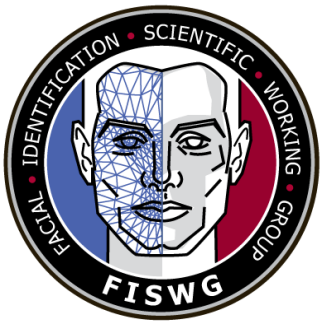

Disclaimer:

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.

Facial Recognition Systems Operation Assurance: Scoring Thresholds

1. Scope

1.1 The scope of this document is to provide a detailed process and examples of how to evaluate scoring thresholds when adjusting operational workflows. Properly executing biometric performance assessment by producing appropriate charts to inspect and review scoring thresholds can be done for many reasons. This document will focus on how this analysis can support a systematic process to determine appropriate facial scoring thresholds to support end user requirements. This document is relevant to systems that operate with automated workflows as well as investigative systems requiring a human practitioner to review a candidate list.

1.2 Understanding how to evaluate facial biometric scoring is critical for both system accuracy and workflows of the human practitioners.

1.3 Topics outside the scope of this document include, but are not necessarily limited to setup, system tuning, workflow management and improvement, and proof-of-concept pilots.

2. Referenced Documents

ASTM Standards: [1]

E2916 Terminology for Digital and Multimedia Evidence Examination

E2825 Standard Guide for Forensic Digital Image Processing

Other Standards:

ANSI/NIST- ITL-1-2011 Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information2

NISTIR 8271 Face Recognition Vendor Test (FRVT) Part 2: Identification

[1] For referenced ASTM standards, visit www.nist.gov/osac/astm-launch-code, or the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For Annual Book of ASTM Standards volume information, refer to the standard's Document Summary page on the ASTM website.

[2] National Institute of Standards and Technology (NIST) standards available from website https://www.nist.gov.

3. Terminology

3.1 *Definitions:*

3.1.1 For terms relating to digital and multimedia evidence, refer to Terminology E2916.

3.2 *Definitions of Terms Specific to This Document:*

3.2.1 *Doppelganger–an apparition or double of a living person.*

3.3 *Acronyms*

3.3.1 *FR*–Facial Recognition

3.3.2 *FRS*–Facial Recognition Systems

3.3.3 *CMC*–Cumulative Match Characteristic

3.3.4 *ROC*–Receiver Operating Characteristics

3.3.5 *DET*–Detection Error Tradeoff

3.3.6 *FMR*–False Match Rate proportion of the completed biometric non-mated comparison trials that result in a false match. This will be referred to as FAR (false acceptance rate) and does not include errors from images which do not create valid templates.

3.3.7 *FNMR*–False Non-Match Rate proportion of the completed biometric mated comparison trials that result in a false non-match. This will be referred to as FRR (false reject rate) and does not include errors from images which do not create valid templates.

3.3.8 *IPD:* Interpupillary Distance

4. Summary of Guide

4.1 This document provides guidelines and techniques to help administrators of facial recognition systems (FRS) produce recognition statistics on the facial recognition systems which can be used to improve overall biometric performance.

4.2 The intended audience of this document is system owners, system users, and system administrators of existing facial recognition systems.

4.3 This document is a continuation of the FISWG documents:

4.3.1 “Understanding and Testing for Facial Recognition Systems Operation Assurance”

4.3.2 “Facial Recognition Systems Operation Assurance: Part 2, Identity Ground Truth”

4.3.3 “Facial Recognition Systems Operation Assurance: Part 3, Image Quality Assessment”

4.3.4 “Facial Recognition Systems Operation Assurance: Part 4, Manual Facial Localization”

4.4 The issues presented in this document form a foundation for other considerations and applications when testing such as system setup and tuning.

5. Significance and Use

5.1 Introduction

5.1.1 For 1:N (one to many) searches, a probe image is searched against a collection of images stored in a gallery. For threshold-based workflows, the search returns candidate(s) that ‘match’ the probe image above a pre-defined threshold. The length of the candidate list is usually set by the user with longer candidate lists requiring more human review effort or resource.

