface includes images of poor, medium and higher quality, we expected that the results of using only better quality images (SCface 3) would be better. The opposite happened, which we interpret as an indication that the aggregation strategies may incorporate useful information from lower-quality images, improving upon the result obtained from better-quality images alone.

Results for the social media scenario are shown in Table 3 and Fig. 10. We also observe gains from the proposed aggregation strategies compared to the baselines.

In general, we observe that aggregating embeddings from multiple images of the same individual is an effective technique for improving recognition performance. The improvements are more pronounced when dealing with multiple low-resolution images, when no single

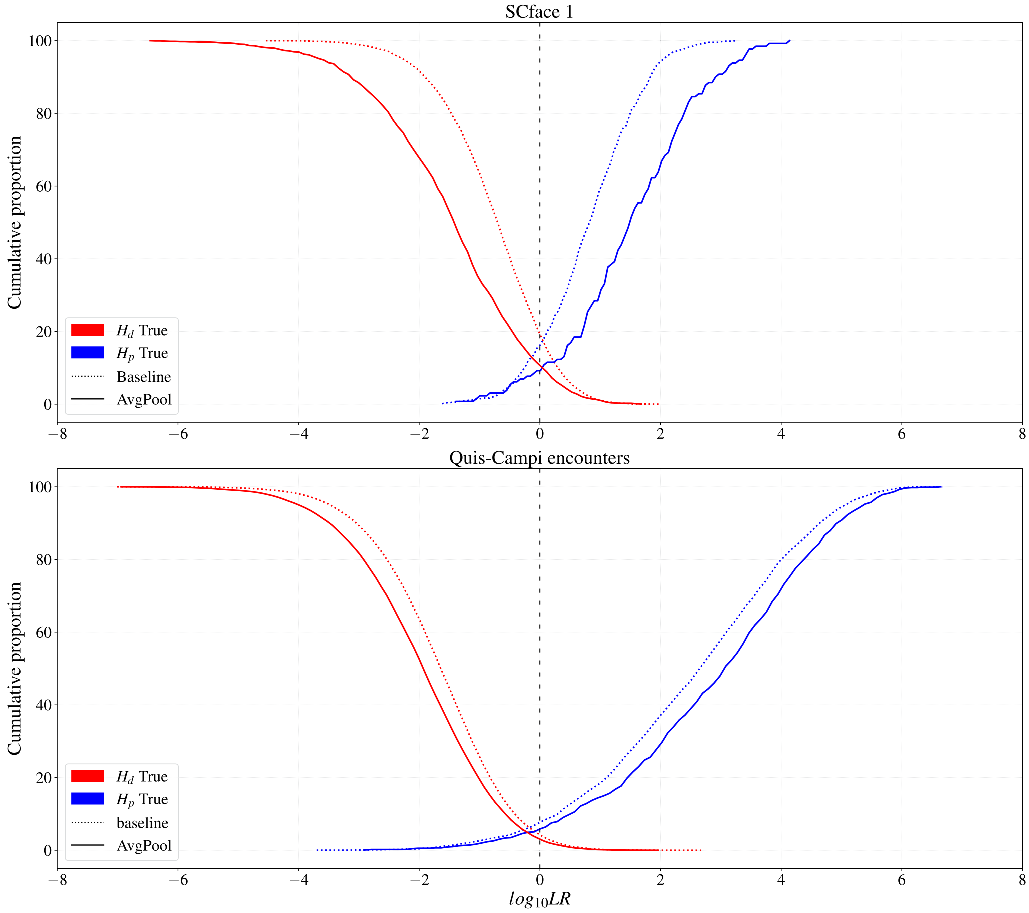


Fig. 9. Tippett plots for the SCface 1 and Quis-Campi encounters datasets. To avoid cluttering the figure, we only show Tippett plots for the baseline and AvgPool strategies. Tippett plots for all strategies are available in the supplementary material.

Table 3
C_{llr} for the social media scenario.

