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: Image Quality Assessment

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 are 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 documents:

- “Understanding and Testing for Face Recognition Systems Operation Assurance”
- “Face Recognition Systems Operation Assurance: Identity Ground Truth”

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 how testing for variations in facial image quality can be used in adjusting operational workflows. Testing for and verifying scoring variations when image quality varies is key so that facial search system workflows can be properly adjusted.

2. Referenced Documents

2.1 ANSI/NIST-ITL Standard Homepage: http://www.nist.gov/itl/iad/ig/ansi_standard.cfm

2.2 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

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 CMC–cumulative match characteristic

3.2.2 DET–detection error tradeoff

3.2.3 FAR–false accept rate

3.2.4 FMR–false match rate

3.2.5 FNMR–false non-match rate

3.2.6 FR–facial recognition

3.2.7 FRR–false reject rate

3.2.8 FRS–facial recognition system

3.2.9 ROC–receiver operating characteristics

4. Background

4.1 When doing *accuracy* profiling, a *critical* step is to understand how image quality affects the processing of the imagery to be enrolled or searched. Regardless of the specific facial biometric algorithm used, there will be some facial imagery which causes facial localization errors which then produce biometric templates of no operational value.

4.2 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.

4.3 Setting up frameworks to do *enrollment* and searching 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.

5. Data Set

5.1 Care should be taken in selecting data sets to profile. 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 Include associated identity truth information.

5.2 The data set used for this document is the Labeled Faces in the Wild (LFW) data set available at: http://vis-www.cs.umass.edu/lfw/ See section "LFW Data Set Information" at http://vis-www.cs.umass.edu/lfw/#information and the referenced paragraph 2.2 document (NISTIR 8271 DRAFT SUPPLEMENT Face Recognition Vendor Test (FRVT) Part 2: Identification) for more information. Conceptually any other facial data set with identity ground truth can be used.

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

5.3.1 Has smaller but consistent image sizes and file formats;

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

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

5.3.4 Has stated identity ground truth errors.

6. Image Quality Assessment Process

6.1 **Step 1 -** Ensure the data set to use has verified ground truth.

6.2 **Step 2 -** Extract the facial image quality from all images to enroll and search. The image quality metrics to use will be vendor dependant. Care should be taken to select the quality metric(s) which have the largest value in doing an image quality assessment. Vendor specific quality metrics will have different numeric ranges and will have vendor specific usability ranges. Consulting with the provider of the facial algorithm is recommended.

6.3 **Step 3 -** Enroll the facial images into a facial gallery for searching.

6.4 **Step 4** - Search the facial images against the facial gallery. The number of candidates returned for this document was 50. This number may vary with agency specifics. Do not use any scoring thresholds.

6.5 **Step 5** - Segment facial search results based on the facial image quality of the probe.

6.6 **Step 6** - For each segmentation group of the search results plot the biometric performance results and compare them. This document uses these plots:

6.6.1 FAR: False accept scoring

6.6.2 FRR: False reject scoring

6.6.3 DET: Detection error tradeoff

6.6.4 CMC: Cumulative Match Curve

6.7 **Step 7:** Determine if targeted image quality assessments can be done before enrollment or searching to see if improper facial localization has occurred which caused facial templates to be created which were of no operational value and should have either been manually reviewed or simply not processed.

