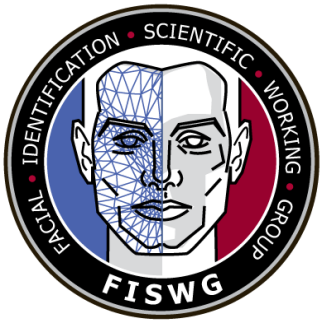


Disclaimer:

As a condition to the use of this document and the information contained herein, the Facial Identification Scientific Working Group (FISWG) requests notification by e-mail before or contemporaneously to the introduction of this document, or any portion thereof, as a marked exhibit offered for or moved into evidence in any judicial, administrative, legislative, or adjudicatory hearing or other proceeding (including discovery proceedings) in the United States or any foreign country. Such notification shall include: 1) the formal name of the proceeding, including docket number or similar identifier; 2) the name and location of the body conducting the hearing or proceeding; and 3) the name, mailing address (if available) and contact information of the party offering or moving the document into evidence. Subsequent to the use of this document in a formal proceeding, it is requested that FISWG be notified as to its use and the outcome of the proceeding. Notifications should be sent to: chair@fiswg.org

Redistribution Policy:

FISWG grants permission for redistribution and use of all publicly posted documents created by FISWG, provided that the following conditions are met:

Redistributions of documents, or parts of documents, must retain the FISWG cover page containing the disclaimer.

Neither the name of FISWG, nor the names of its contributors, may be used to endorse or promote products derived from its documents.

Any reference or quote from a FISWG document must include the version number (or creation date) of the document and mention if the document is in a draft status.

Face Recognition Systems Operation Assurance: Identity Ground Truth

Purpose

This document provides guidelines and techniques to help administrators of automated face recognition systems (FRS) produce advanced and accurate recognition statistics on the face recognition systems.

The intended audience of this document is system owners, system users, and system administrators of existing automated face recognition systems. Outside the scope of this document include, but not necessarily limited to, system setup, system tuning, workflow management and improvement, and proof of concept pilots.

This document is a follow on from the FISWG document: "Understanding and Testing for Face Recognition Systems Operation Assurance" (version 1.0, 2020.12.11)

The issues presented in this document form a base for other considerations and advanced topics when testing (e.g., system setup and tuning) which will be covered in future FISWG documents.

1. Scope

1.1 The scope of this document is to provide a detailed process and examples of testing and repairing identity ground truth in facial data sets which is a critical initial step before recognition statistics are created and reviewed. This document does not address facial accuracy but is focused solely on testing and correcting identity ground truth. Any facial algorithm can be used with these processes. It is assumed that all facial images create proper templates. The facial data set used in this document is the "Labeled Faces in the Wild" (LFW) but conceptually any other facial data set with identity ground truth can be used.

2. Referenced Documents

2.1 American National Standards Institute/National Institute of Science and Technology-Information Technology Laboratory Standard (ANSI/NIST-ITL Standard): http://www.nist.gov/itl/iad/ig/ansi_standard.cfm

2.2 Chi Jin, Ruochun Jin, Kai Chen, Yong Dou, "A Community Detection Approach to Cleaning Extremely Large Face Database", Computational Intelligence and Neuroscience, vol. 2018, Article ID 4512473, 10 pages, 2018. https://doi.org/10.1155/2018/4512473

2.3 Gallo, Ignazio & Nawaz, Shah & Calefati, Alessandro & Piccoli, Gabriele & Zamberletti, Alessandro. (2018). A Pipeline to Improve Face Recognition Datasets and Applications. https://doi.org/10.1109/IVCNZ.2018.8634724

2.4 Leys, C., et al., Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, Journal of Experimental Social Psychology (2013), http://dx.doi.org/10.1016/j.jesp.2013.03.013

2.5 NIST/SEMATECH e-Handbook of Statistical Methods, https://doi.org/10.18434/M32189

2.6 P. Grother, M. Ngan, K. Hanaoka "NISTIR 8271 DRAFT SUPPLEMENT Face Recognition Vendor Test (FRVT) Part 2: Identification" https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf

2.7 V. Varkarakis and P. Corcoran, "Dataset Cleaning — A Cross Validation Methodology for Large Facial Datasets using Face Recognition," 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), Athlone, Ireland, 2020, pp. 1-6, https://doi.org/10.1109/QoMEX48832.2020.9123123

2.8 Yager, Neil and Ted Dunstone. "The Biometric Menagerie." IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010): 220-230. https://ieeexplore.ieee.org/document/4711054

3. Terminology

3.1 *Definitions*

3.1.1 *doppelganger, n*–an apparition or double of a living person.

3.1.2 *false accept rate, n–see definition false match rate definition.*

3.1.3 *false match rate, n–the proportion of the completed biometric non-mated comparison trials that result in a false match. This is also referred to as false acceptance rate and does not include errors from images which do not create valid templates.*

3.1.4 *false non-match rate. n–the proportion of the completed biometric mated comparison trials that result in a false non-match. This is also referred to as false reject rate and does not include errors from images which do not create valid templates.*

3.2 *Acronyms*

3.2.1 *DET–Detection error tradeoff*

3.2.2 *FAR–False acceptance rate*

3.2.3 *FMR–False match rate*

3.2.4 *FNMR–False non-match rate*

3.2.5 *FR–Facial recognition*

3.2.6 *FRR–False reject rate*

3.2.7 *FRS–Facial recognition system*

3.2.8 *ROC–Receiver operating characteristics*

4. Background

4.1 When doing accuracy profiling, there is always one key aspect which must be addressed first: what is the identity ground truth within the images? All data sets will potentially have some corruption in the identity ground truth with the data. Detecting and correcting this so pristine results can be reviewed is always a critical portion of profiling.

4.2 This type of identity ground truth verification will potentially exist in all data sets no matter where the data sets come from. This is an iterative process as the agency learns the algorithms, the data, and how the two interact with each other. If proper care is not given in these early stages, then incorrect assumptions on the outcomes will be made. It's critical to understand this process with an investigative mindset before the agency gets to the operational data sets which may have identity corruption and image quality issues that may be large but not uncommon in operational deployments. If the agency gets to the operational data set without a firm awareness and knowledge base on the how the core algorithms work with verified data, then the agency could be incorrectly assessing and measuring performance of the FRS. Agencies need to lay the groundwork to know and trust the algorithms before they get to the possibly unmanaged and unknown operational data.

4.3 Most of the work in these processes is on creating the testing frameworks and understanding how to repeatedly run tests, make corrections, and do retesting with what has been learned. Once the frameworks and the processing are understood, then the agency can make diligent progress, but it takes time and focus. The outcomes are worth the time spent as you begin to understand how the data interacts with the algorithms which give the agency the ability to trust the solution and not just assume the data is invalid.

4.4 Setting up frameworks to do enrollment and searching while recording results is fairly mechanical as you learn the facial algorithms and the data sets to develop proper profiling. Understanding the data and building frameworks to analytically qualify the results is not trivial but must be done so effective operational metrics can be derived and applied.

4.5 Before doing this analysis on operational data, it is recommended that the agency develop and test the framework on experimental datasets. After some experience is gained in this process and confidence that the process is correct, one could then assess the operational dataset.

4.6 This document describes procedures to assess an experimental dataset, which can be replicated before assessing operational datasets.

