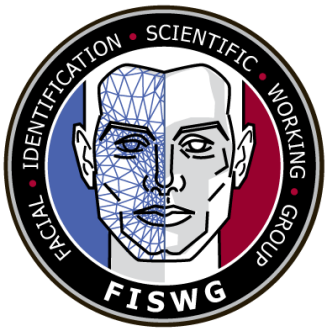


Disclaimer:

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.

Version 2.0 2026.06.05

Introduction to Face Recognition Technology Evaluation (FRTE) Testing

1. Scope

1.1 This document is intended to give an operational perspective on how biometric performance test reports can be created, interpreted, and applied to agencies which operate or intend to deploy an FRS (Facial Recognition System). The commonly used facial biometric accuracy charts of Receiver Operating Characteristic (ROC), Detection Error Tradeoff (DET), and Cumulative Match Characteristics (CMC) are explained using simplistic definitions and examples.

1.2 This document references National Institute Standards and Technology (NIST) as an authoritative testing agency; other agencies may do similar testing. This document also presents how NIST FRTE tests can be applied to agency FRS systems and where NIST FRTE tests may not be directly relevant to operational systems.

1.3 The intended audience of this document is agency decision makers and operational facial examiners. This document focuses on how these communities can use the NIST FRTE tests to:

1.3.1 Understand commonly used terms from these tests.

1.3.2 Gain practical knowledge of how these tests are reported.

1.3.3 Understand how these tests can add value for current or future FRS procurements.

1.3.4 Enhance or augment understanding of an operational FRS – including understanding strengths and limitations.

1.3.5 Establish the limits of the tests in terms of facial biometric technology.

1.3.6 Generate realistic expectations about what can be derived from these tests.

2. Referenced Documents

2.1 NIST

ANSI/NIST-ITL-1-2011 Update 2015: Data Format for the Interchange of Fingerprint, Facial and Other Biometric Information[1]

Face Recognition Technology Evaluation (FRTE) Part 2: Identification[2]

Receiver Operating Characteristic[3]

3. Terminology

3.1 Definitions:

[1] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-290e3.pdf

[2] https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf

[3] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

3.1.1 Imposter pair*, n*—a pair of images which are not of the same identity (non-mate)

3.1.2 Genuine pair, *n*—a pair of images which are of the same identity (mate)

3.1.3 Ground truth, *n*—where all imposter pairs and genuine pairs have been verified to be correct

3.2 Acronyms:

3.2.1 CMC, *n*—Cumulative Match Characteristic

3.2.2 DET, *n*—Detection Error Tradeoff

3.2.3 FMR, *n*—False Match Rate

3.2.4 FNMR, *n*—False Non-Match Rate

3.2.5 FRS, *n*—Facial Recognition System

3.2.6 ROC, *n*—Receiver Operating Characteristic

3.2.7 TMR, *n*—True Match Rate

3.2.8 TPIR, *n*—True Positive Identification Rate

4. Summary of Practice

4.1 NIST has conducted the FRTE program since the mid-1990s. This testing and others similar to NIST FRTE help to improve the accuracy of facial biometrics and are the gold standard in how to evaluate performance of facial biometric algorithms. The

NIST FRTE evaluations have shown that commercially available facial biometric algorithmic search performance has dramatically improved over the last thirty years.

4.2 Understanding the NIST FRTE reports can be daunting for someone not versed in the style of FRTE reports and the terms and charts used in these reports. This is the focus of this document.

5. Significance and Use

5.1 Simplistic Test Scenario

5.1.1 A facial algorithm that generates a score within the range of zero to 100. Zero is the lowest similarity score possible while 100 is the highest similarity score possible when doing a search or verification.

5.1.2 A facial collection of 1000 image pairs of known mates and a collection of 1000 image pairs of imposters. The images are passport quality frontal poses.

5.1.3 The 2000 (i.e., 2 x 1000 images) pairs are enrolled in a facial gallery. There are no failures to enroll, and all facial images are verified to be properly enrolled.

5.2 Ideal Scoring

5.2.1 Assume 1000 sets of genuine pair (mates) images are compared to each other and then 1000 sets of imposter pair (non-mates) images are compared to each other. All the genuine pair comparisons yield a score of 100 (green), and all the imposter pairs yield

a score of 0 (red). These are plotted as scores on the X-axis and the count of genuine pair and imposter pair comparisons on the Y-axis.