7. Image Quality Assessment Process Step Outputs

7.1 Step 2 Outputs:

7.1.1 When this was done with the LFW data set the following facial metrics were recorded:

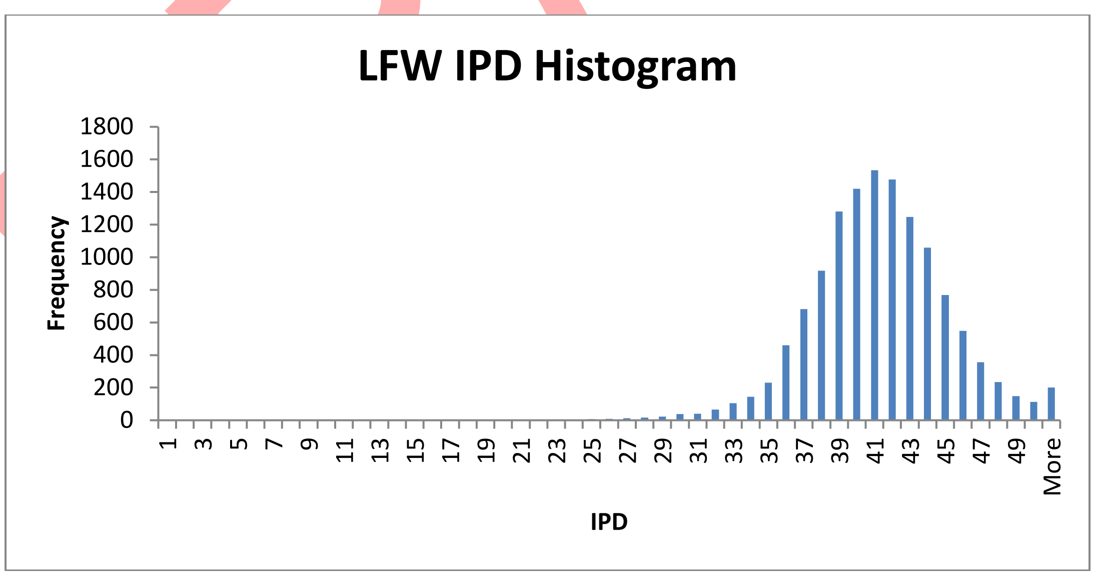


Figure 1: LFW Interpupillary Distance (pixels)

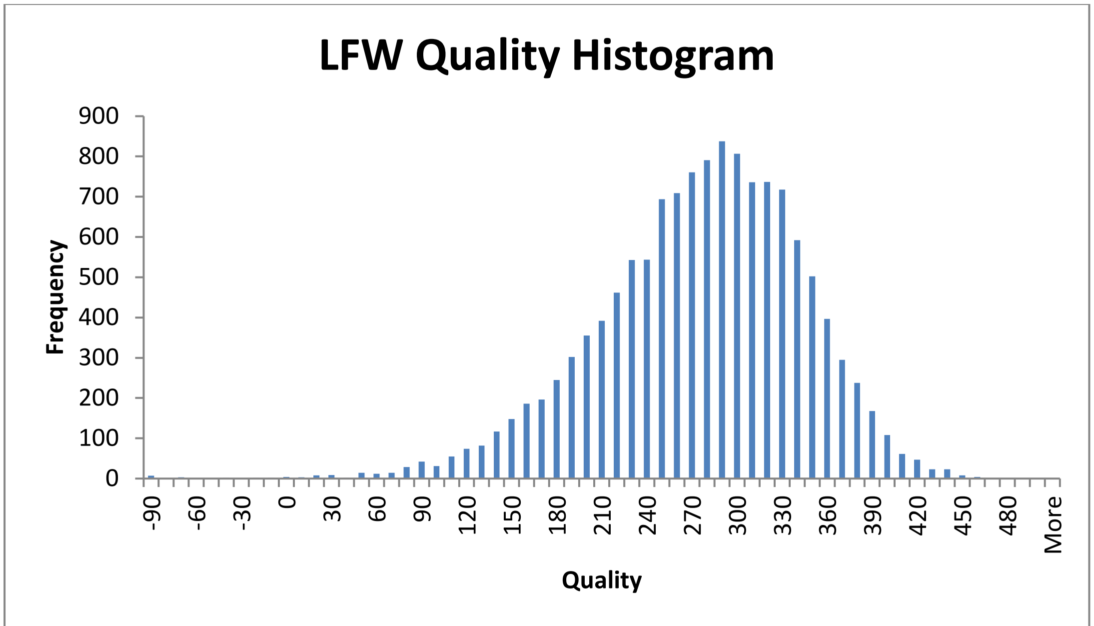


Figure 2: LFW Quality (linear Y axis)