5. Data Set

5.1 Care should be taken in selecting data sets to profile experimentally. It is recommended to select data sets which:

5.1.1 Have operational relevancy;

5.1.2 Have consistent image quality aspects: type of capture, size of images, subject poses, etc.;

5.1.3 Have sufficient identities and images to test with - this decision will be agency specific; and

5.1.4 Includes associated identity ground truth information which links each image to a unique identity.

5.2 The data set used for this document is the LFW data set available at: http://vis-www.cs.umass.edu/lfw/ See Appendix 2: "**LFW Data Set Information**" for more details. Conceptually any other facial data set with identity ground truth can be used.

5.2.1 LFW is a widely used open source data set which will work well for this specific document serving as an introductory data set. Information on this data set includes:

5.2.1.1 Has smaller but consistent image sizes and file formats;

5.2.1.2 Has over 5,700 identities and over 13,000 images;

5.2.1.3 Has a wide range of subjects: sex, pose, lighting, etc.; and

5.2.1.4 Has stated identity ground truth errors.

5.3 A key point in the LFW errata is that there are known errors in the LFW data set. While the LFW URL addresses these errors, this document will show how to locate them and give examples on how determining identity ground truths in relating images to identities is critical and needs to be addressed in any operational testing scenario.

6. Ground Truth Process

6.1 **Step 1:** Enroll the facial images into a facial gallery for searching.

6.2 **Step 2:** Search the facial images against the facial gallery. The number of candidates returned for this document was 100. This number may vary with agency specifics and the biometric algorithm deployed. It is recommended to test with a larger number of candidates than what may be operationally used so that deeper accuracy investigations can be analyzed. Do not use scoring thresholds.

6.3 **Step 3:** Analyze the scoring to delineate every candidate in all 1:N searches:

6.3.1 Image file name;

6.3.2 Image identity;

6.3.3 Score;

6.3.4 Rank;

6.3.5 Mate scoring; and

6.3.6 Imposter scoring.

6.3.7 When this was done for this document the following accuracy curves were obtained:

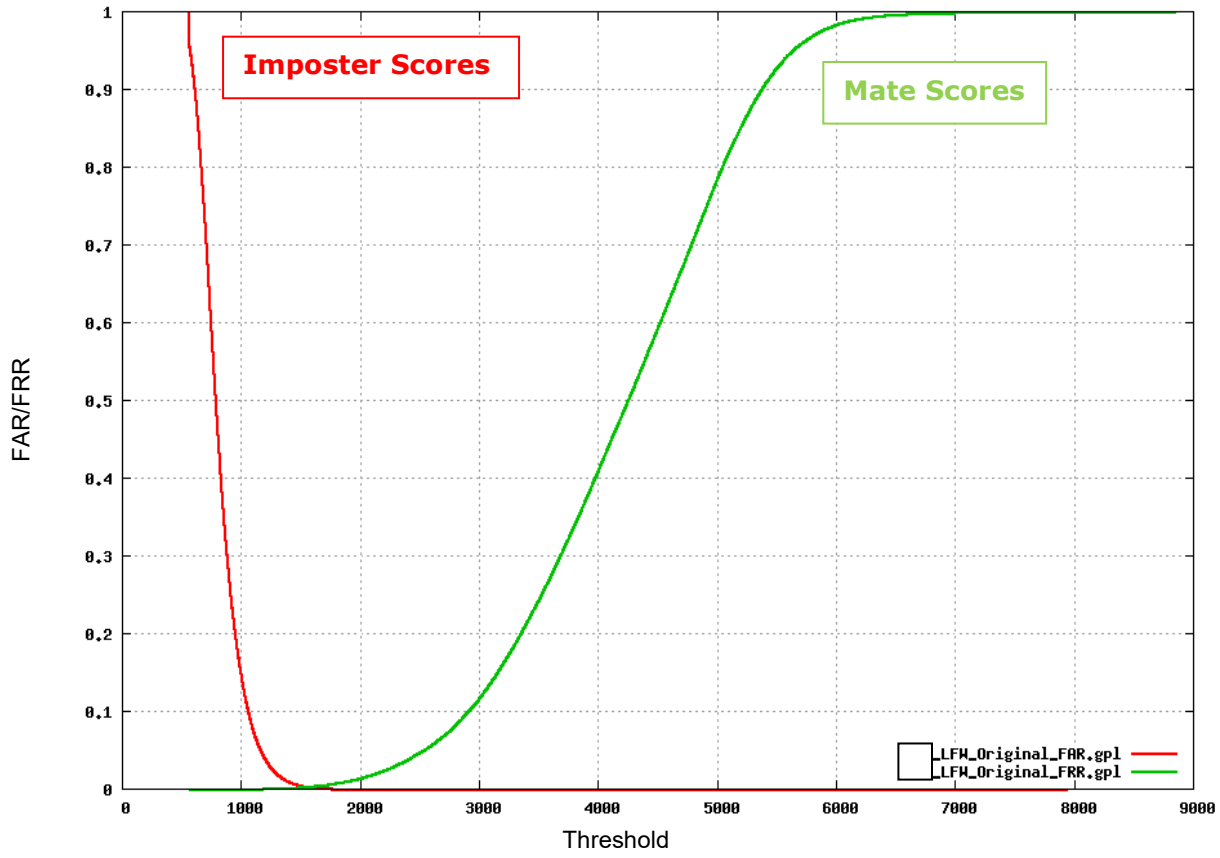


FIG. 1 FAR and FRR from uncorrected imagery

6.3.7.1 In Figure 1 the FAR and FRR scores are presented on a linear Y-axis. This can tend to hide the identity errors.

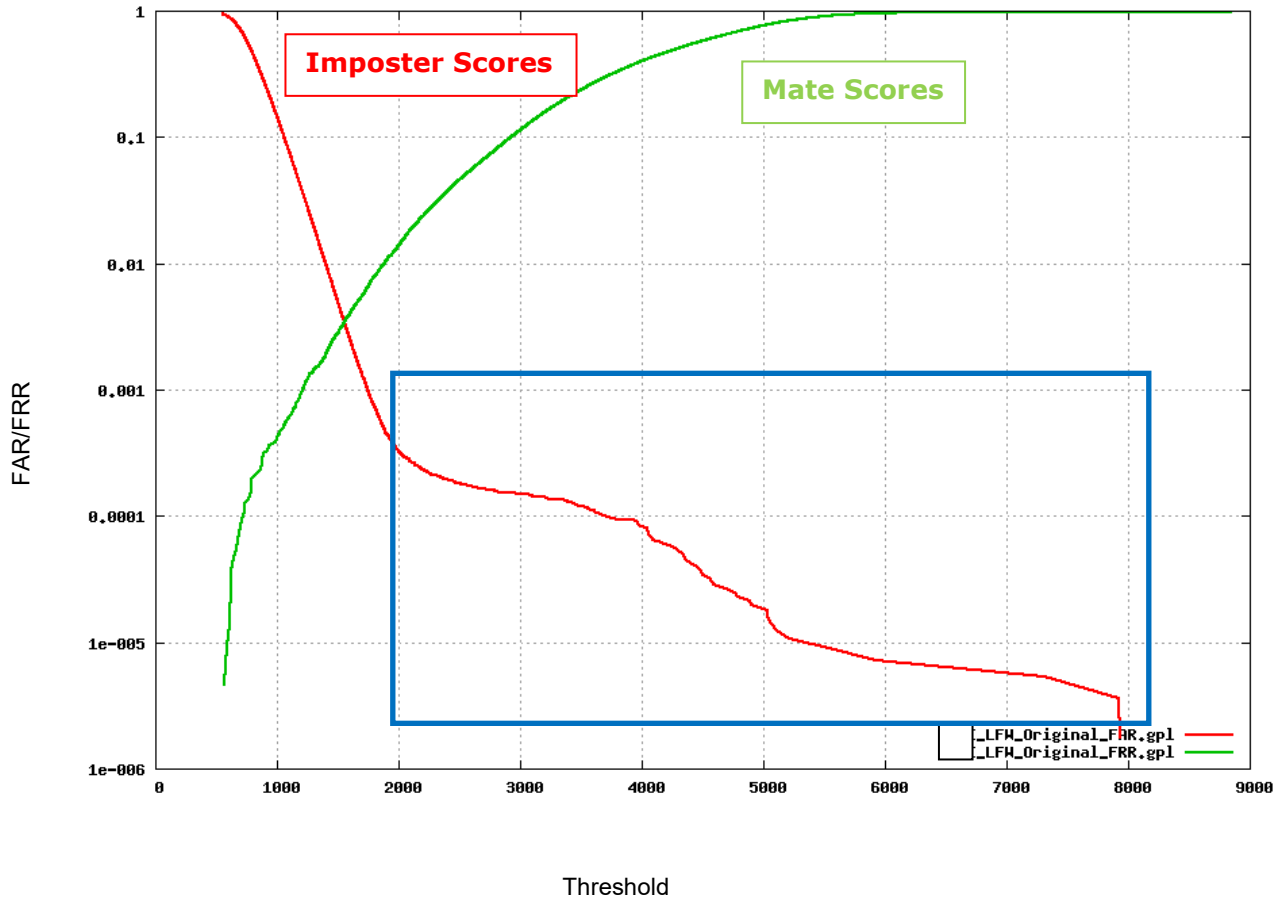


FIG. 2 FAR and FRR from uncorrected imagery

6.3.7.2 In Figure 2 the FAR and FRR scores are presented on a logarithmic Y-axis. This shows that there may be identity errors in the high scoring FAR scores (blue area). The sudden increase in FAR scores indicates potential mates that are incorrectly labeled imposters.

6.3.8 Other methods to select incorrectly labeled imposters are publicly available. See Appendix 1: "**Alternative Methods**".

6.4 **Step 4:** Analyze the scoring to resolve high scoring imposters to see if identity errors do exist in the data. These steps need to be done through manual reviews of the high scoring imposter pairs. If errors are located corrections to the identity image sets need to be done to manually or potentially remove from the data set.

6.5 **Step 5:** Iterate between Steps 1-4 as long as identity errors are suspected. The expectation is that several passes will be needed to achieve an objective measure of correct ground truth.

NOTE: In the LFW data set several iterations were done and approximately 90 identities were modified. After these corrections were made the following accuracy plots were derived.

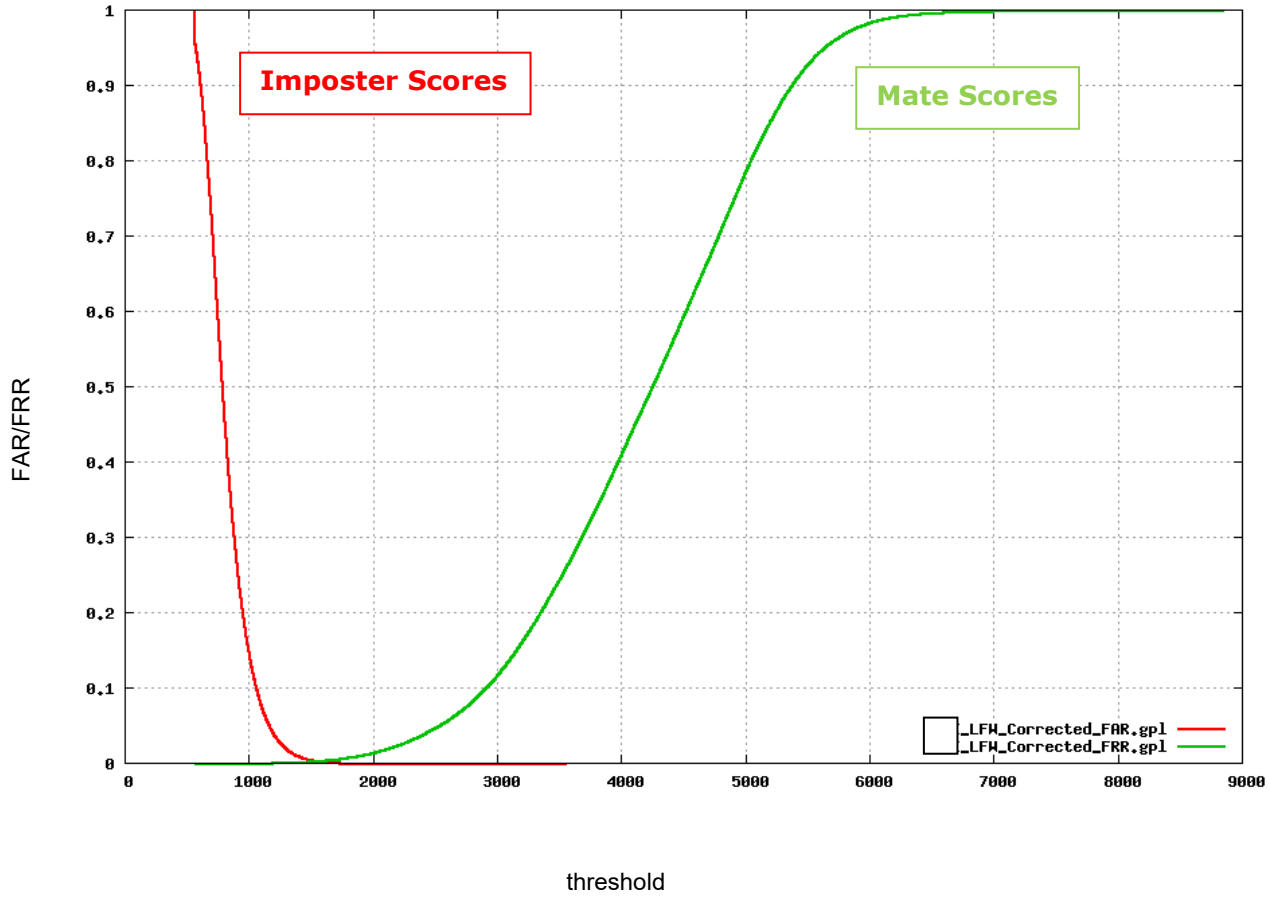


FIG. 3 FAR and FRR from corrected imagery

In Figure 3 the FAR and FRR scores are presented on a linear Y-axis. This can tend to hide the identity errors.

FIG. 4 FAR and FRR from uncorrected imagery

In Figure 4 the FAR and FRR scores are presented on a logarithmic Y-axis. This shows there may be several more identity errors in the FAR scoring (blue area).

One can find the following images by investigating the high score imposters.