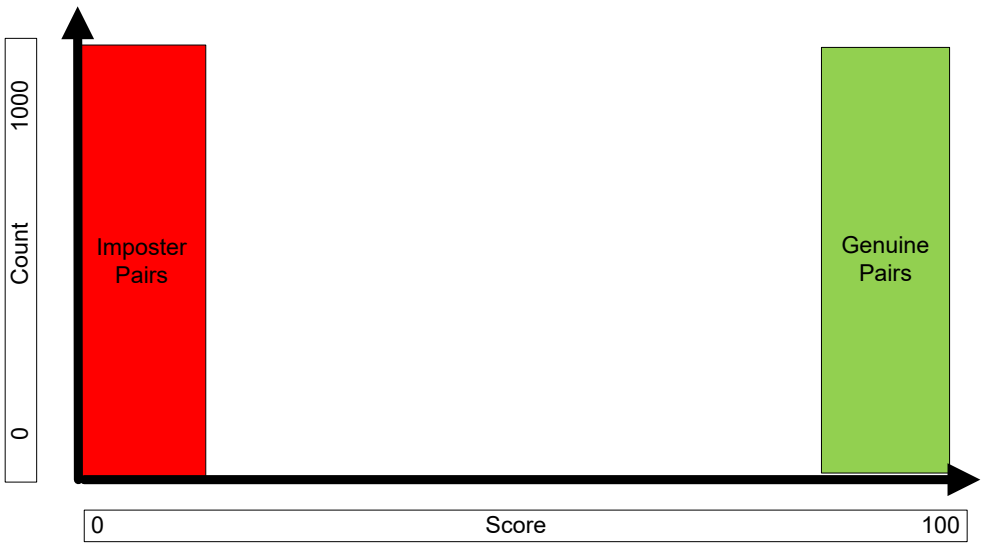


Figure 1: Ideal Scoring

5.2.2 Figure 1 presents all the imposter pairs stacked in the red section with a score of 0 and all the genuine pairs stacked in the green section with a score of 100. This can be regarded as ideal algorithm performance and scoring.

5.3 Facial biometric algorithms will rarely produce a similarity score of 0 for an imposter and 100 for a genuine for every biometric comparison. These comparisons will normally produce similarity scores in the range of 0 to 100. Similarity scores of 0 and 100 may be produced, but usually scores will range between 0 to 100 with higher scores indicating a stronger similarity and lower scores exhibiting weaker similarity.

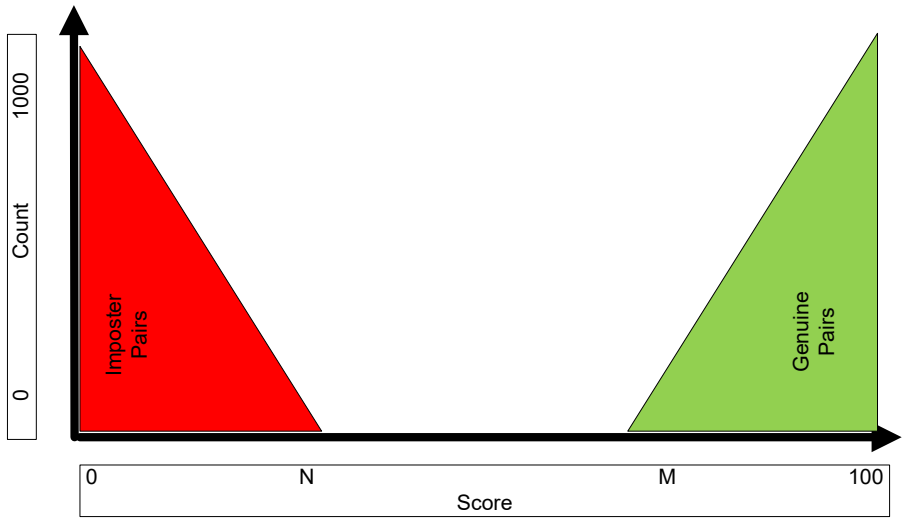


Figure 2: Excellent Scoring

5.3.1 Figure 2 presents all the imposter pairs included in the red section with a score of 0 to N and all the genuine pairs included in the green section with a score of M to 100. N is the highest scoring imposter pair and M is the lowest scoring genuine pair. These are still excellent results because there is no crossover in scoring between the genuine and imposter pair scores.

5.4 Operational Scoring

5.4.1 The results presented in Figure 3 show some imposter pairs producing relatively high scores and some genuine pairs producing relatively low scores. As a result, these score ranges between the M (lower score genuine pairs) and N (higher score imposter pairs) scores now overlap.