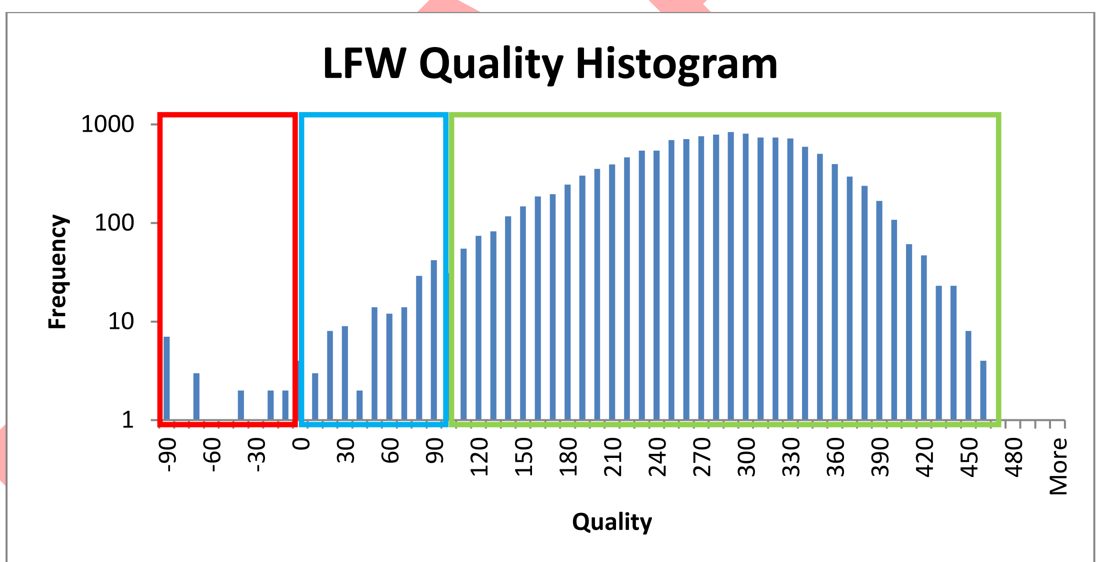


Figure 3: LFW Quality (logarithmic Y axis)

7.1.2 The LFW images were then enrolled and searched. Based on the image quality histograms three segments were selected:

7.1.2.1 All search results

7.1.2.2 All searches with a quality less than 0

7.1.2.3 All searches with a quality between 0 and 100

7.1.2.4 All searches with a quality greater than 100

7.2 Step 6 Outputs:

7.2.1 Once the accuracy plots were created, they could be analyzed.

7.2.2 **FAR and FRR were analyzed first.**

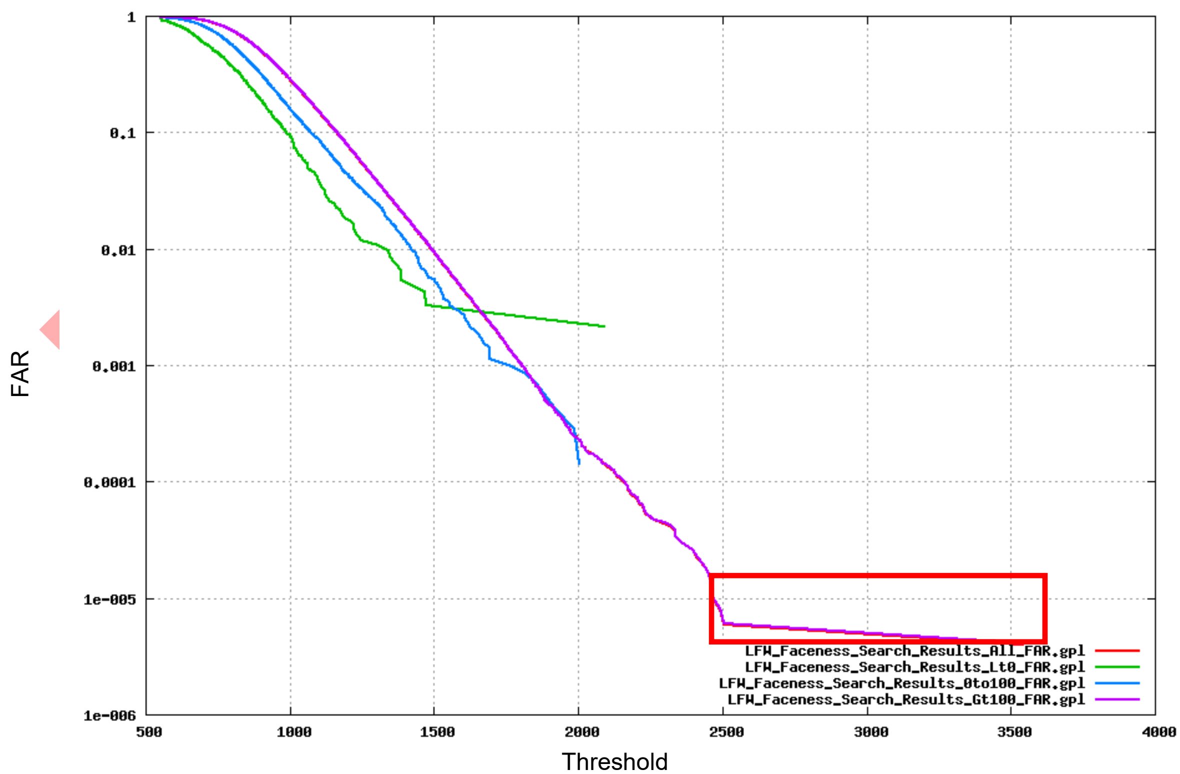


Figure 4: LFW FAR Imposter Scores (logarithmic Y axis)

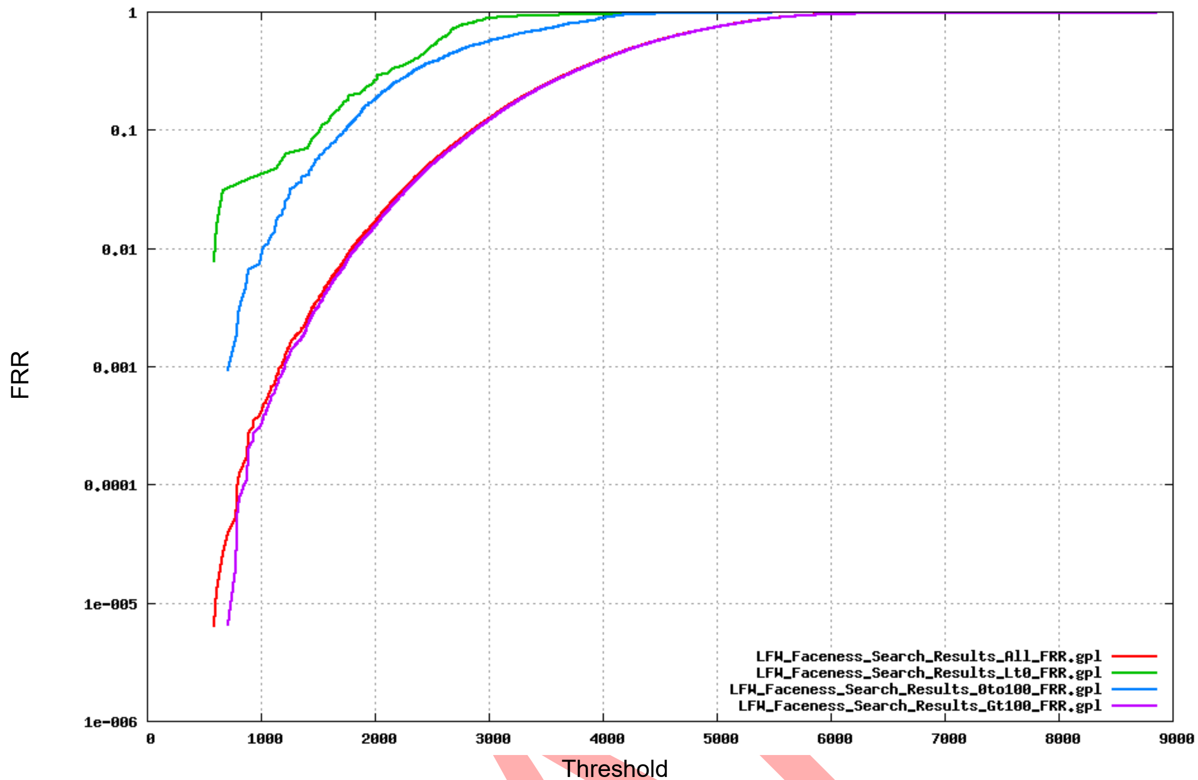


Figure 5: LFW FRR Mate Scores (logarithmic Y axis)