5.1.2 The candidates are normally sorted to have the highest score as the first (Rank 1) candidate and the lowest score as the last candidate (Rank N).

5.1.3 Ideally, a probe with a mate(s) in the gallery (mated pair) will return a high score(s) and a probe without a mate in the gallery (non-mated pair) will return low scores, allowing a clear discrimination in the meaning and inference of the results in the candidate list. If you had a perfect biometric algorithm and searched a large number of probes against a gallery with known mates and non-mates and plotted the candidate scores you would come up with a distribution shown below in Figure 1.

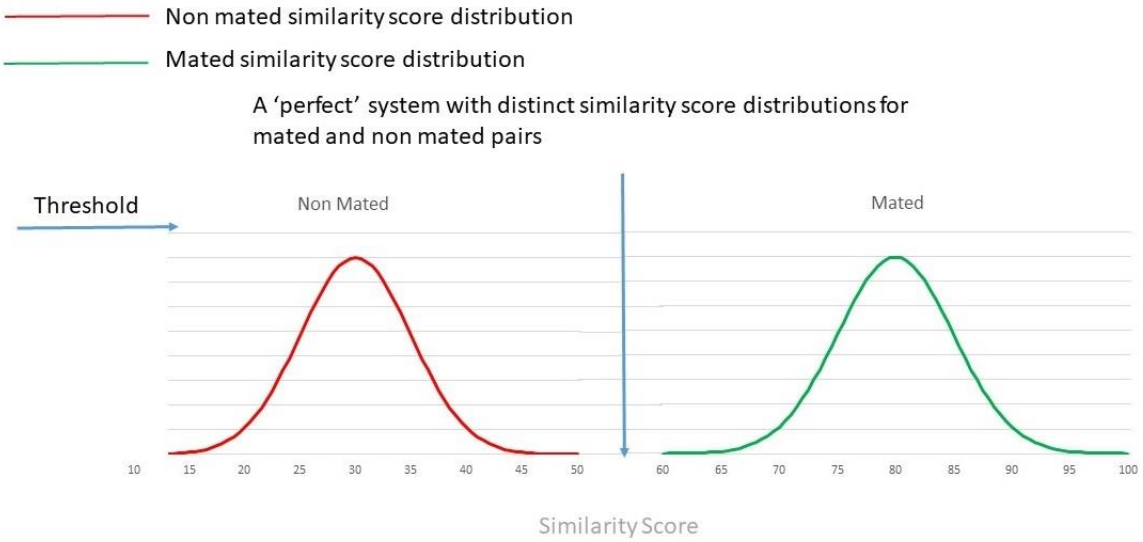

Figure 1: Distribution of similarity scores for mated & non-mated image pairs for a ‘perfect’ biometric system

- This plot shows that all the mates have a high score and all the non-mates have a low score. For this, a simple threshold score can be used to determine a mate and this would support an automated workflow scenario with no manual review.
- Unfortunately, there is no such thing as a perfect biometric algorithm. Some mates score low and some non-mates score high, which result from various conditions. The result of this is that there is a cross-over in the distributions of scores for mated and non-mated pairs and a resulting range of scores, sometimes referred to as “Yellow Resolve” where the score alone can’t accurately and reliably be used to determine a mate from a non-mate (see Figure 2). A facial comparison practitioner must therefore review the candidate list and make a potential mate/non-mate decision.