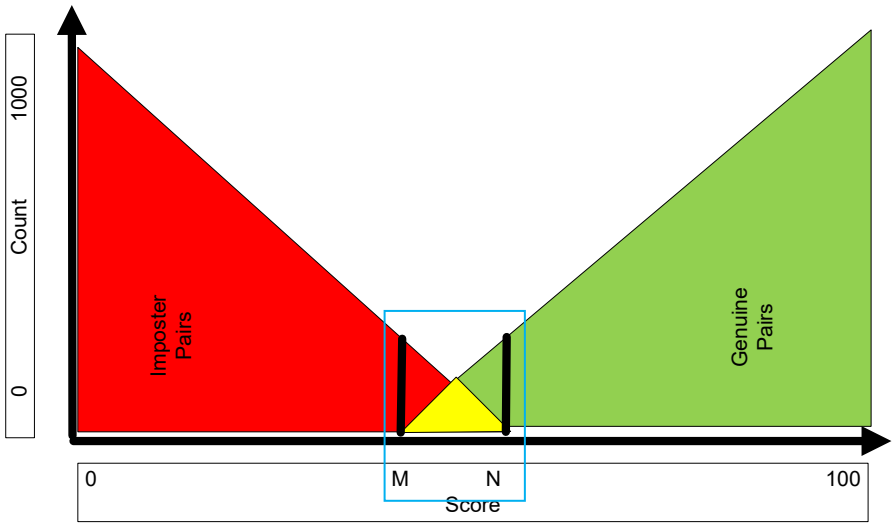


Figure 3: Operational Scoring

5.4.2 Figure 3 shows the imposter pairs included in the red section with a score of 0 to N and the genuine pairs included in the green section with a score of M to 100. The ambiguity present here is that the highest imposter score N is greater than the lowest genuine score M. This area of ambiguity is usually called “manual resolve” since the comparisons in this area need to be manually reviewed by a facial practitioner to determine a true genuine or imposter.

5.5 Manual Resolve

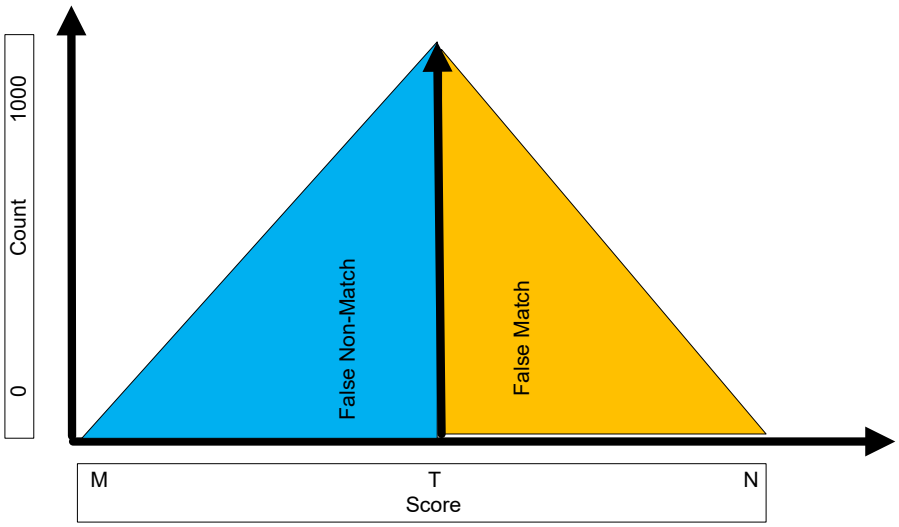


Figure 4: Manual Resolve Area

5.5.1 Figure 4 displays the impact of setting a threshold score (T) in the manual resolve area of interest. In an automated FRS any pair of images that achieves a score above the threshold value T could be considered a match regardless of whether the pair of images is a genuine pair or an imposter pair. Likewise, any pair of images that achieves a score below the threshold value T could be considered a non-match regardless of whether the pair of images is an imposter pair or a genuine pair. The blue area represents False Non-Match where true genuine pairs have scores below the score T. The orange area represents False Match where an imposter pair comparison has a score greater than the score T.

5.6 The manual resolve area (i.e., the range between M and N) is separated from the true match area (genuine pairs) and the true non-match (imposter pair) score ranges.