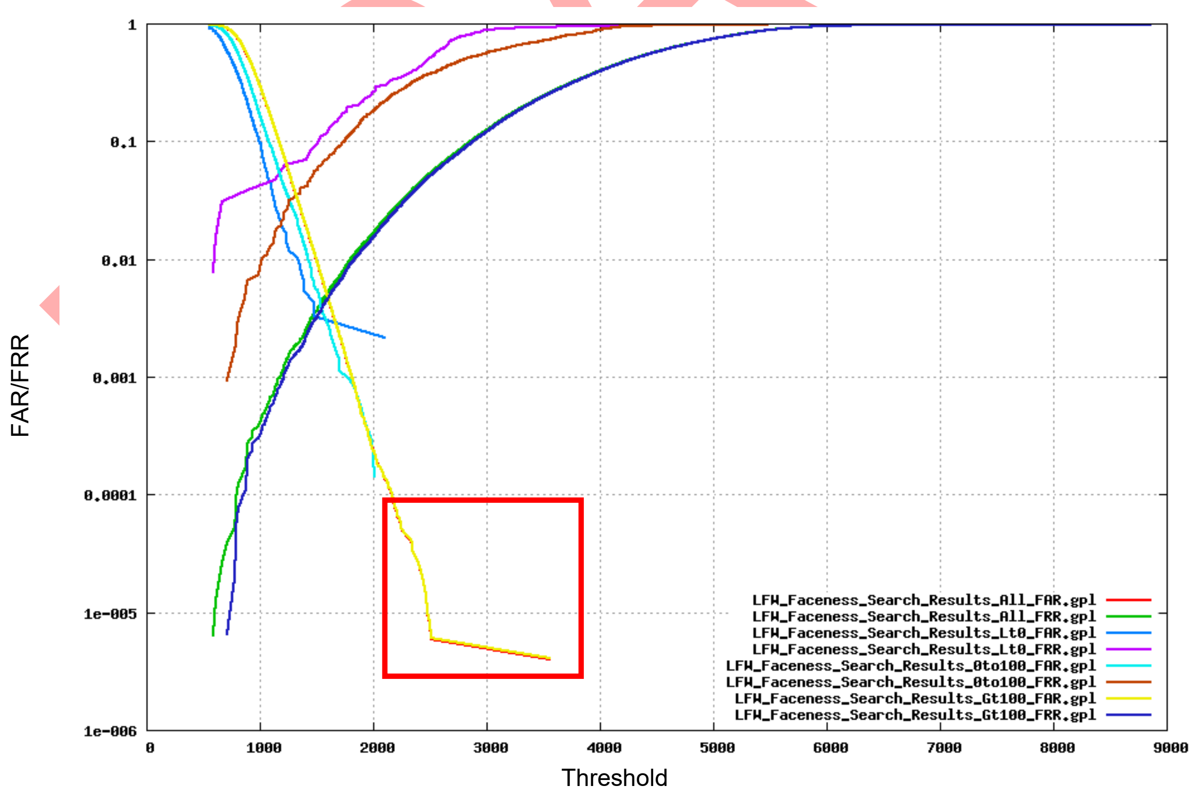


Figure 6: LFW Mate and Imposter Scores (FAR and FRR)

7.2.3 **Notes on Step 6 FAR and FRR analysis**:

7.2.3.1 The overall imposter (FAR) scoring in Figures 4 and 6 was relatively consistent regardless of probe image quality with high values of ~2500.