Twins	
	
	


The images above are siblings


The images above are close appearances


The images above are close appearances

As these high scoring imposters were investigated, it became apparent that the high scores in the corrected FAR were heavily influenced by:

- Twins
- Siblings
- What are referred to as “doppelgangers”

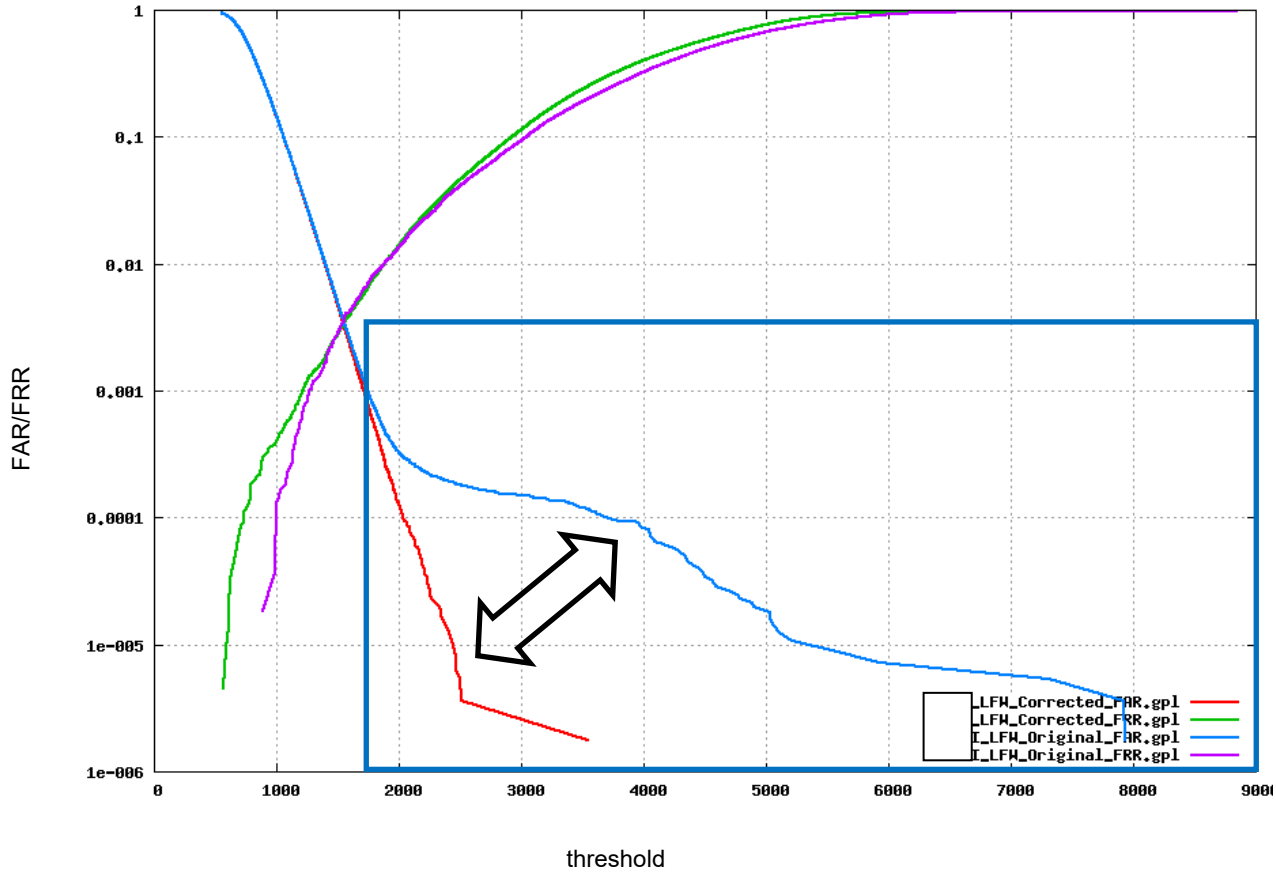


FIG. 5 FAR and FRR from original and corrected LFW imagery

In Figure 5 the FAR and FRR from original vs. corrected imagery are presented on a logarithmic Y-axis. This shows the scoring variances after the identity errors were corrected (blue area). This shows how a few improper identities can dramatically affect the FAR and FRR scoring profiles.

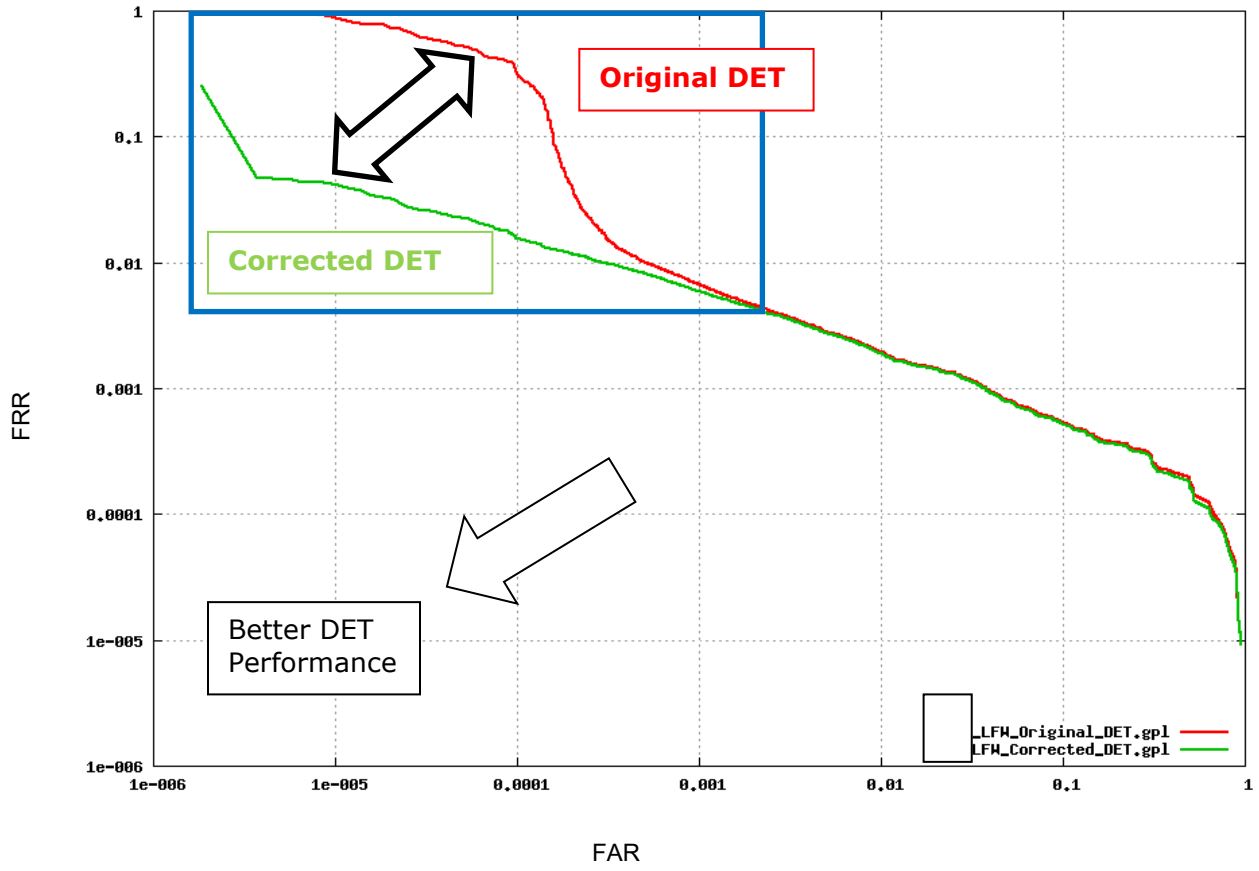


FIG 6 DET comparisons between original and corrected LFW imagery

In Figure 6 the DET curve from original vs. corrected imagery are presented. This shows the scoring variances after the identity errors were corrected (blue area). This again shows how a few improper identities can dramatically affect the FAR and FRR scoring profiles.