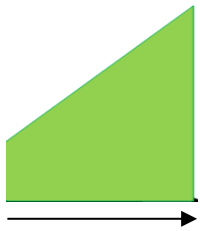
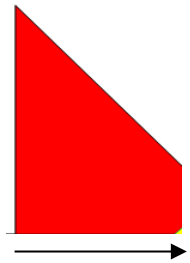
This is the true match area. Notice this area represents genuine pairs with a score when no imposter pairs are present. Notice that N is the highest score of a false match (imposter pair).	
This is the true non-match area. Notice this area represents imposter pairs with a score where there are no genuine pairs present. Notice that M is the lowest score of a false non-match (genuine pair).	

Figure 5: True Match and Non-match Areas

5.6.1 The examples shown in the sections above are core to the FRTE evaluations. These can be summarized as a collection of facial images comprising genuine pairs and imposter pairs used to derive similarity scores. These similarity scores are then used to derive scoring profiles which show the following areas:

- True match (green area) also called genuine pairs
- True non-match (red area) also called imposter pairs
- False Match (orange area)
- False Non-Match (blue area)

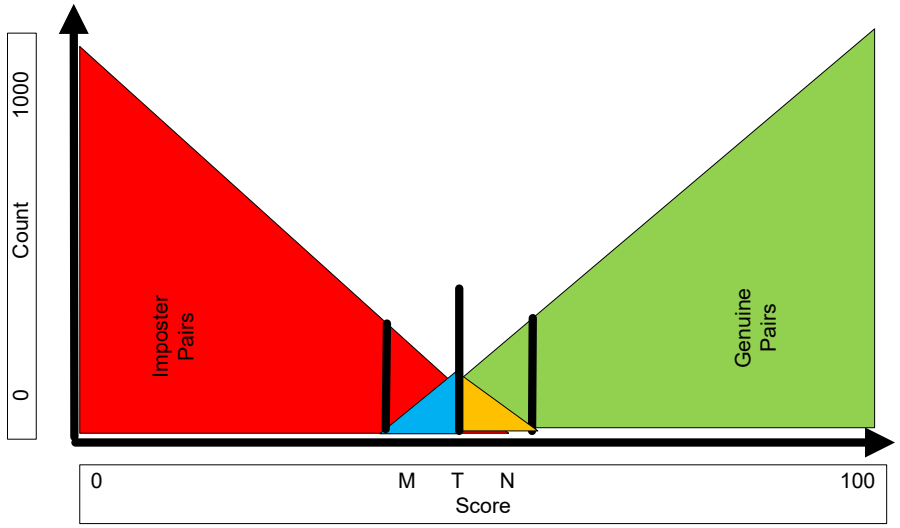


Figure 6: Operational Scoring

5.7 The examples shown in the sections above assume verified ground truth data is present and is then used to measure system accuracy. If the ground truth of operational data sets selected for this testing is unknown and may contain errors, this will impact the simplicity intended to be presented here.

5.8 The four areas depicted in Figure 6 are usually represented in a series of charts that NIST creates from the tests they conduct. NIST has many more results that add detail to their tests, but these are the four scoring profiles key to the core accuracy issues in evaluating facial biometric recognition algorithms.

6. Basic Accuracy Charts

6.1 Similarity Score Distributions: Figure 7 shows the scores of all comparisons plotted as genuine pair and imposter pair scores. This is the same as shown in Figure 6 but reflects a more realistic distribution of scores.