	Adience clean	BFW clean
Baseline	0.035	0.081
AvgScore	0.010	0.037
AvgPool	**0.007**	0.027
CSPool	0.009	0.031
Ser-FiqPool	0.008	**0.023**

image is of good enough quality to provide good probative value. We also note that the number of images that could have their embeddings aggregated is an important factor and warrants further investigation. These two observations are especially relevant for realistic forensic conditions, where CCTV video with a large number of low quality frames is usually available, such as the recordings of the invasions of the Brazilian Federal Government buildings in January 2023 [55]. Even “naïve” approaches such as *AvgPool* can offer substantial performance improvements relative to considering a single trace image. This strategy also has the advantage of not requiring the estimation of facial image quality.

The approaches based on quality-weighted aggregation - Ser-FiqPool and CSPool - also demonstrate good performance and, for some combinations of dataset and quality metric, provided the best performance. In some combinations, though, quality-based aggregation was detrimental to performance (e.g., CSPool on SCface1). This suggests a complex interplay between the quality metrics and the face recognition model used to extract the embeddings and may warrant further investigation.

We note two important limitations to the present study.

Firstly, the available datasets are of limited representativity to real forensic casework, with just a few frames from CCTV footage, in the case of SCface and Quis-Campi, and with no images representative of standard reference images, with controlled pose, illumination and facial expression in the case of Adience and BFW. We also note that, in the case of the social media datasets, there are multiple images from the same session for various identities, making it more difficult to generalize the results of our experiments with these datasets for casework.

The second limitation relates to the amount of data (both the number of identities and the number of images per identity) to train the PLDA model, especially in the experiments with CCTV images. As a complex multivariate model, the limited amount of data available to train the PLDA model probably represents a major limitation to better performance.

We aim to address these limitations in future work by collecting new data and assessing the aggregation approaches on images from CCTV videos that are more representative of casework conditions.

7. Conclusions

We presented aggregation strategies to improve face recognition performance under challenging conditions usually found in forensic casework when multiple images of the same person are available. Our results indicate an effective approach for dealing with low-quality images frequently found in forensic casework. Future work will employ data that are more representative of forensic casework, including the amount needed to train the statistical models.

We also presented an initial exploration for calculating common-source likelihood ratios from DNN embeddings extracted from facial images using scores that consider both the similarity and the typicality of the embeddings. Future works will deepen the investigation of