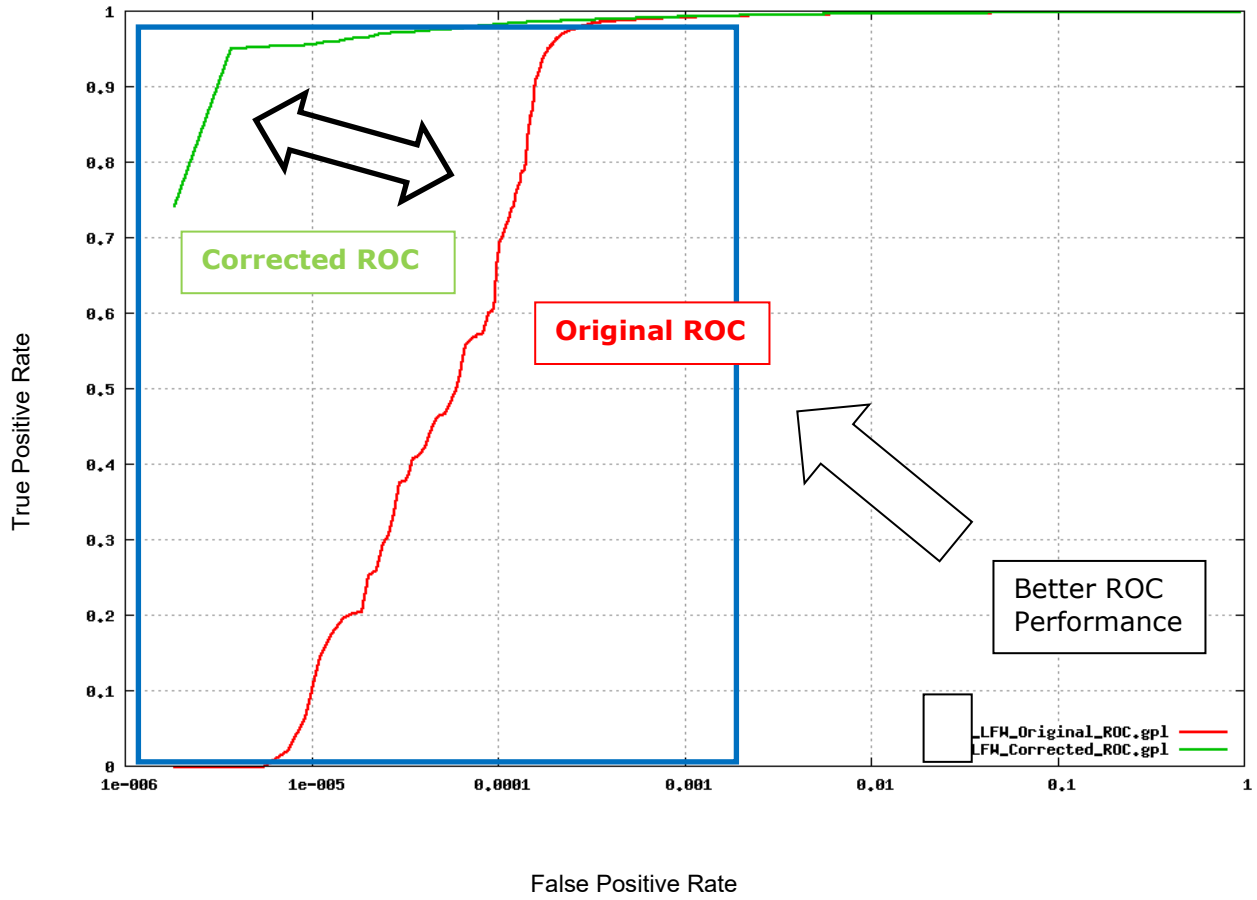
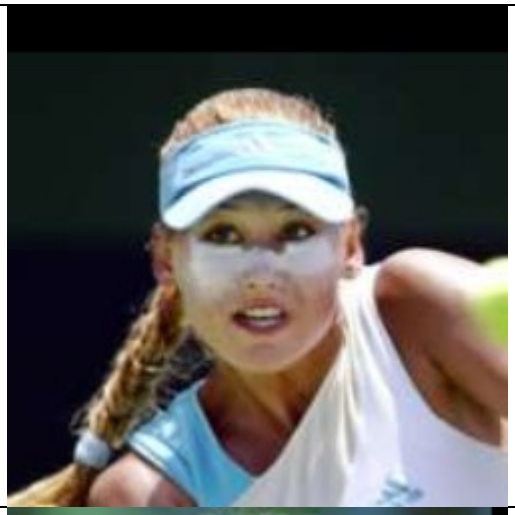


FIG. 7 ROC comparisons between original and corrected imagery

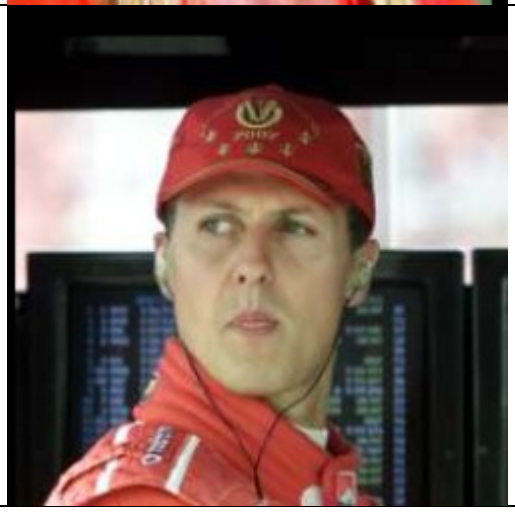
In Figure 7 the ROC curve from original vs. corrected imagery are presented with a logarithmic X-axis. This shows the scoring variances after the identity errors were corrected (blue area).

Investigating low scoring mates - the following images were confirmed identities but scored very low.

Low Scoring Mates








7. Outcomes

7.1 Based on this data set and the testing process documented here:

7.1.1 The LFW data set tested had identity errors which need to be adjusted to ensure proper scoring analysis could be done. From the default data set downloaded ~90 identities should have been corrected.

7.1.2 Iterative processes to properly analyze and modify this data set were required which focused on using the FAR/FRR curves to locate the identity errors.

7.1.3 Correcting the identity errors improved the FAR/FRR scoring analysis

7.1.4 Twins, siblings, and doppelgangers did affect the scoring analysis. Two twins, several siblings, and doppelgangers were located, causing high FAR scores.

7.1.5 FAR, FRR, DET and CMC curves were utilized in these processes.

7.1.6 Critical scoring analysis required the presentation of the scoring in both linear and logarithmic presentations to see the imposter scores which had identity errors.

7.1.7 A variety of methods can be used to resolve identity errors in facial datasets. FRS administrators should be aware of the advantages and disadvantages of each method before selecting and applying a method, especially on operational datasets.

Appendix A: Alternative Methods

A1. General/Initial assessment

A1.1 The initial assessment objectives are twofold: first, to acquire a general sense of how the FRS interacts with the data, and second, to select the identities or images for detailed review.

A1.2 The initial assessment assumes that similarity scores, both genuine and imposter are distributed according to predictable and smooth monomodal curves and that errors in the dataset are expected to introduce "unnatural" multimodal variations in the curves obtained from the sets of scores.

A1.3 Because the specific shape of this curve (e.g., Gaussian, exponential, log-normal) varies from system to system and also depends on the dataset itself, it is important to develop a general sense of the interaction between the FRS and the data.

A1.4 It is important to note that the effectiveness of these procedures, which are ultimately based on the distribution of scores, is fundamentally dependent on the accuracy of algorithm used to generate these scores. Algorithms whose scores are highly discriminative are more effective in identifying outliers.

A1.5 An initial assessment could then be based on curves derived from genuine and imposter scores distributions as, e.g., the FAR/FRR curve, or directly inspecting both distributions of scores.

A1.6 FAR/FRR curve plots both of these rates as a function of threshold. An example of such a curve for the LFW dataset is shown below.