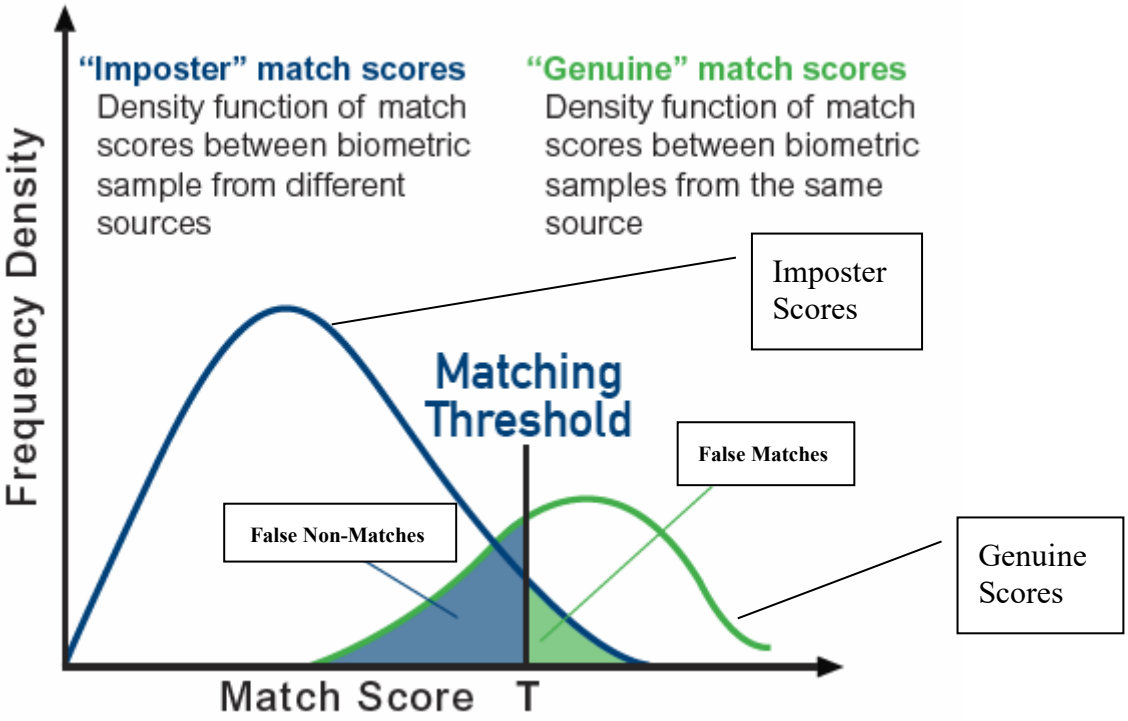


Figure 7: Similarity Score Distributions

6.1.1 This illustrates how genuine and imposter similarity scores are distributed in practice.

- The genuine pair scores are concentrated to the right (high score)
- The imposter pair scores are concentrated to the left (low score)
- The middle manual resolve area shows False Matches and False Non-Matches

6.2 ROC curve

6.2.1 This type of plot compares the True Match Rate (TMR) versus the False Match Rate (FMR). The ROC curve aggregates statistics of the TMR to the FMR. The performance of a verification system is summarized using the ROC curve.

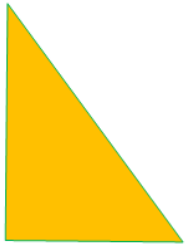
This is the true match area. Notice this area represents genuine pairs with a score when no imposter pairs are present. Notice that left edge of the area is the highest score of an imposter pair.	
This is the false match area. Notice this area represents imposter pairs with a score above the threshold which differentiates between a genuine and an imposter pair.	

Figure 8: ROC Curve Score Components

6.2.2 These areas are then plotted with an algorithm that varies the threshold score and creates the percentages at each score.