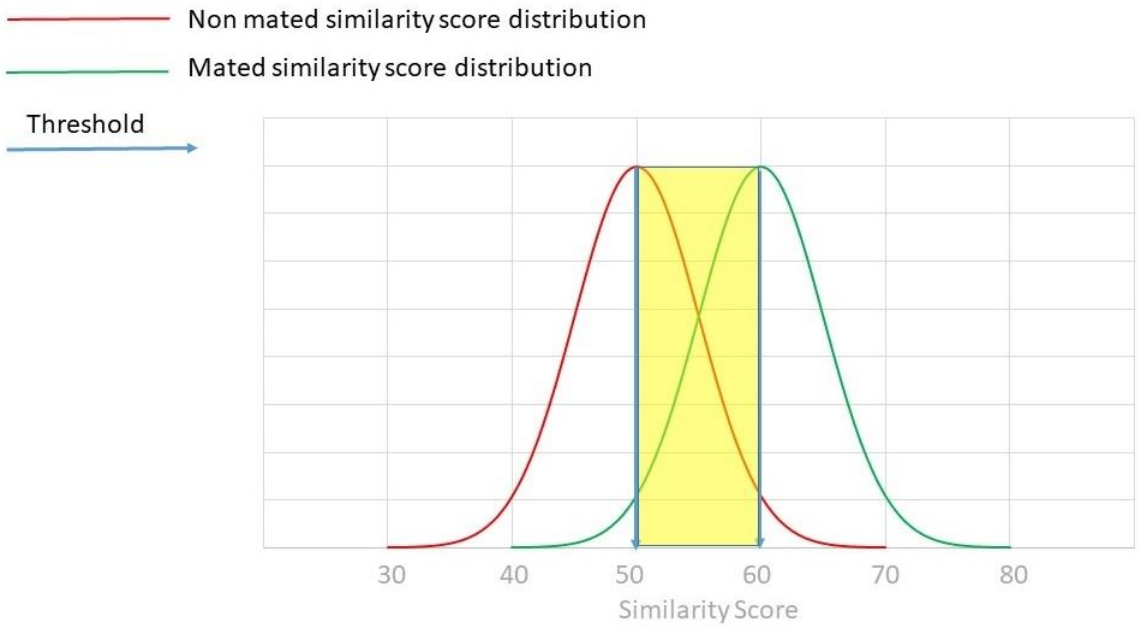

Figure 2: Example distribution of similarity scores for mated & non-mated image pairs showing ‘yellow resolve’ zone (not generated from real data)

- The numeric ranges and meaning of the score distribution within this ‘yellow resolve’ range need to be understood by the operators of the solution.
- The solution will have specific requirements for accuracy in terms of human practitioner resource availability balanced against the acceptability of incorrect matches (the False Accept Rate, requiring more time and effort to

review) and the risk of missing a mated image (the False Reject Rate), (see Figure 3).

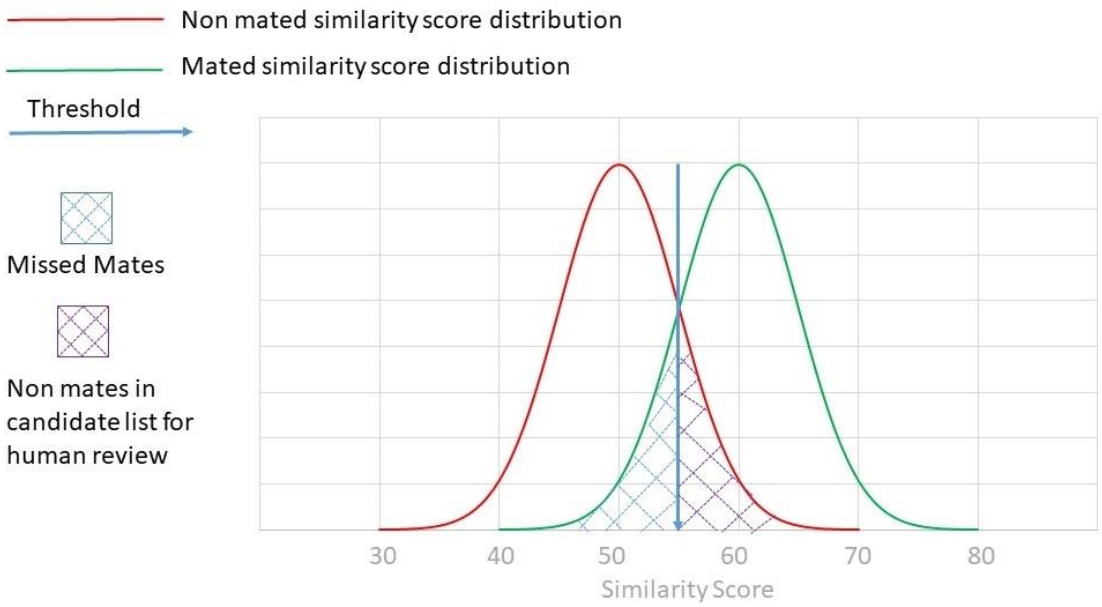

Figure 3: Example distribution of similarity scores for mated & non-mated image pairs showing the proportion of False Rejects and False Accepts for a particular threshold setting (not generated from real data)