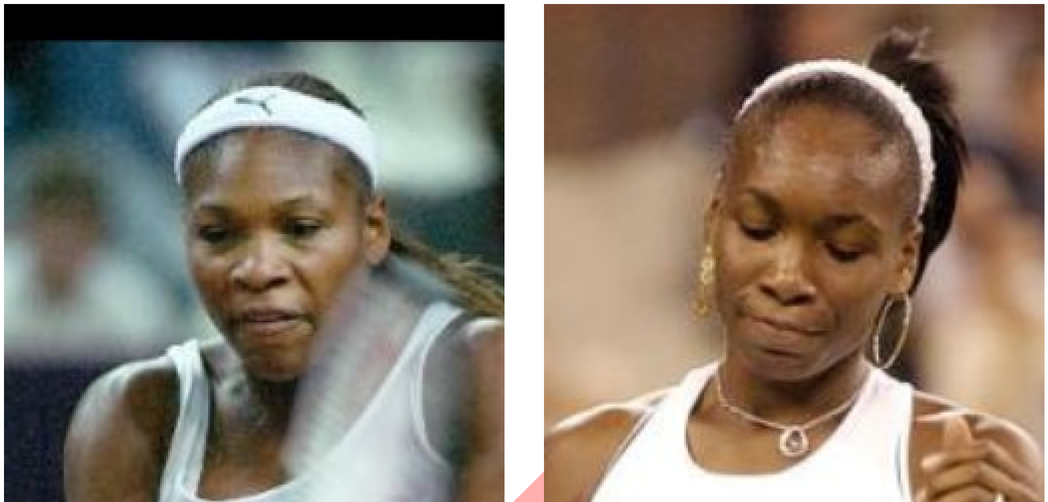
7.2.3.2 The higher imposter (FAR) scoring in Figures 4 and 6 (red box) was due to the presence of known twins and several siblings and doppelgangers.

7.2.3.3 The highest imposter (FAR) score of ~3500 in Figures 4 and 6 (red box) was due to known twins in the LFW data set.

7.2.3.4 The overall mate scoring (FRR) in Figures 5 and 6 did vary with probe image quality with the lowest values of ~750.