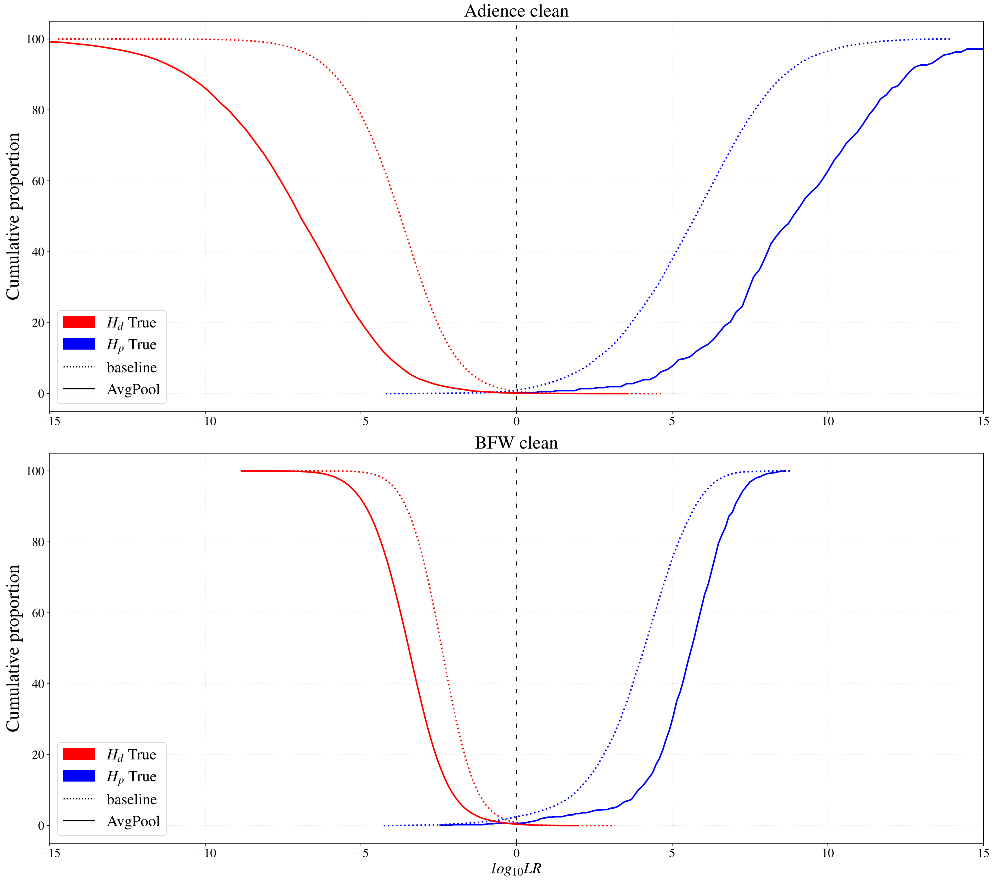


Fig. 10. Tippett plots for the Adience clean and BFW clean datasets. To avoid cluttering the figure, we only show Tippett plots for the baseline and AvgPool strategies. Tippett plots for all strategies are available in the supplementary material.

calculating common-source likelihood ratios for facial comparisons.

Funding source: This research was partially supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the contributions of Paulo Max Gil Innocencio Reis, Gustavo Henrique Machado de Arruda and Janine Zancanaro da Silva for the fruitful discussions regarding the experiments. The authors also acknowledge the discussions with Dr. Geoffrey S. Morrison and Dr. Phil Weber on the statistical models employed in the field of forensic voice comparison.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scijus.2024.07.006.

References

[1] M.P.J. Ashby, The value of CCTV surveillance cameras as an investigative tool: an empirical analysis, Eur. J. Crim. Policy Res. 23 (3) (2017) 441–459, https://doi.org/10.1007/s10610-017-9341-6.

[2] 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 Biometrics 6 (6) (2017) 487–494.

[3] N. Bacci, T.M.R. Houlton, N. Briers, M. Steyn, Validation of forensic facial comparison by morphological analysis in photographic and CCTV samples, Int. J. Legal Med. 135 (5) (2021) 1965–1981, https://doi.org/10.1007/s00414-021-02512-3.

[4] F.I.S.W.G. (FISWG), Facial comparison overview and methodology guidelines (2019).

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

[6] ENFSI, Enfsi-bpm-di-01 - best practice manual for facial image comparison (01 2018). https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf.

[7] S. Willis, A. Ligertwood, J. Molina, C. Berger, G. Zadora, A. Nordgaard, B. Rasmusson, L. Lunt, C. Champod, A. Biedermann, T. Hicks, F. Taroni, X. Zhu, Enfsi guideline for evaluative reporting in forensic science (03 2015). https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

[8] P.J. Phillips, A.N. Yates, Y. Hu, C.A. Hahn, E. Noyes, K. Jackson, J.G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen, C.D. Castillo, R. Chellappa, D. White, A.J. O'Toole, Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms, Proceedings of the National Academy of Sciences 115 (24) (2018) 6171–6176. doi:10.1073/pnas.1721355115. doi: 10.1073/pnas.1721355115.

[9] C.A. Hahn, L.L. Tang, A.N. Yates, P.J. Phillips, Forensic facial examiners versus super-recognizers: Evaluating behavior beyond accuracy, Appl. Cognit. Psychol. 36 (6) (2022) 1209–1218, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.4003, doi: 10.1002/acp.4003. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.4003.

