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 read 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 Characteristic (CMC) are explained using simplistic definitions and examples.

1.2 This document uses the National Institute of Standards and Technology (NIST) as an authoritative testing agency; other agencies may do similar testing. Several opinions are also presented on how NIST Face Recognition Technology Evaluation (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 of future FRS procurements.

1.3.4 Enhance or augment understanding of an operational FRS

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) 1:N Identification[2]

Introduction to ROC Curves[3]

Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification[4]

3. Terminology

[1] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-290e3.pdf
[2] https://pages.nist.gov/frvt/html/frvt1N.html
[3] http://gim.unmc.edu/dxtests/roc2.htm
[4] https://www.nist.gov/system/files/documents/2017/08/07/frvt_report_2017_08_07.pdf

3.1 *Definitions:*

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

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

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

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 *FNR, n*—False Negative Rate

3.2.6 *FPR, n*—False Positive Rate

3.2.7 *FRS, n*—Facial Recognition System

3.2.8 *ROC, n*—Receiver Operating Characteristic

3.2.9 *TMR, n*—True Match Rate

3.2.10 *TPIR, n*—True Positive Identification Rate

4. Summary of Practice

4.1.1 National Institute of Standards and Technology (NIST) has conducted the Facial Recognition Technology Evaluation (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 are one of the primary reasons why 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 Assume the following:

5.1.1.1 A facial algorithm that generates a score within the range of zero to 10. Zero is the lowest similarity score possible while 10 is the highest similarity score possible when doing a search or verification.

5.1.1.2 A facial collection of 1000 image pairs of known mates. The images are passport quality frontal poses.

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

5.2 Perfect Scoring

5.2.1 Assume you compare 1000 sets of genuine pair (mates) images to each other and then compare 1000 sets of imposter pair (non-mates) images to each other. The genuine pair comparisons yield a score of 10 (green) and the imposter pairs yield a score of 0 (red). These are plotted as scores on the X-axis and count of genuine pair / imposter pair comparisons on the Y-axis.

Figure 1: Perfect Scoring: This illustration shows idealized separation between imposter pairs and genuine pairs.

5.2.2 Figure 1 shows 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 10. This can be regarded as perfect algorithm performance and scoring since the imposter pairs all scored 0 and the genuine pairs all scored 10.

5.3 Excellent but not Perfect Scoring

5.3.1 Facial biometric algorithms will rarely produce perfect scoring as described above. Rather the comparisons will produce similarity scores in the range of 0 to 10. Similarity scores of 0 and 10 may be produced, but usually scores will range between 0 to 10 with higher scores indicating a stronger similarity and lower scores exhibiting less similarity (see Figure 2).

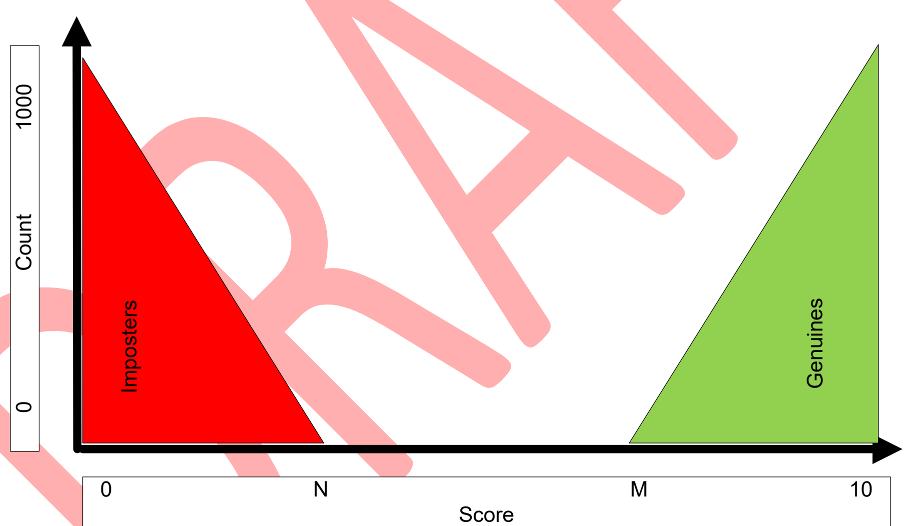


Figure 2: Excellent, but not perfect, scoring: Example of high-performing results where genuine and imposter pair distributions remain fully separated.