LFW Twin

LFW High Scoring Imposters



The images above are siblings

The images above are doppelgangers

7.2.4 **DET was analyzed next.**

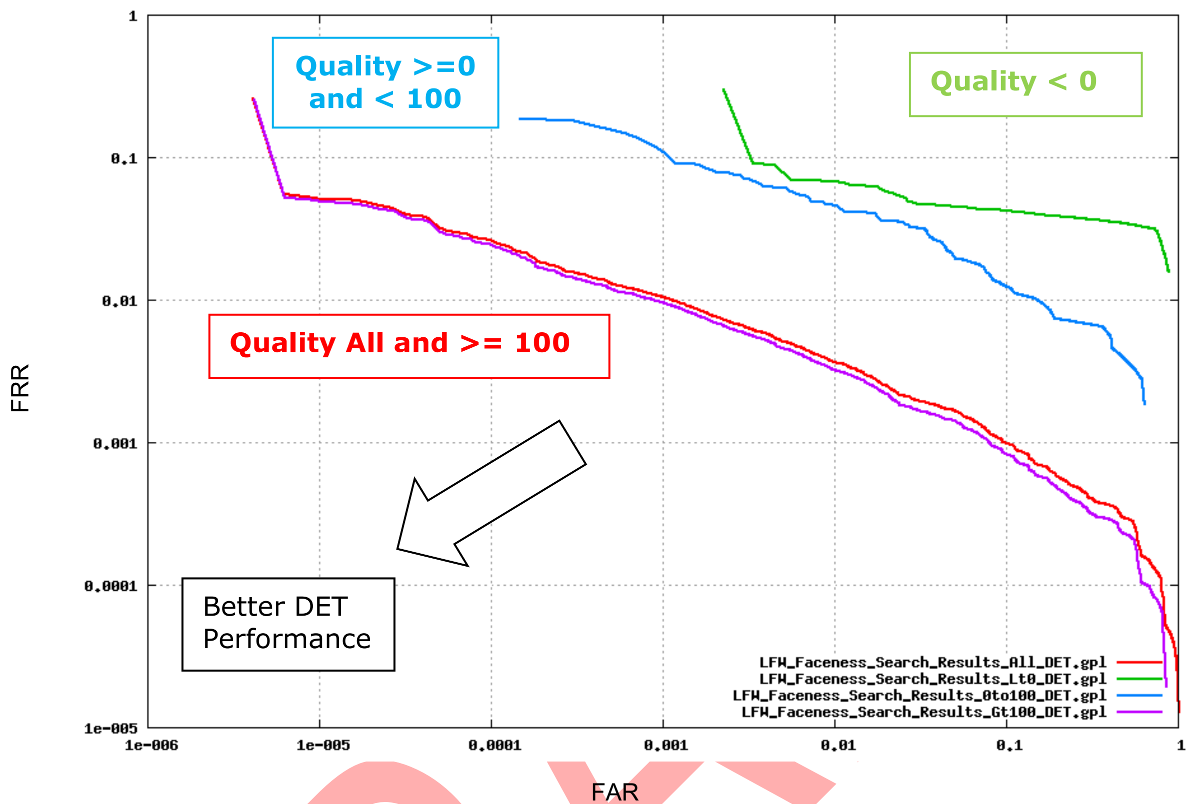


Figure 7: LFW DET Curve

7.2.5 **Notes on Step 6 DET analysis:**

7.2.5.1 The DET curve performance in Figure 7 was better with higher probe image quality.

7.2.6 **CMC was analyzed next.**

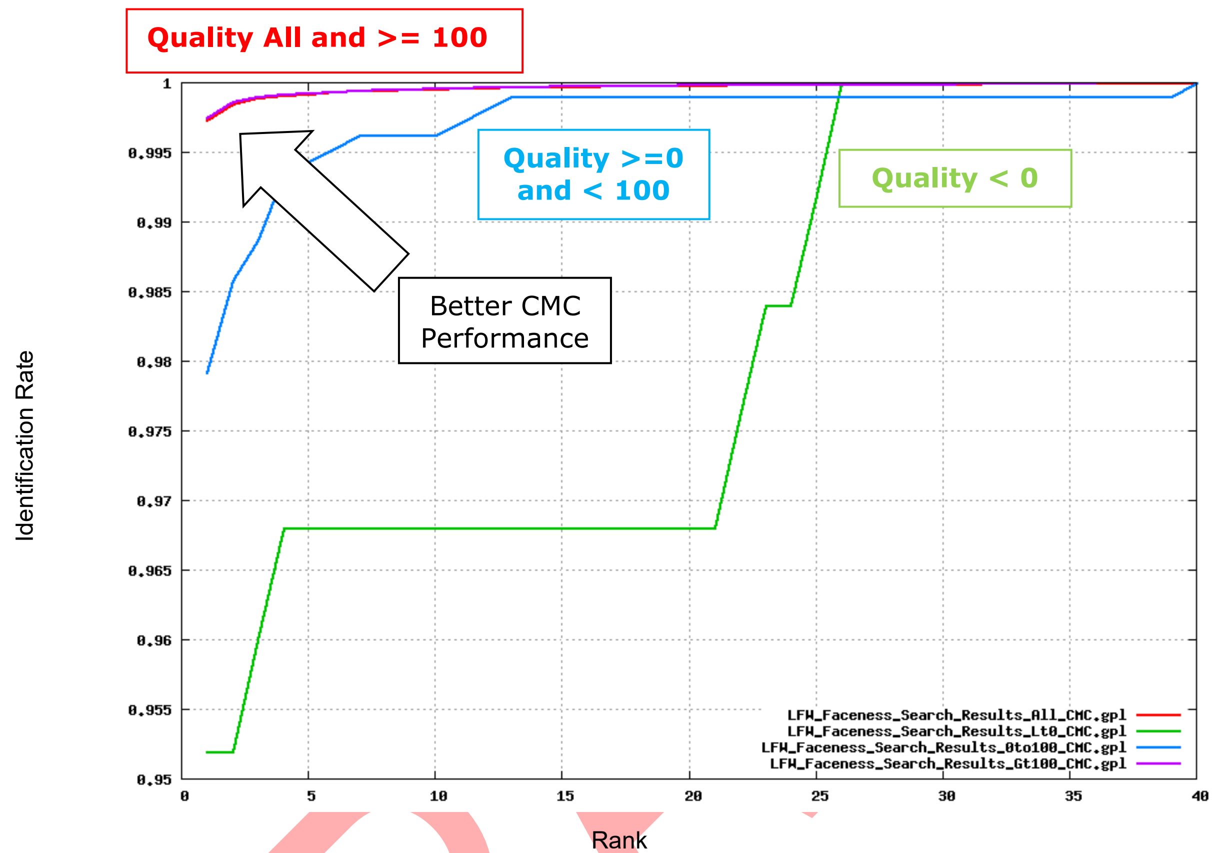


Figure 8: LFW CMC Curve

7.2.7 **Notes on Step 6 CMC analysis:**

7.2.7.1 The CMC curve performance in Figure 8 was better with higher probe image quality.

7.2.8 Perhaps the largest outcome of this testing is presented in the CMC curves and can be described as follows:

7.2.8.1 The lowest image quality plots (< 0) have a 100% CMC point at a search results candidate of ~27.

7.2.8.2 The medium image quality plots (>= 0 and < 100) has a 100% CMC point at a search results candidate of ~13.

7.2.8.3 The highest image quality plots (>= 100) has a 100% CMC point at a search results candidate of ~10.

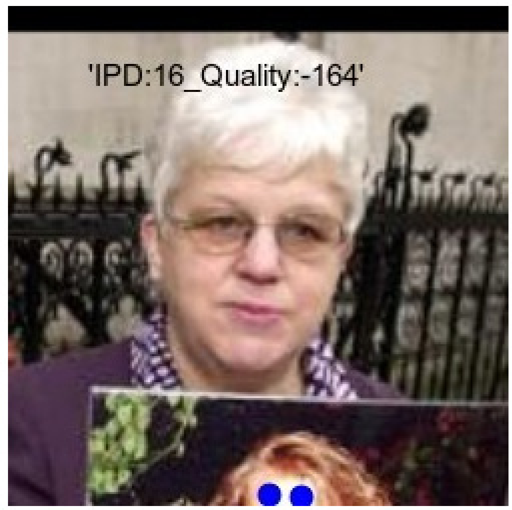
7.2.8.4 So, in order to achieve the 100% CMC point for all image quality variations the number of candidates a human examiner must review is ~27.

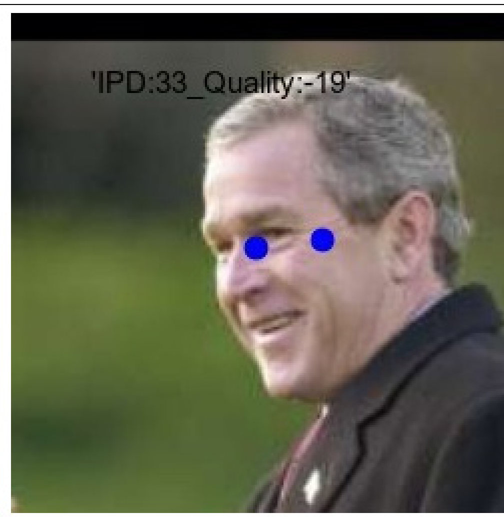
7.2.8.5 If the overall image quality can be improved through proper image quality assessment during enroll and searching, then it would be possible to lower number of candidates a human examiner must review to achieve a desired 100% CMC point.

7.3 **Step 7 Outputs:**

7.3.1 The search probes where then analyzed to locate low quality imagery so the candidate lists returned could be reviewed to see if any mates were returned.

This probe had 51 mates in the gallery. Only one was returned at rank 3. The probe quality was -186.	