[10] M.J. Saks, J.J. Koehler, The coming paradigm shift in forensic identification science, Science 309 (5736) (2005) 892–895, https://doi.org/10.1126/science.1111565.

[11] N.R. Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, DC, 2009, https://doi.org/10.17226/12589. URL https://nap.nationalacademies.org/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward.

[12] E.S. Lander, P.W. Group, others, Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods, Tech. rep., President's Council of Advisors on Science and Technology (US), publisher: President's Council of Advisors on Science and Technology (US) (2016).

[13] G.S. Morrison, Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science, Forensic Sci. Int.: Synergy 5 (2022) 100270, https://doi.org/10.1016/j.fsisyn.2022.100270.

[14] D. Meuwly, Forensic individualisation from biometric data, Sci. Just. 46 (4) (2006) 205–213.

[15] C. Neumann, J. Hendricks, M. Ausdemore, Statistical support for conclusions in fingerprint examinations, in: Handbook of Forensic Statistics, Chapman and Hall/CRC, 2020, pp. 277–324.

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

[17] 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.

[18] G.S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Science & Justice 61 (3) (2021) 299–309.

[19] A. Ruifrok, P. Vergeer, A.M. Rodrigues, From facial images of different quality to score based lr, Forensic Sci. Int. (2022) 111201.

[20] T. Ali, Biometric score calibration for forensic face recognition, University of Twente, 2014. Jun. Ph.D. thesis.

[21] M. Jacquet, C. Champod, Automated face recognition in forensic science: Review and perspectives, Forensic Sci. Int. 307 (2020) 110–124. URL https://www.sciencedirect.com/science/article/pii/S0379073819305365.

[22] A.L. Mölder, I. Enlund Åström, E. Leitet, Development of a score-to-likelihood ratio model for facial recognition using authentic criminalistic data, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), 2020, pp. 1–6. doi: 10.1109/IWBF49977.2020.9107954.

[23] P. Grother, M. Ngan, K. Hanaoka, Face Recognition Vendor Test (FRVT) part 2: identification Draft Supplement, National Institute of Standards and Technology, Gaithersburg, MD, Dec. 2022. Tech. Rep. NIST IR 8271 Draft Supplement.

[24] T. d. F. Pereira, D. Schmidli, Y. Linghu, X. Zhang, S. Marcel, M. Günther, Eight years of face recognition research: Reproducibility, achievements and open issues (2022). doi:10.48550/ARXIV.2208.04040. https://arxiv.org/abs/2208.04040.

[25] P.J. Phillips, A.J. O'toole, Comparison of human and computer performance across face recognition experiments, Image Vis. Comput. 32 (1) (2014) 74–85.

[26] N.I. of Forensic Science Australia New Zealand, An introductory guide to evaluative reporting (06 2017). https://www.anzpaa.org.au/ArticleDocuments/220/An%20Introductory%20Guide%20to%20Evaluative%20Reporting.PDF.aspx.

[27] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality, Science & Justice 58 (1) (2018) 47–58, https://doi.org/10.1016/j.scijus.2017.06.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1355030617300849.

[28] C. Neumann, M. Ausdemore, Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios', Law, Probability and Risk 19 (1) (2020) 21–42, https://doi.org/10.1093/lpr/mgaa006, arXiv:https://academic.oup.com/lpr/article-pdf/19/1/21/33390883/mgaa006.pdf.

[29] S. Ioffe, Probabilistic Linear Discriminant Analysis, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 531–542.

[30] T. Ali, L. Spreeuwers, R. Veldhuis, D. Meuwly, Effect of calibration data on forensic likelihood ratio from a face recognition system, in: in: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE, 2013, pp. 1–8.

[31] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, J. Ortega-Garcia, Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems, Forensic Sci. Int. 155 (2) (2005) 126–140.

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

[33] C. Zeinstra, R. Veldhuis, L. Spreeuwers, Grid-based likelihood ratio classifiers for the comparison of facial marks, IEEE Trans. Inf. Forensics Secur. 13 (1) (2018) 253–264.