- Setting the system-operating threshold should be aligned to the Concept of Operations (ConOps) and the level of human resource available. For a high throughput ConOps where human review must be minimized, a high threshold setting will eliminate False Accepts, but there is a trade-off as the proportion of False Rejects is increased (see Figure 4). For a high security ConOps, a low system threshold will ensure that mated pairs are returned in the candidate list but the trade-off is that the proportion of False Accepts is increased requiring a high level of human review resource (see Figure 5).

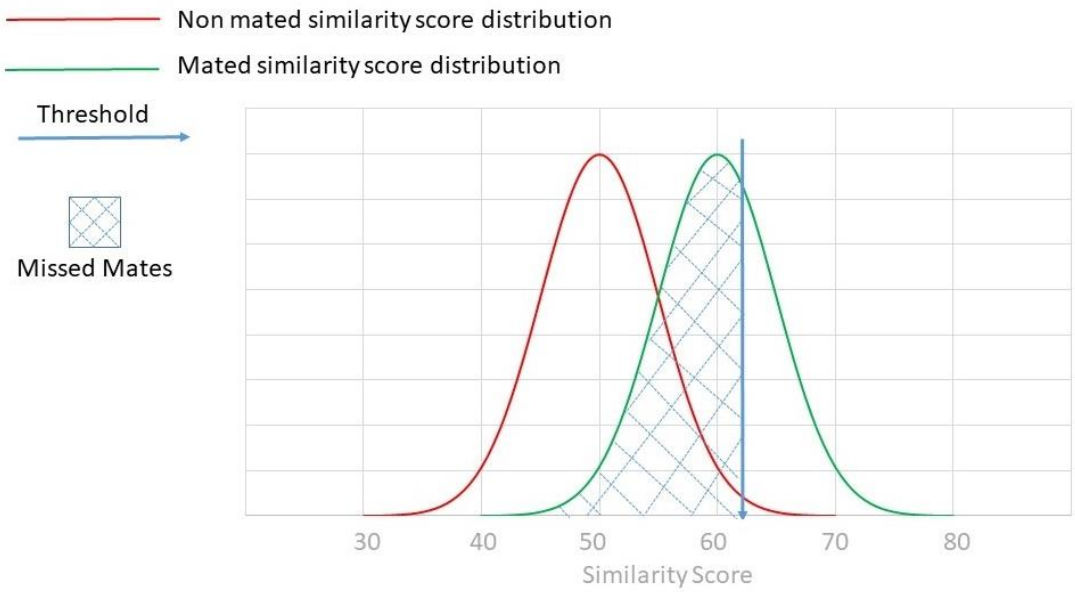

Figure 4: Example high Throughput ConOps where a high threshold eliminates False Accepts but results in a high proportion of missed mates (not generated from real data)