This mate which was returned at a rank 3 with a score of 1213. Note the bad eye locations which were similar. The probe quality was -163.	
This probe had 0 mates in the gallery. The probe quality was -164. This example shows how the incorrect face in the image was localized.	
This probe had 0 mates in the gallery. The probe quality was -123. This example shows how improper facial localization occurred.	

This probe had 7 mates in the gallery. Only two were returned. The probe quality was -84.	
This mate which was returned at a rank 26 with a score of 575.	
This probe had 530 mates in the gallery. No mates were returned. The probe quality was -19.	

This probe had 8 mates in the gallery. No mates were returned. The probe quality was 22.	

8. Image Quality Assessment Outcomes

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

8.1.1 Facial search performance is affected *by facial image quality which will be vendor dependent. Various image quality metrics needed to be analyzed to profile the facial image quality into specific ranges of interest.*

8.1.2 *The ability to* segment search results based on facial image quality is shown.

8.1.3 Mates and imposter scoring profiles can change with mates being more susceptible to image quality issues.

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

8.1.5 How the facial image quality scoring and resultant operational ramifications need to be profiled to properly locate low scoring mates in search results which have lower image quality.

8.1.6 Detecting *and correcting low quality facial imagery w*hich could be manually corrected could improve overall facial accuracy.

8.1.7 Enrolling and searching all *facial* imagery without attention to image quality will negatively affect facial workflows.

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