[34] A.L. Mölder, I. Enlund Åström, E. Leitet, Development of a score-to-likelihood ratio model for facial recognition using authentic criminalistic data, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), 2020, pp. 1–6.

[35] M. Jacquet, Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal, Thèse de Doctorat, Université de Lausanne, Faculté de droit, des sciences criminelles et d'administration publique, Lausanne (Dec. 2021).

[36] I.D. Raji, G. Fried, About face: A survey of facial recognition evaluation (2021). arXiv:2102.00813.

[37] J. Neves, J. Moreno, H. Proença, Quis-campi: an annotated multi-biometrics data feed from surveillance scenarios, IET Biometrics 7 (2017), 10.

[38] M. Grgic, K. Delac, S. Grgic, Scface — surveillance cameras face database, Multimedia Tools Appl. 51 (3) (2011) 863–879.

[39] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2170–2179.

[40] J.P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, S. Timoner, Face recognition: Too bias, or not too bias? IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2020 (2020) 1–10.

[41] C. Jin, R. Jin, K. Chen, Y. Dou, A community detection approach to cleaning extremely large face database, Comput. Intell. Neurosci. 2018 (2018) 1–10, https://doi.org/10.1155/2018/4512473.

[42] J. Guo, J. Deng, A. Lattas, S. Zafeiriou, Sample and computation redistribution for efficient face detection, CoRR abs/2105.04714 (2021) arXiv:2105.04714. URL https://arxiv.org/abs/2105.04714.

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

[44] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, University of Massachusetts, Amherst, 2007. October. Tech. Rep. 07–49.

[45] P. Terhörst, J.N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020 (2020) 5650–5659.

[46] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation. in: AAAI, 2016.

[47] D. Garcia-Romero, C.Y. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: Interspeech 2011, interspeech 2011, ISCA, 2011. doi:10.21437/interspeech.2011-53. https://doi.org/10.21437/Interspeech.2011-53.

[48] G.S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, A. Lozano-Díez, Statistical Models in Forensic Voice Comparison (ch. 20), in: D. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, 1st Edition, Chapman and Hall/CRC, New York, NY, USA, 2020, p. 47. doi: 10.1201/9780367527709.

[49] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio, Speech, Lang. Process. 15 (7) (2007) 2104–2115, https://doi.org/10.1109/TASL.2007.902747.

[50] S. Pigeon, P. Druyts, P. Verlinde, Applying logistic regression to the fusion of the nist'99 1-speaker submissions, Digital Signal Processing 10 (1) (2000) 237–248, https://doi.org/10.1006/dspr.1999.0358. URL https://www.sciencedirect.com/science/article/pii/S1051200499903585.

[51] G.S. Morrison, Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio, Austral. J. Forens. Sci. 45 (2) (2013) 173–197.

[52] B.X. Wang, V. Hughes, System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison, in: Proc. Interspeech 2021, 2021, pp. 381–385. doi:10.21437/Interspeech.2021-267.

[53] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors, Sci. Just. 58 (3) (2018) 200–218, https://doi.org/10.1016/j.scijus.2017.12.005, https://linkinghub.elsevier.com/retrieve/pii/S1355030617301582.

[54] C. Tippett, V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones, M. S. Lampert, The evidential value of the comparison of paint flakes from sources other than vehicles, J. Forensic Sci. Soc. 8 (2) (1968) 61–65.

[55] C.B. Oliveira, J.C. Neves, R.O. Ribeiro, D. Menotti, People tracking methods applied to planalto palace security videos, in: Anais Estendidos da XXXVI Conference on Graphics, Patterns and Images (SIBRAPI Estendido 2023), SIBRAPI Estendido 2023, Sociedade Brasileira de Computação - SBC, 2023. doi:10.5753/sibgrapi.est.2023.27462. doi: 10.5753/sibgrapi.est.2023.27462.