5.3.2 Figure 2 shows 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 10. N is the highest score for an imposter pair, and M is the lowest score for a genuine pair.

These are 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 now overlap.

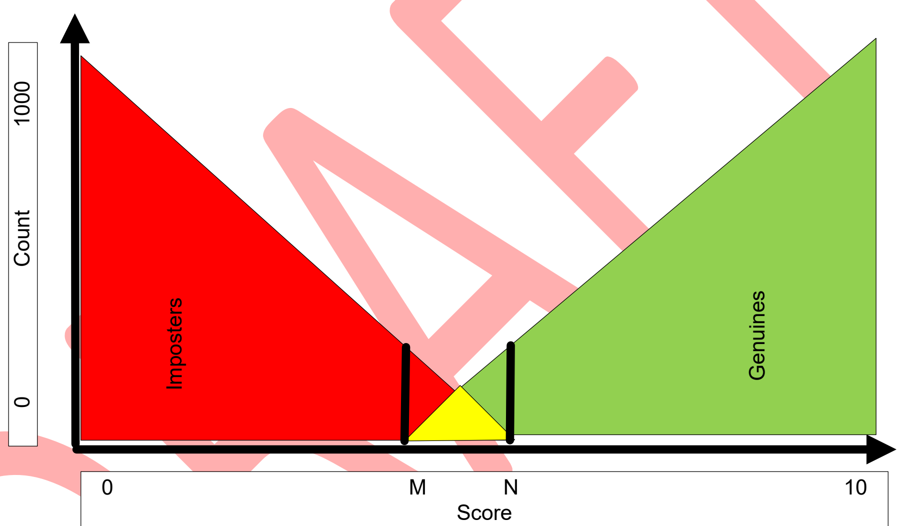


Figure 3: Operational Scoring: Depicts a realistic distribution where imposter and genuine pair scores overlap. The region of overlap represents scores that may require manual review to determine whether they correspond to genuine or imposter pairs.

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 10. The issue here is that the highest imposter pair score N is greater than the lowest genuine pair score M. This area of interest is usually called “manual resolve” since this area needs to be manually reviewed by a practitioner.

5.5 The Areas of Interest

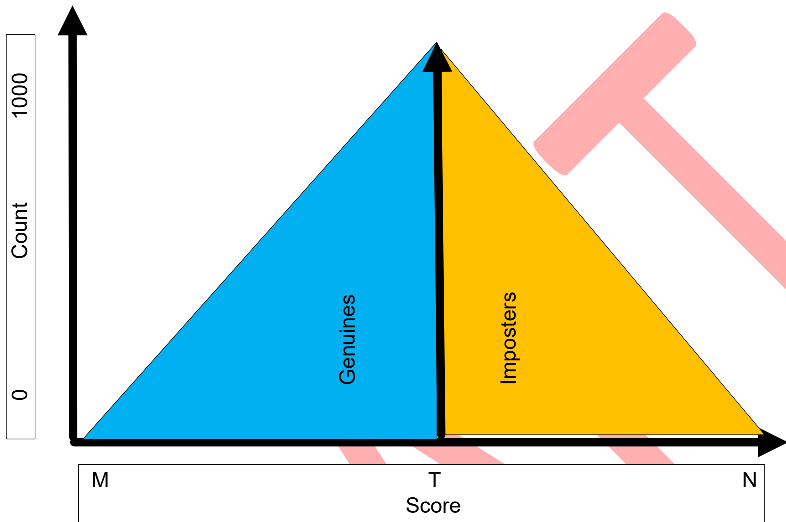
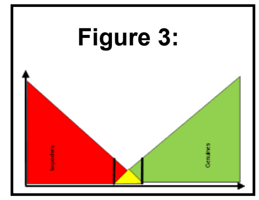


Figure 4: Areas of Interest: Shows how a decision threshold (T) divides the overlap region into false accepts and false rejects.