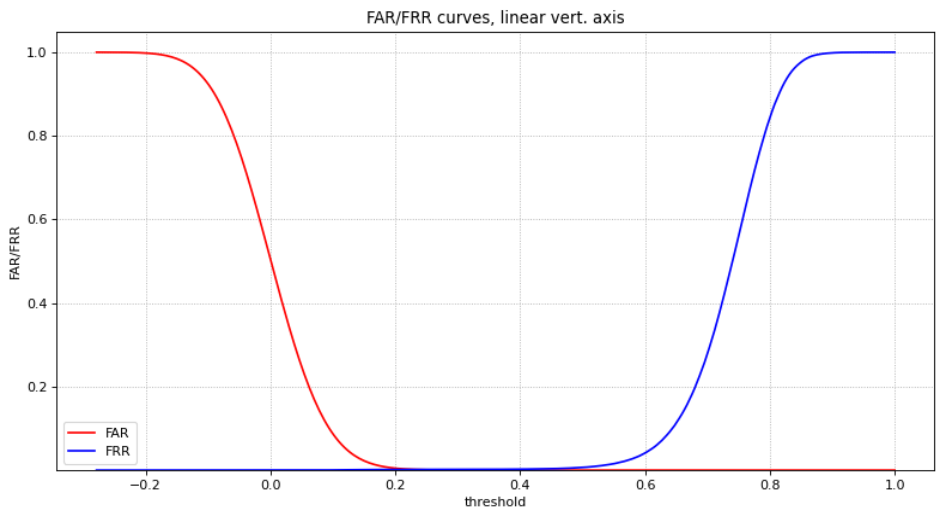


FIG. A1 FAR and FRR from uncorrected imagery

A1.7 In this plot, the vertical axis is linear, but plotting the same data with the vertical axis in a logarithmic scale facilitates the inspection of the curves in the lowest values.

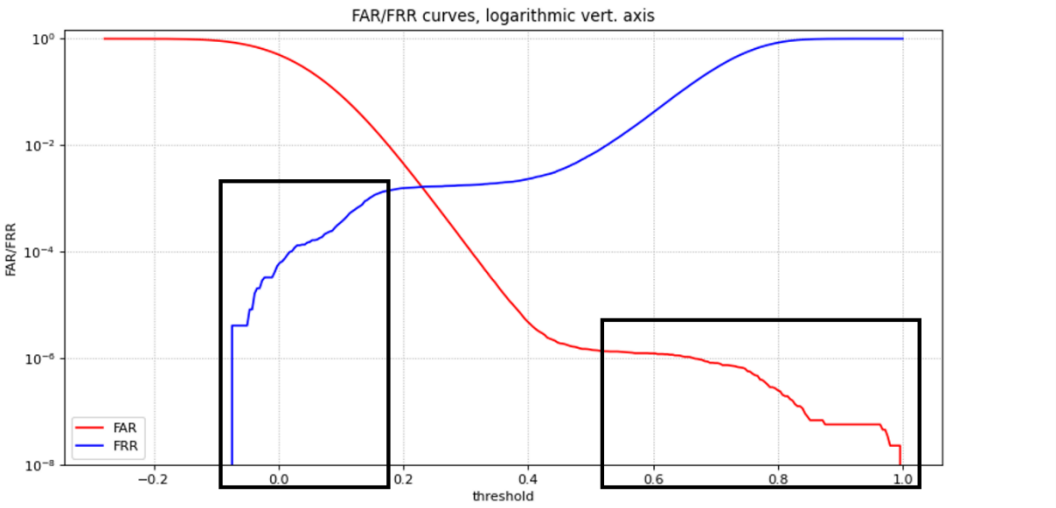


FIG. A2 FAR and FRR from uncorrected imagery (Logr axis)

A1.8 In this plot, it is now possible to observe variations on the shape of the FAR and FRR curves that violate the assumption of smoothness of a monomodal distribution curve (framed portions). This raises suspicions that errors exist in the dataset, resulting in more imposter scores with high values and more genuine scores with low values.

A1.9 This assessment can be further explored using the distributions of scores directly. The plot below is based on the same data.

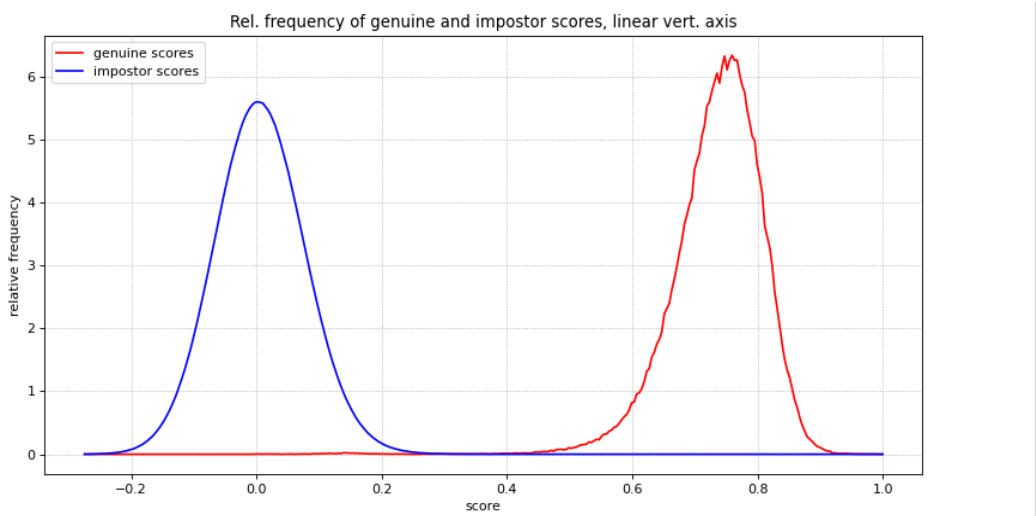


FIG. A3 Genuine and Imposter Scoring Histograms uncorrected imagery

A1.10 The vertical axis's linear scale again makes it challenging to inspect the lower values, but using a logarithmic vertical axis allows one to observe a violation of the assumption of smoothness of a monomodal distribution curve (framed portions).

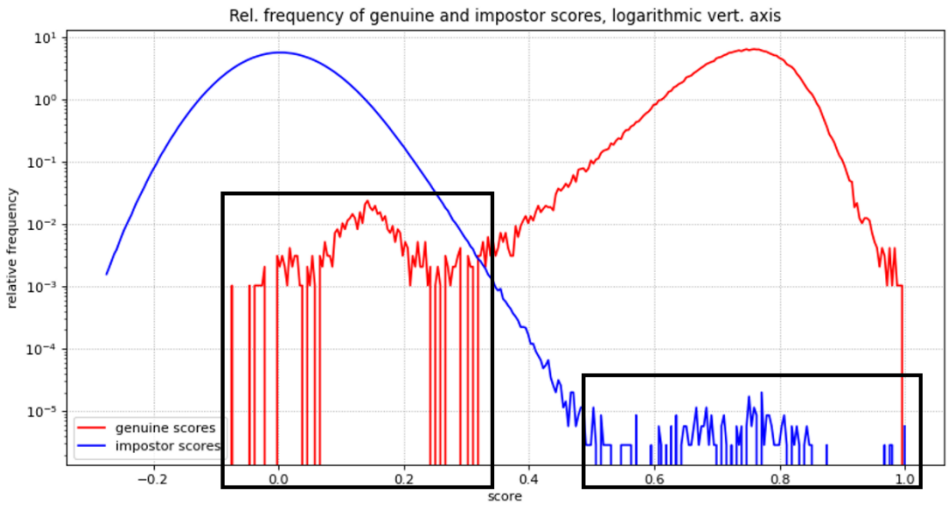


FIG. A4 Genuine and Imposter Scoring Histograms uncorrected imagery (Logr axis)

A1.11 Both the higher imposter scores and the lower genuine scores could be selected from this plot to a detailed review of the images involved. Imposter scores should be selected for further inspection in highest to lowest order, and the opposite for genuine scores.