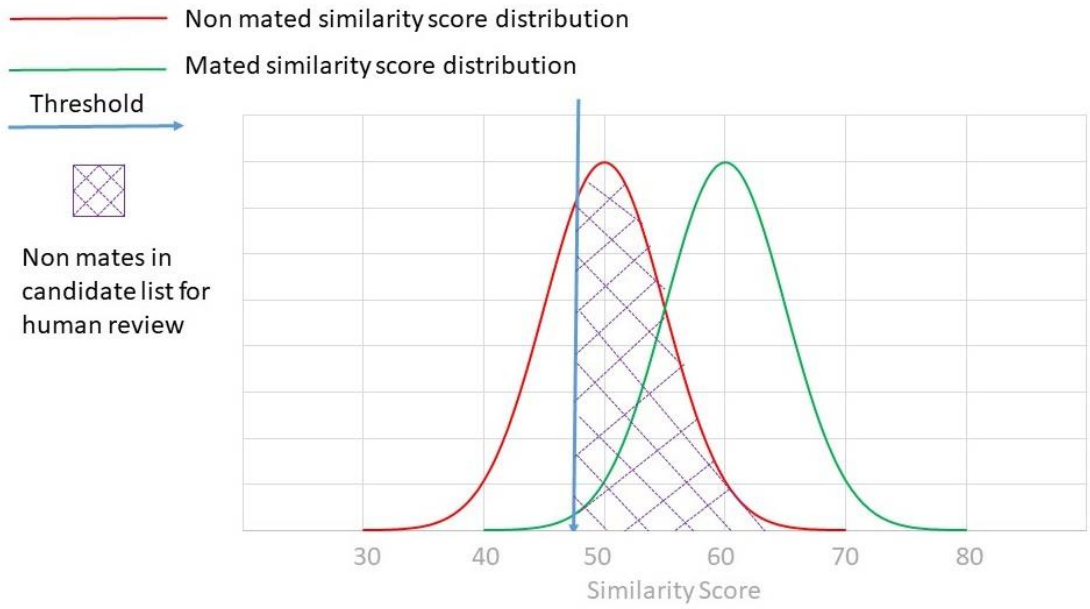

Figure 5: Example high Security ConOps where a low threshold maximizes the return of True Mates but results in a high proportion of False Accepts for human review (not generated with real data)

5.2 Considerations

5.2.1 The tuning and operational implementation of the mate, non-mate, and yellow resolve areas are dependent on the areas mentioned previously:

- The quality of the facial data (both probe and reference images)
- The performance of the algorithm
- The requirements of the solution

5.2.2 Factors impacting the facial data quality include:

- Were the images captured in controlled or uncontrolled environments?
- Were different capture systems used to populate the gallery?

5.2.3 Factors impacting the facial algorithm include:

- How do the varying types of image quality interact with each other when doing 1:N-based searching?
- How selective is the algorithm in properly discriminating the mates and non-mates?

5.2.4 Factors impacting the requirements of the solution include:

- What is least favorable? A false accept (e.g., a high scoring non-mate) or a false reject (e.g., a low scoring mate)?
- How will the yellow resolve area be managed? This usually involves manual review of the search to determine if a mate exists. This manual review takes a varying amount of time to determine and depending on agency protocols; two separate examinations may be required by distinct practitioners. This adds time to the search, which can then be translated into a labor cost and a cost in terms of operational search throughput.

5.2.5 A key reference document to review is:

NISTIR 8271 DRAFT SUPPLEMENT Face Recognition Vendor Test (FRVT) Part 2: Identification

Low similarity scores: In thousands of mugshot cases the correct gallery image is returned at rank 1 but its similarity score is nevertheless low, below some operationally required score threshold. This is not so important when face recognition is used for “lead generation” in investigational applications because human reviewers are specifically required to review potentially long candidate lists and the threshold is effectively 0. In applications where search volumes are

> higher and labor is not available to review the results from searches, a higher threshold must be applied. This reduces the length of candidate lists and false positive identification rates at the expense of increased false negative miss rates. The tradeoff between the two error rates is reported extensively later.

From: https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf or https://doi.org/10.6028/NIST.IR.8271

5.3 Important Notes

5.3.1 Care should be taken in selecting data sets to assess. It is recommended to select data sets which:

- Have operational relevancy if actual operational data is not available.
- Have consistent image quality aspects: type of capture, size of images, subject poses, etc. Many operational datasets do not contain imagery that is consistent in terms of quality aspects. This needs to be taken into consideration and it may be necessary to sample your data to ensure that the test dataset is representative of the range of image quality.
- Have sufficient identities (including mated pairs) and images to test with. This decision will be agency specific. This includes associated identity ground truth information.

5.3.2 The data set used for this document is the LFW (Labeled Faces in the Wild) data set available at: http://vis-www.cs.umass.edu/lfw/ See section “LFW Data Set Information” for more details in referenced document [2]. Conceptually any other facial data set with identity ground truth can be used.