5.5.1 Figure 4 displays the impact of setting a threshold score (T) in the area of interest. In an automated process any pair of images that achieves a score above the threshold value T will be declared a match regardless of whether the pair of images is a genuine pair or not. Likewise, any pair of images that achieves a score below the threshold value T will be declared a non-match regardless of whether the pair of images is an imposter pair or not. The blue area represents false rejects where true genuine pairs have scores below the score threshold. The orange area represents false accepts where an imposter pair comparison has a score greater than the score threshold.

5.6 At this stage the manual resolve area (i.e., the range between M and N) is separated from the genuine and imposter areas.



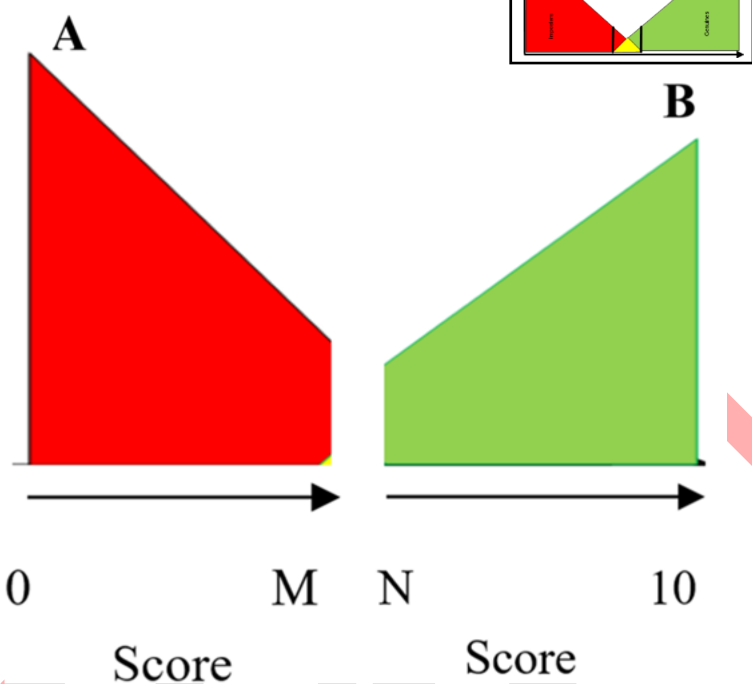


Figure 5: Excerpt from Figure 3: Non Match (A) and True Match (B) Areas.

5.6.1 The examples shown in Figures 4 and 5 are core to the FRTE evaluations. The two areas illustrated in Figure 5 represent, respectively, the non match (A) and true match (B) areas. The non match area (A) represents imposter pairs with a score where there are no genuine pairs present. Notice that M is the lowest score of a false reject. The true match area (B) represents genuine pairs with a score where there are no imposter pairs present. Notice that N is the highest score of a false accept. The examples shown can be summarized as a collection of facial images comprised of genuine and imposter pairs that are used to derive similarity scores. These similarity scores are then used to derive scoring profiles which are shown in Figure 6:

5.6.1.1 True positive (green area) also called genuine

5.6.1.2 True non match (red area) also called imposter

5.6.1.3 False positive (orange area) also called “False Match” or FMR

5.6.1.4 False negative (blue area) also called “False Non-Match” or FNMR

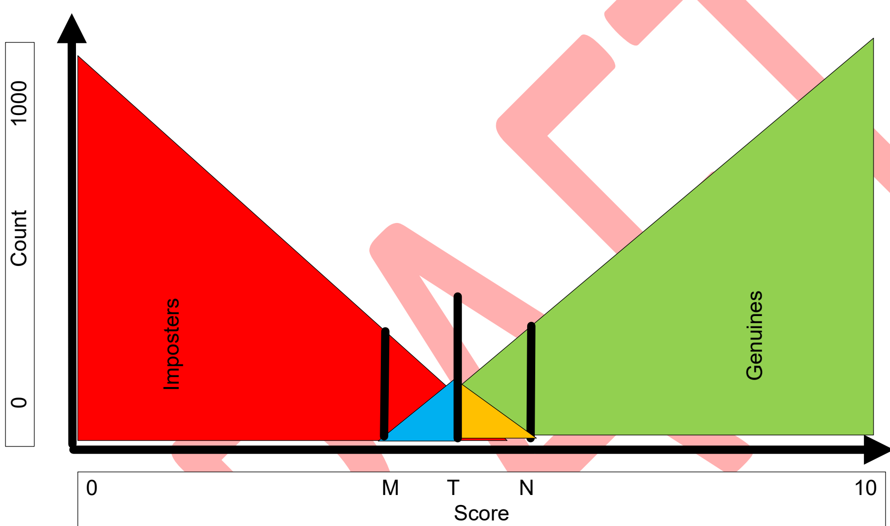


Figure 6: Operational Scoring: Combines the principal regions derived from similarity scoring: true matches (green), true non-matches (red), false positives (orange), and false negatives (blue). These categories form the basis for most biometric accuracy metrics.

5.7 The examples shown in the sections above assume known ground truth data is used to measure system accuracy. The ground truth of operational data sets is unknown and may incorporate errors, which 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 which NIST creates from the tests they conduct. NIST has many more results which add detail to their tests, but these are the four scoring profiles key to the accuracy issue in facial biometric matching algorithms.

6. Basic Accuracy Charts

6.1 Similarity Score Distributions: This chart shows the scores of all comparisons plotted as genuine pair (mates) and imposter pair (non-mates) scores. This is the same as the colored charts shown in the previous section but reflects a more realistic distribution of scores.