A1.12 Apart from the visual assessment of the curves derived from the distribution of scores, statistical methods for identifying outliers in each set of scores can be used.

A1.13 One such method involves selecting outliers based on multiples of the median absolute deviation from median (MAD) of each set of scores. In univariate statistics, the MAD is a robust dispersion measure in the presence of outliers, contrary to the more commonly used standard deviation around the mean. [3]

A1.14 In practical terms, for genuine scores, those scores that lie below a multiple of MAD from the median should be considered as outliers and then manually verified. For imposter scores, the outliers are those higher than a multiple of MAD from the median. Larger multiples of MAD will result in smaller sets of outliers to be manually verified, while smaller multiples of MAD will result in a larger number of scores being considered as outliers.

A1.15 It is essential to notice that both the MAD and median statistics should be calculated independently for each set of scores, genuine and imposters.

A1.16 The specific multiple of MAD for each set should be determined experimentally, on a case by case basis. For the worked example, it was determined through a controlled experiment with the algorithm and the LFW dataset.

A1.17 For this specific purpose, controlled errors were introduced in the dataset, which added 764 errors in the genuine scores and the same quantity in the imposter scores. These experimental sets were verified using the criteria of multiples of MAD, which resulted in the number of six MADs as being adequate to identify the controlled errors as outliers in both genuine and imposter scores.

A1.18 The remainder of the analysis was conducted in the original LFW dataset, without the artificially introduced errors.

A1.19 The figure below illustrates the procedure for selecting outliers in the imposter scores, in the original LFW dataset, after the appropriate multiple of MAD was determined. The procedure for genuine scores is analogous.

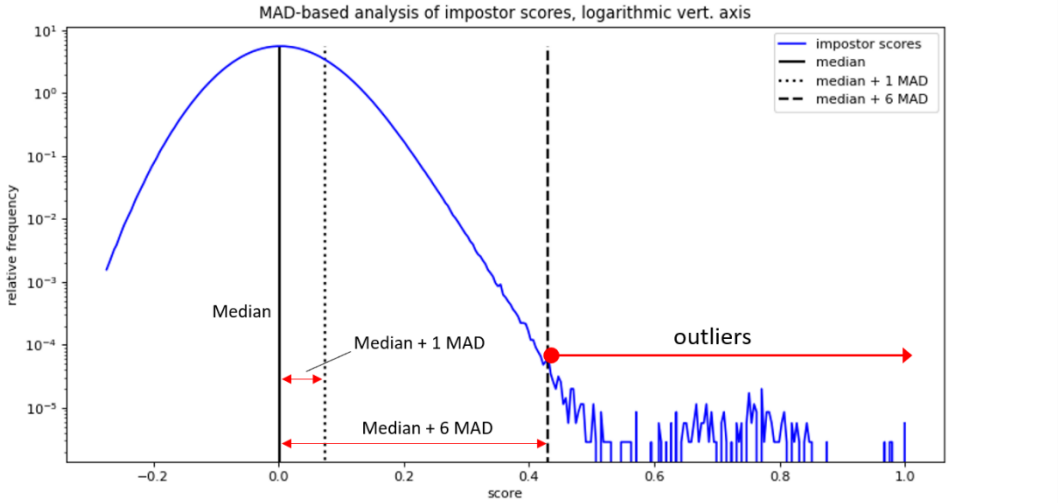


FIG. A5 MAD Based Imposter Scores uncorrected imagery (Logr axis)

A1.20 In the case of the original LFW dataset, without prior corrections of the known errors, the number of six MADs for the genuine and imposter scores sets resulted in the selection of 474 and 220 image pairs as outliers in each set, respectively.

A1.21 The amount of detailed inspection is subject to the available workforce. Ideally, it should be done until all image pairs identified as outliers are reviewed, and the FRS administrator is assured of the dataset integrity and correctness.

A2. Specific Assessments

A2.1. The **general**/initial assessment will result in two sets of image pairs (allegedly genuine and allegedly imposters) that must be manually reviewed to check for identity or other kinds of errors.

A2.2. Depending on the number of image pairs, a direct assessment of all pairs can be a viable approach. However, if the number of pairs is relatively large, some strategies can be employed to optimize this reviewing process.

A2.3. One such strategy is to verify if some images appear more frequently in each of the sets selected for review. These images should be inspected first.

A2.4. Results

A2.5. The selected pairs of images were inspected, and all the known errors published on the LFW website were identified.

A2.6. Some of the high imposter scores were confirmed to be from different persons. In most cases, this was caused by the presence of twins, siblings, or close appearances.

A2.7. For the low genuine scores, some were verified as being the same person, but variations in pose, make-up, facial expression, and age were factors the negatively affected the scores.

A2.8. After correcting the errors, the FAR/FRR and the scores distributions plots show a smoother monomodal behavior.

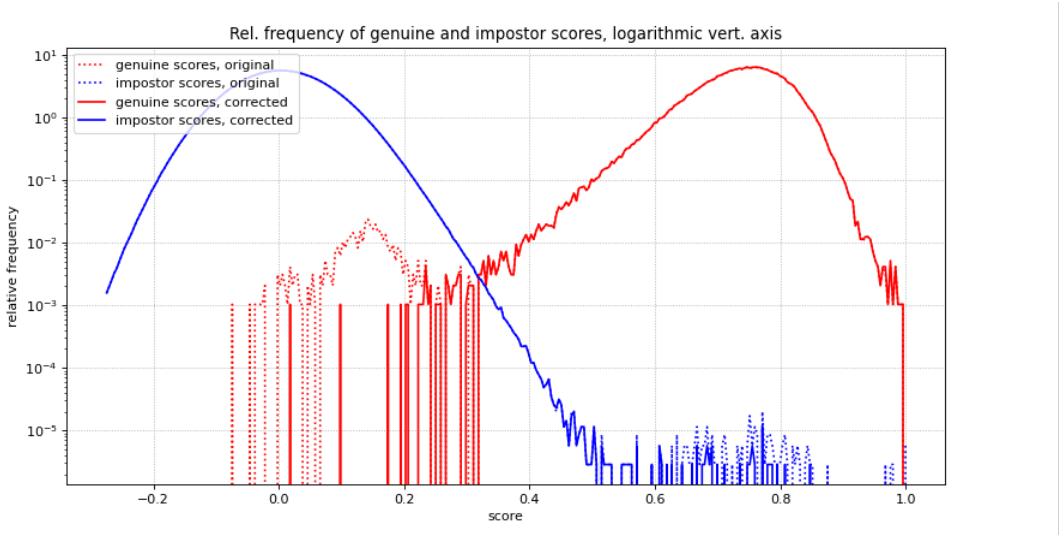


FIG. A6 Genuine and Imposter Scoring Histograms corrected imagery (Logr axis)

A2.9. The distributions of scores improved noticeably, with fewer outliers both in the genuine and imposter sets after the corrections were applied.

A2.10. This improvement can also be observed in the FAR/FRR plot, which is directly related to the scores distributions. After the corrections, EER improved by an order of magnitude.