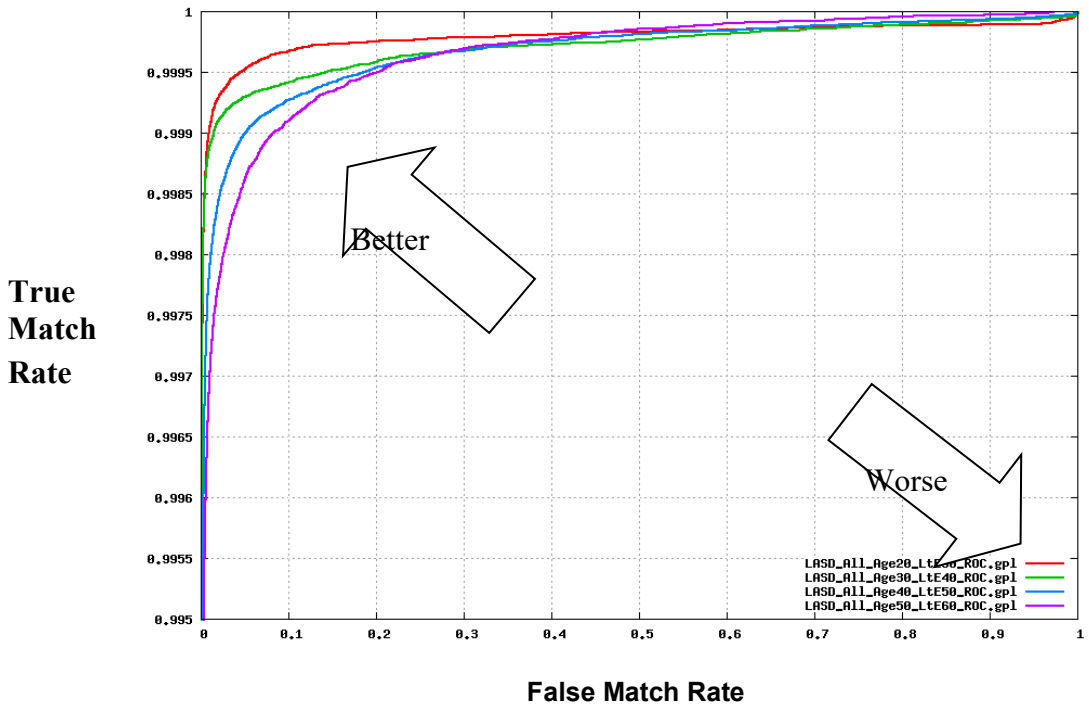


Figure 9: Multiple ROC Curves

6.3 Figure 9 Illustrates ROC curves for multiple algorithms, each showing the relationship between FMR and TMR across varying threshold settings. Curves that trace closer to the top-left border indicate higher accuracy and better discrimination between genuine and imposter pairs, while those nearer the diagonal represent weaker performance

6.4 A ROC curve demonstrates several things:

6.4.1 It shows the tradeoff between False Match Rate and True Match Rate: any increase in True Match Rate will be accompanied by a decrease in False Match Rate.

6.4.2 The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

6.4.3 The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

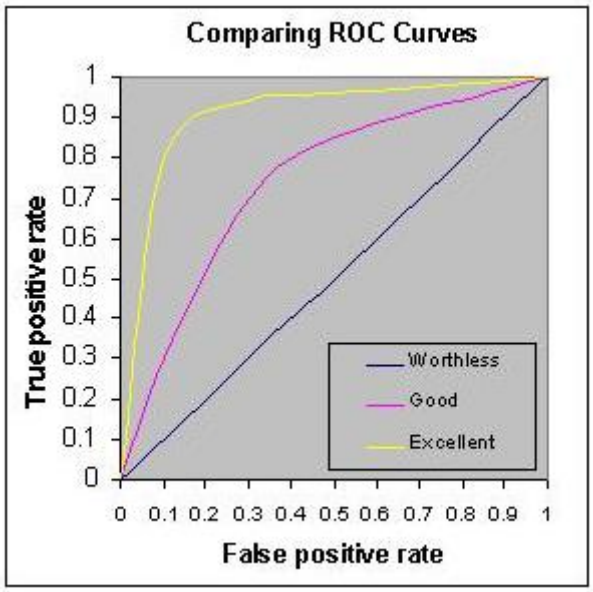
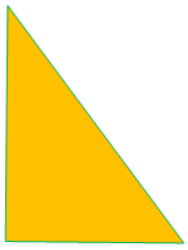


Figure 10: ROC Curves

6.5 Detection Error Tradeoff (DET) curve

6.6 This type of plot compares the False Non-Match Rate (FNMR) versus the FMR. These areas are then plotted with an algorithm that varies the threshold score and creates the percentages at each score.

This is the false match area. Notice this area represents imposter pairs with a score above the threshold score which differentiates between a genuine pair and an imposter pair	