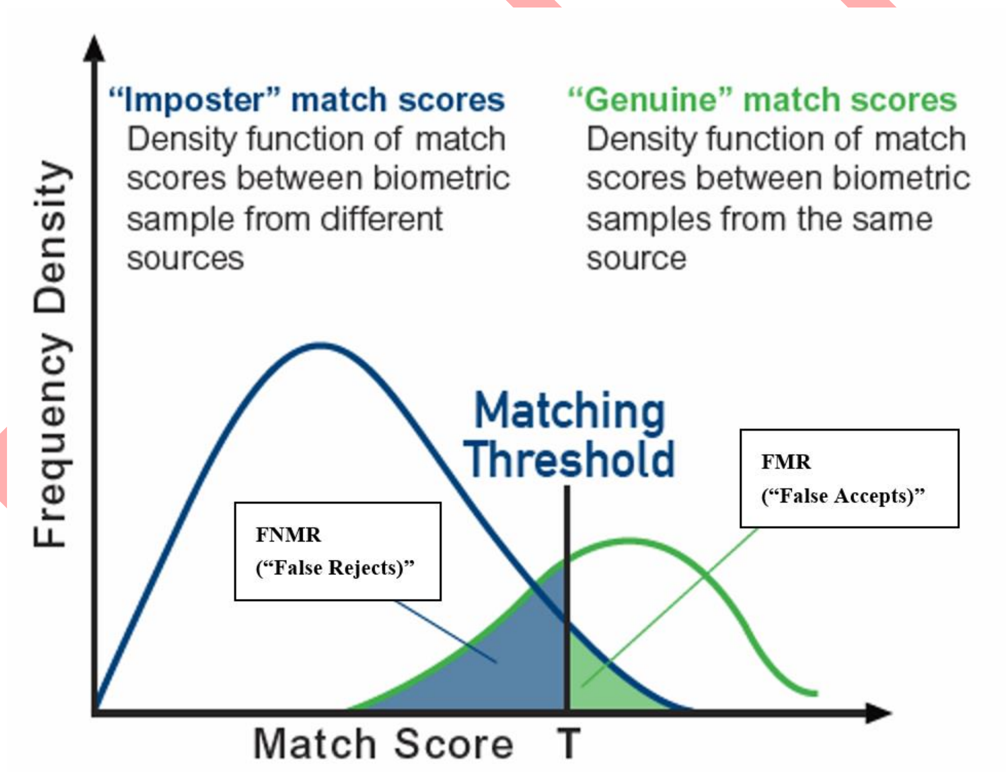


Figure 7: Match Score Distribution: Illustrates how genuine and imposter scores are distributed in practice. Genuine pairs cluster toward higher similarity scores, imposter pairs toward lower, and overlap defines the region where false accepts and false rejects occur.

6.2 Note:

6.2.1 The genuine pair (mates) scores are concentrated to the right (high score).

6.2.2 The imposter pair (non-mates) scores are concentrated to the left (low score).

6.2.3 The crossover areas false accept (FMR), and false rejects (FNMR) are in the middle.

6.2.4 To further explore the manual resolve area, consider the data in Table 1 and Figure 8. This data reflects the output of an idealized algorithm when the algorithm is presented with 100,000 imposter pairs and 100,000 genuine pairs. In this example, the idealized algorithm generates match scores between 0 and 100.

6.2.4.1 The data recorded in Table 1 offers an opportunity to explore the effect that setting different threshold values has on FMR and TMR. To aid this exploration, Figure 9 offers an enlarged view of the area of interest (center portion of Figure 8).

Match score	Number of Imposter pairs	Number of Genuine Pairs
0	10834	0
5	20000	0
10	20000	0
15	17500	0
20	15000	0
25	10000	0
30	5000	0
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000
70	0	5000
75	0	10000
80	0	15000
85	0	17500
90	0	20000
95	0	20000
100	0	10834

TABLE 1 – This table reports the number of imposter pairs and genuine pairs which returned a given match score for an idealized algorithm. Figure 8 shows the Match Score Distribution for this data.