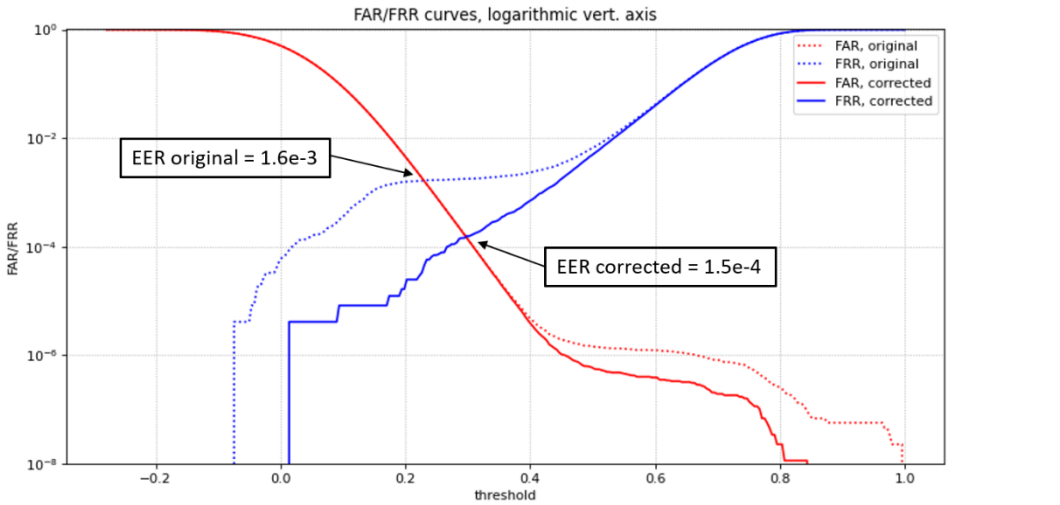


FIG. A7 FAR and FRR from uncorrected/corrected imagery (Logr axis)

A2.11. Finally, the DET plot also reveals improvement in all operational points.

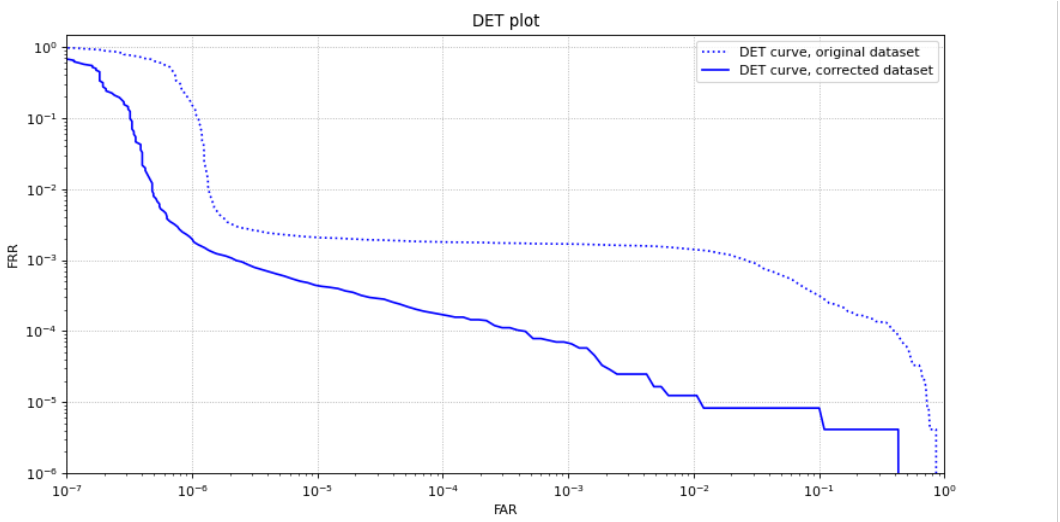


FIG. A8 DET from uncorrected/corrected imagery

A2.12. The whole process can be repeated until the FRS administrator is assured of the dataset integrity and correctness or while there are resources available.

A3. Other Methods

A3.1. The verification of integrity and correctness of biometric datasets is an evolving area in the scientific literature. In this section, some methods and approaches that can be used to tackle this problem are referenced.

A3.2. The chapter on Exploratory Data Analysis (EDA) of the NIST/SEMATECH e-Handbook of Statistical Methods [4] is recommended. The book describes EDA as "(…) an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to (i) maximize insight into a data set; (…) (iv) detect outliers and anomalies; (…)". Although not specific to inspecting facial datasets, many of the concepts presented in this chapter can be used for this purpose.

A3.3. In [5], the authors describe a method specifically designed to sort out identity errors in large facial datasets, similar to the MAD analysis described in the worked example. Their method is based on a two-layered thresholding process to select outliers. First, identity outliers are selected, and, for each selected identity, images considered as outliers are manually reviewed.

A3.4. In [6], the authors present the concept of Biometric Menagerie and propose a framework for evaluating biometric systems. This framework is based on the relationship between a person's genuine and imposter scores and could be explored to select identities for further inspection.

A3.5. Apart from methods that aid in selecting identities or images for manual review, some methods are proposed in the literature to automatically clean large facial datasets, with a large number of images assigned to each identity.

A3.6. In one such approach, presented in [7], a clustering algorithm, Density-Based Spatial Clustering of Applications with Noise – DBSCAN, clusters similar images in each identity set, retaining the cluster with most images.

A3.7. Another automatic method [8] employs the community detection algorithm to identify and delete mislabeled images while preserving diversity in each identity's images.

A3.8. The reader should be aware that these fully automated methods will remove some images from the dataset without human reviewing.

A3.9. Although there is a limited number of commercially available software specialized for this task, most scientific literature methods can be implemented in software devoted to statistics and mathematics.

.

Appendix B: LFW Data Set Information

B1. This is a widely used open source data set which will work well for this specific document. Information on this data set includes:

DISCLAIMER:

Labeled Faces in the Wild is a public benchmark for face verification, also known as pair matching. No matter what the performance of an algorithm on LFW, it should not be used to conclude that an algorithm is suitable for any commercial purpose. There are many reasons for this. Here is a non-exhaustive list:

- Face verification and other forms of face recognition are very different problems. For example, it is very difficult to extrapolate from performance on verification to performance on 1:N recognition.
- Many groups are not well represented in LFW. For example, there are very few children, no babies, very few people over the age of 80, and a relatively small proportion of women. In addition, many ethnicities have very minor representation or none at all.
- While theoretically LFW could be used to assess performance for certain subgroups, the database was not designed to have enough data for strong statistical conclusions about subgroups. Simply put, LFW is not large enough to provide evidence that a particular piece of software has been thoroughly tested.
- Additional conditions, such as poor lighting, extreme pose, strong occlusions, low resolution, and other important factors do not constitute a major part of LFW. These are important areas of evaluation, especially for algorithms designed to recognize images “in the wild”.

For all of these reasons, we would like to emphasize that LFW was published to help the research community make advances in face verification, not to provide a thorough vetting of commercial algorithms before deployment.

While there are many resources available for assessing face recognition algorithms, such as the Face Recognition Vendor Tests run by the USA National Institute of Standards and Technology (NIST), the understanding of how to best test face recognition algorithms for commercial use is a rapidly evolving area. Some of us are actively involved in developing these new standards and will continue to make them publicly available when they are ready.

This document includes a cover page with the FISWG disclaimer

Welcome to Labeled Faces in the Wild, a database of face photographs designed for studying the problem of unconstrained face recognition. The data set contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector. More details can be found in the technical report at http://vis-www.cs.umass.edu/lfw/lfw.pdf

Critical to this document's purpose is the errata found on the LFW website. A list of errors can be viewed at: http://vis-www.cs.umass.edu/lfw/

FISWG documents can be found at: www.FISWG.org