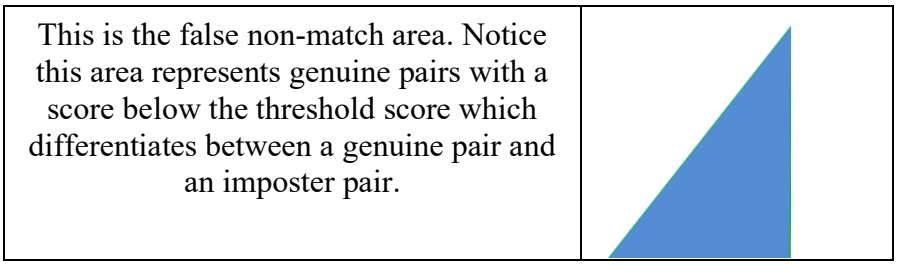


Figure 11: DET Curve Scoring Regions

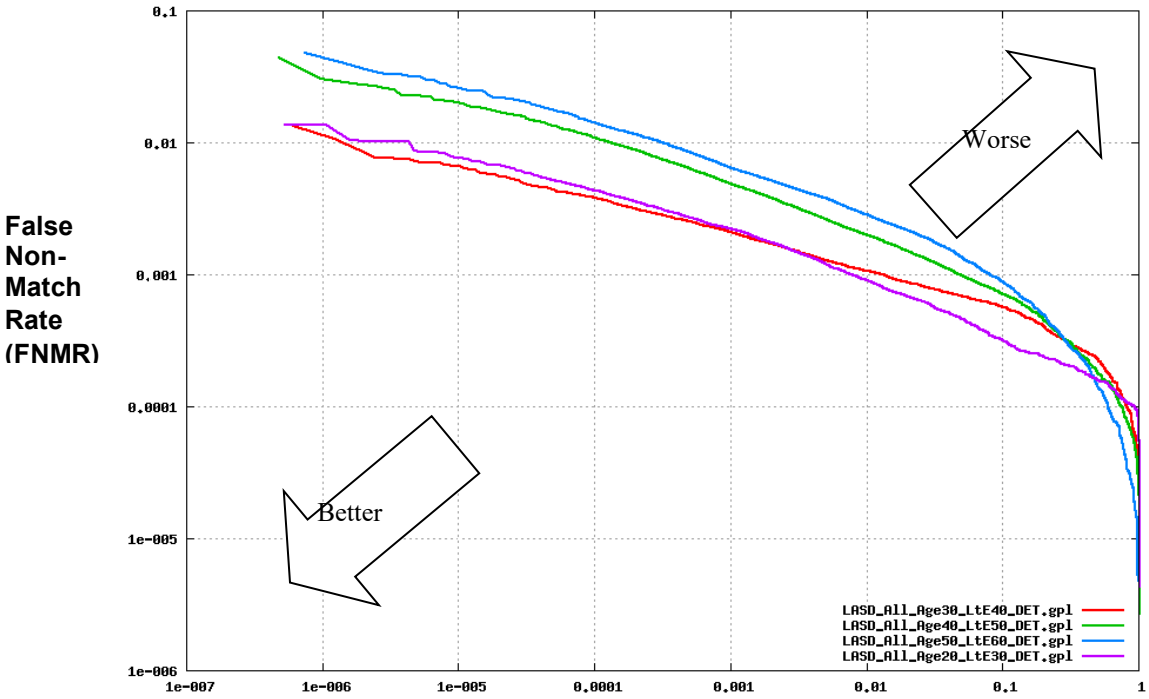


Figure 12: Various DET Curves

6.7 A DET curve demonstrates several things:

6.7.1 It shows the tradeoff between the FMR and FNMR for all threshold values.

6.7.2 The closer the curve follows the bottom border and the flatter the curve is the more desirable result.

6.7.3 The higher the curve is from the bottom border, and the higher the curve increases to the left-hand border, the less accurate the test.

6.8 CMC curve.

6.9 Performance of a closed-set identification system (where every search is known to contain a genuine pair in the database) is summarized using a CMC curve. This plots the True Positive Identification Rate (TPIR) against the candidate rank by displaying the probability of observing the correct identity within the top search candidate ranks.

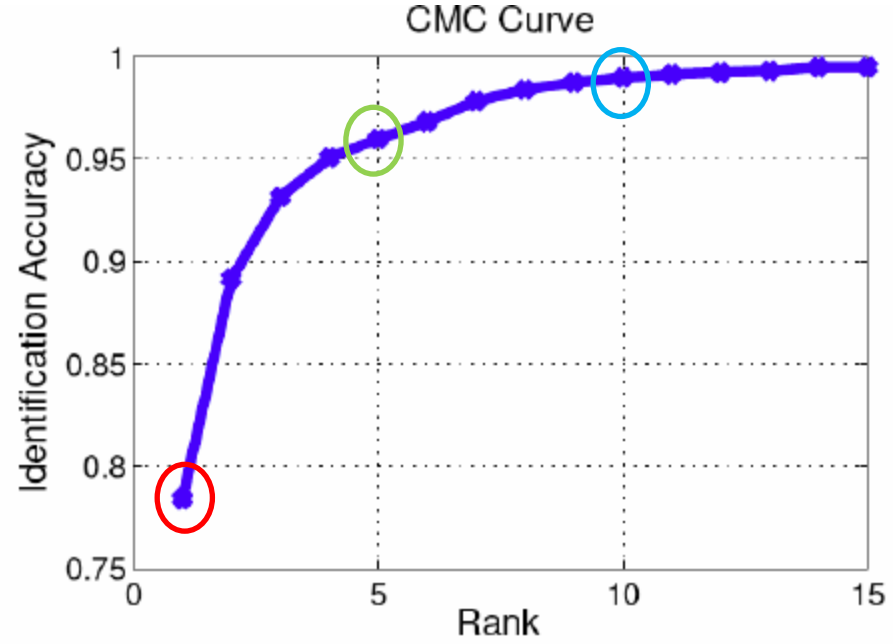


Figure 13: CMC Curve

6.9.1 In Figure 13 above:

- The chance of a genuine mate in the candidate results at **rank 1 is ~77%.**
- The chance of a genuine mate in the candidate results at **rank 5 is ~96%.**