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

- Smaller but consistent image sizes and file formats
- Over 5,700 identities and over 13,000 images
- Has a wide range of subjects: sex, pose, lighting, etc.
- Stated identity ground truth errors

6. Procedure

6.1 Yellow Resolve Determination Process

6.1.1 Ensure the data set to use has verified ground truth and that any manual localization to the facial images was performed.

6.1.2 Enroll the facial images into a facial gallery for searching.

6.1.3 Search all the facial images against the facial gallery (NxM). For this test 50 candidates were returned for each search, but 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 apply a score threshold for this test phase.

6.1.4 Plot the mated (FRR) and non-mated scores (FAR)

- Determine an approximate high score imposter from the FAR. This should be the highest true imposter score plus some small percentage added for safety.
- Determine a low mate score by analyzing the FRR. There are numerous ways to do this. For this document the following approach is being demonstrated.
- Determine the approximate equal error rate where the FRR score crosses the FAR. This score should be lowered by some small percentage added for safety.
- The yellow resolve scoring range is therefore slightly higher than the highest scoring imposter and slightly lower than the equal error rate score.
- Determine how many searches have a candidate score which is in the yellow resolve range. This represents the number of “searches” which would need manual review.
- Determine how many images have a true mate lower score than the yellow resolve range. This represents the number of “mates” which would have been missed.

6.2 Process Outcomes

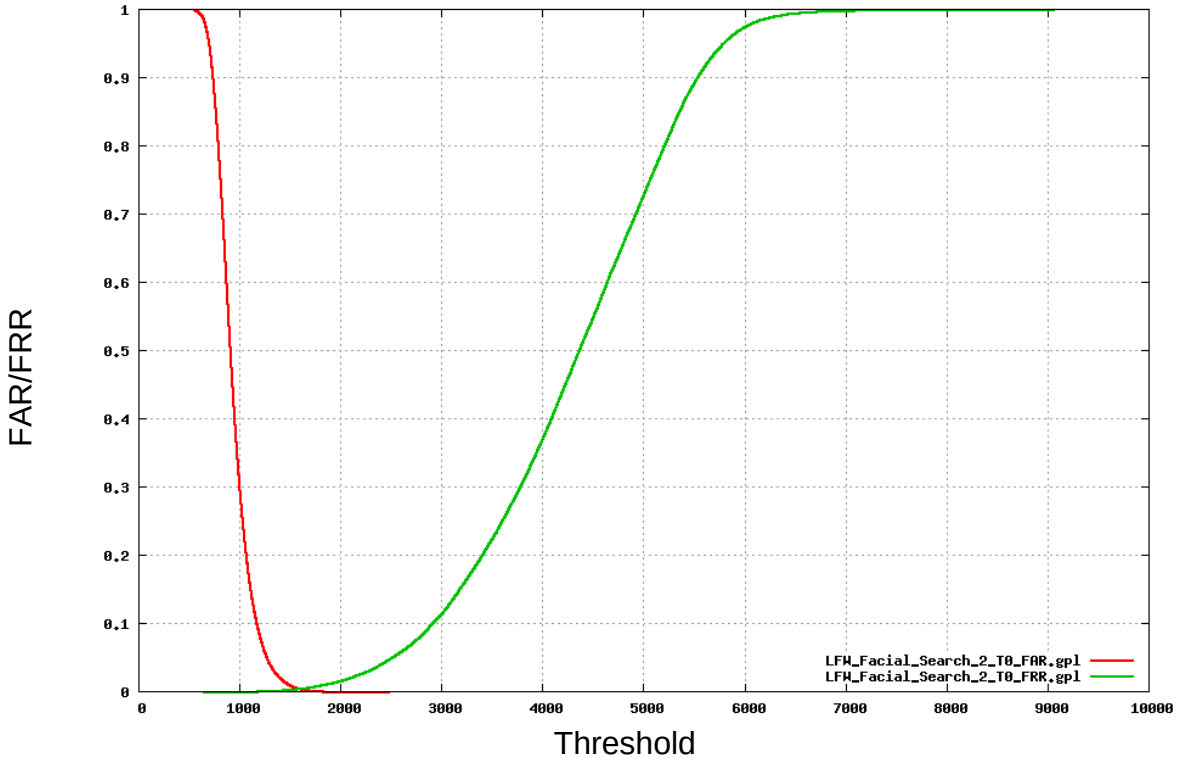

Figure 3: FAR/FRR Scoring Linear Axis

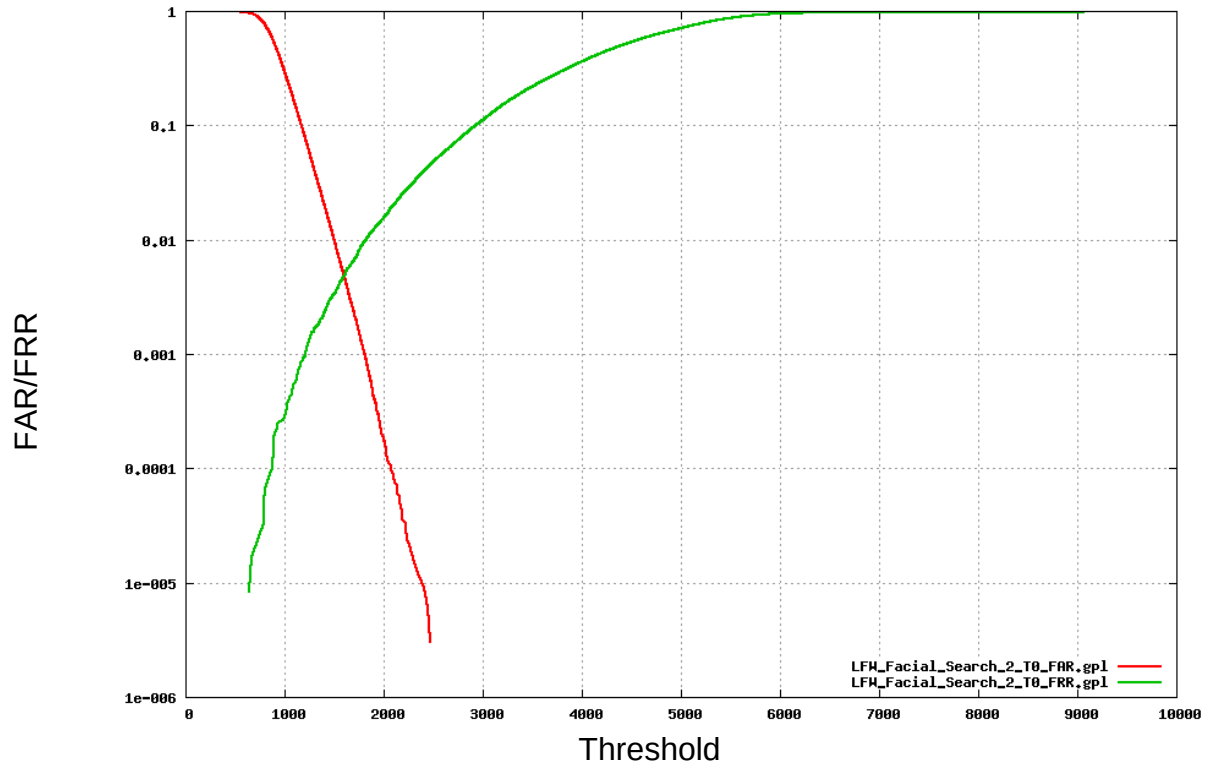

Figure 4: FAR/FRR Scoring Logarithmic Axis

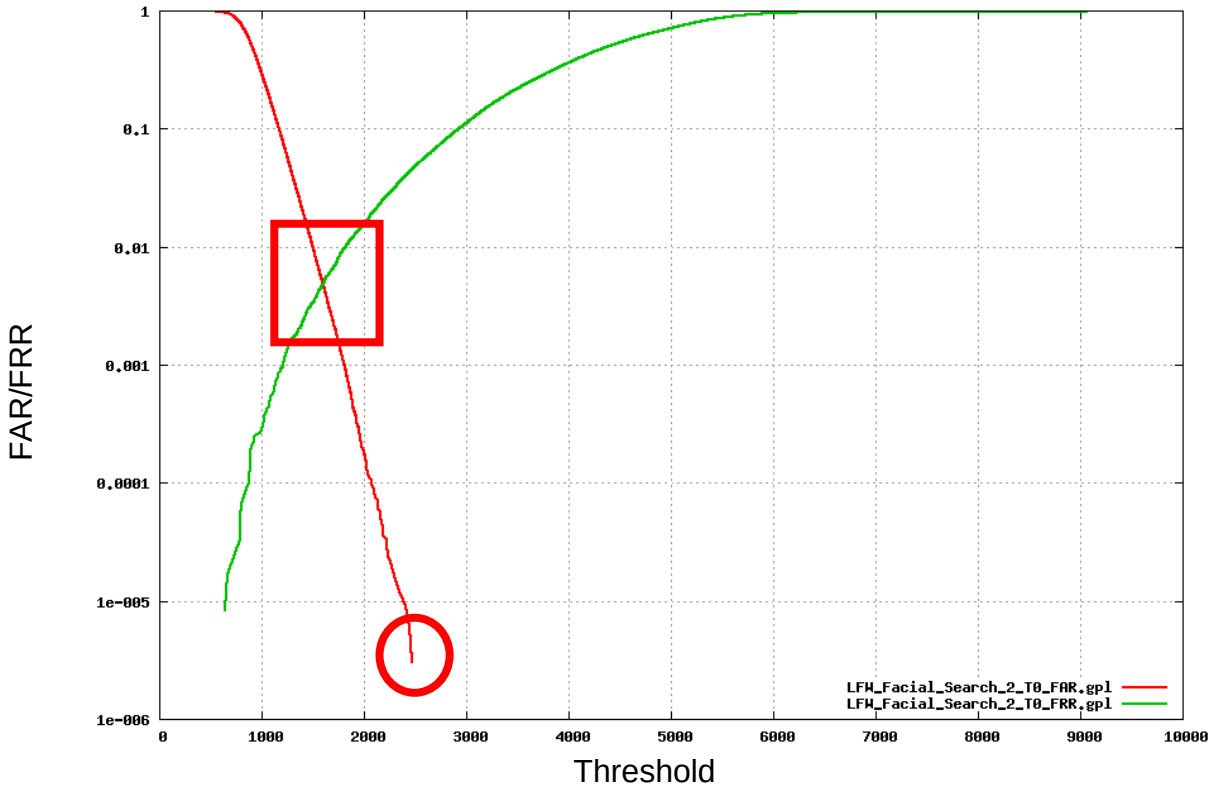

Figure 5: FAR/FRR Scoring Logarithmic Axis

Notes:

- The highest FAR score (red circle) is ~2460.
- The FAR/FRR crossover score (red square) is ~1550.
- So given these values a yellow resolve range could be assumed to be:
- Low yellow: 1395 (1550 – 10%)
- High yellow: 2700 (2460 + 10%)
- This score range can then be applied to all the search results given the searches with 50 candidates.
- Based on the sample set if we set the score range of 1395 to 2700 the results were as follows:
- 250 searches were in the yellow resolve range of 1395 and 2700
- 289 mates were missed with a score less than 1395
- However with a tighter score range of 1550 to 2460 the results showed many more mates were missed:

- 136 searches were in the yellow resolve range of 1550 and 2460
- 512 mates were missed with a score less than 1550
- The yellow resolve score range can be modified to gauge how many searches would need manual adjudication by human practitioners versus how many false positives and false negatives occurred.

6.3 Outcomes

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

- FAR and FRR curves were utilized in these processes.
- Analyzing the behavior and interactions of the FAR and FRR is an effective evidence-based method to determine yellow resolve scoring thresholds
- Once yellow resolve thresholds are determined the number of searches which would need manual review and the number of missed mates can be compared.
- The yellow resolve thresholds can then be modified to refine how the agency can balance manual reviews versus missed mates.