- The chance of a genuine mate in the candidate results at **rank 10 is ~99%.**

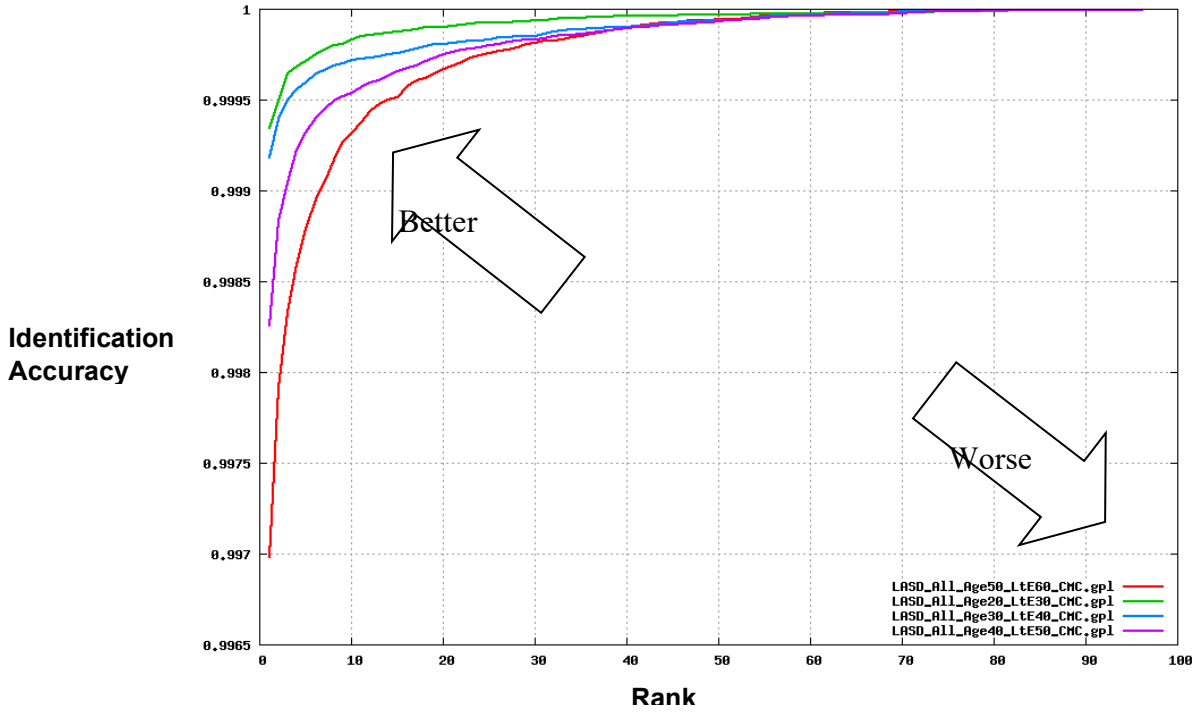


Figure 14: Various CMC Curve

6.10 A CMC demonstrates several things.

6.10.1 The CMC is used as a measure of 1: N identification system performance. It judges the ranking capabilities of an identification system.

6.10.2 The closer the curve follows the top left corner and the top border of the CMC curve, the more accurate the test.

7. Overall Testing Process

7.1 To generate the DET, ROC, and CMC charts used by FRTE the following process steps can be summarized.

7.2 Biometric samples with known ground truth are compared against each other:

- Each probe biometric sample is compared against all gallery samples
- Genuine and impostor scores are generated
- The resulting scores are sorted and ranked

7.3 TMR and FNMR are computed at multiple thresholds

7.4 Create and plot the ROC Curve

7.5 FMR and FNMR are computed at multiple thresholds

7.6 Create and plot the DET Curve

7.7 Determine the candidate ranks at which a true match occurs

7.8 Create and plot the CMC Curve

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