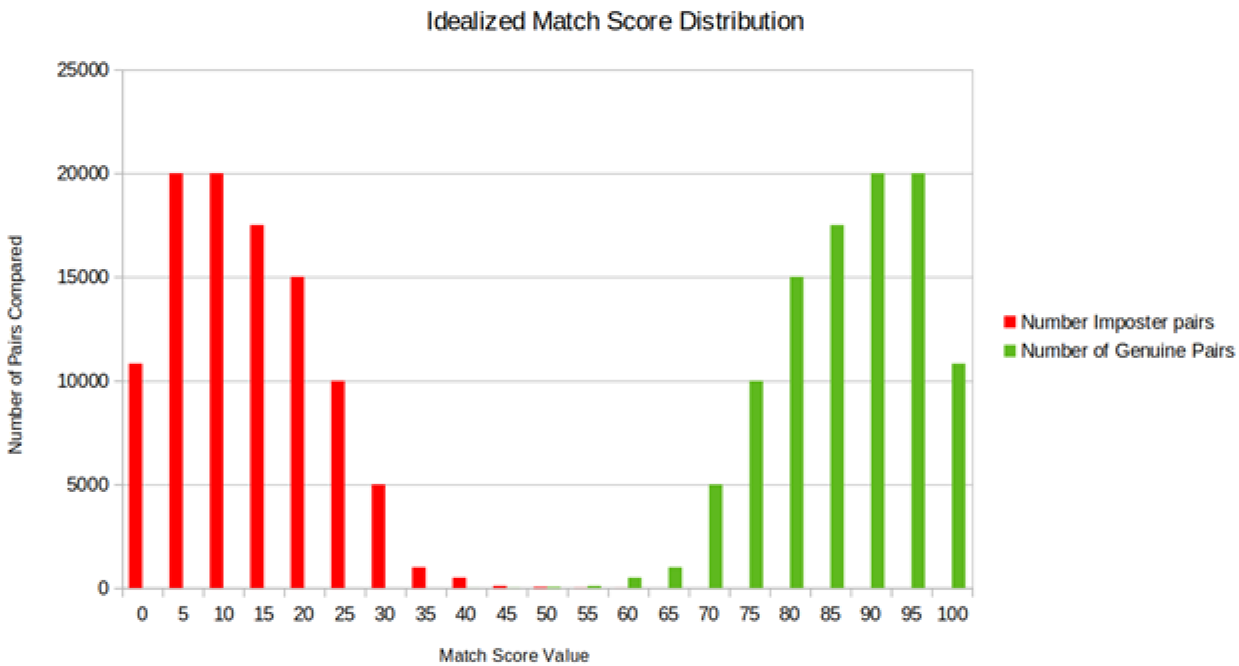


Figure 8 – Match Score Distribution for the data in Table 1.

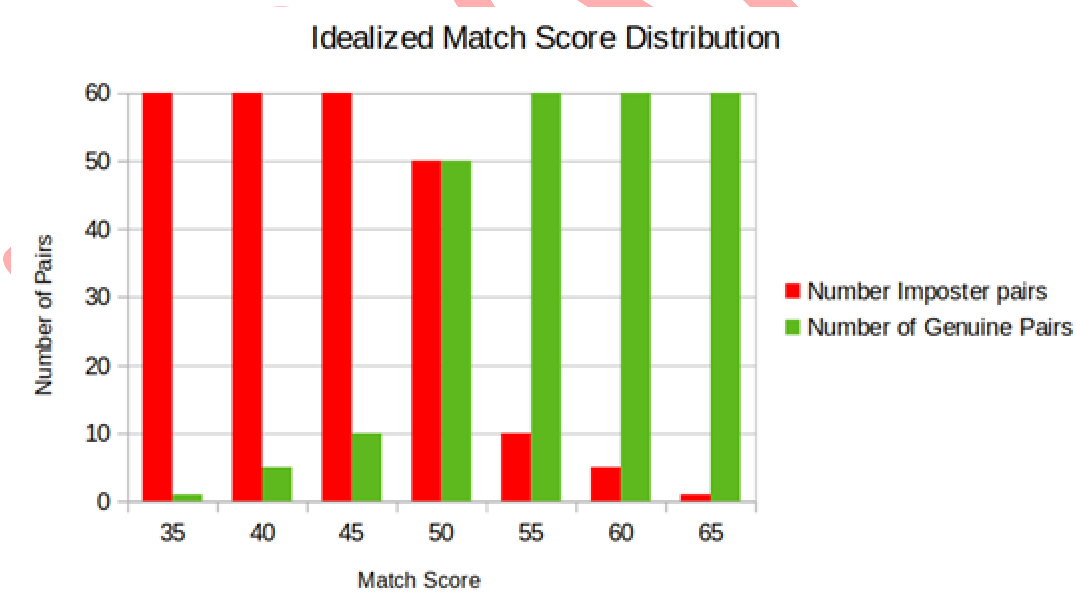


Figure 9 – The central portion of Figure 8, allowing one to visualize the small values present there.

6.2.5 As recorded in Table 1 and seen in Figures 8 and 9, the idealized algorithm generates symmetrical match score distributions, with 50 imposter pairs and 50 genuine pairs returning a match score of 50. Table 1 shows that while 100 imposter pairs return a match score of 45, only 10 genuine pairs return this value. Likewise, while 500 imposter pairs return a match score of 40, only 5 genuine pairs do so. At a match score of 35, 1000 imposter pairs occur, but only one genuine pair does so. No genuine pairs return a match score below 35.

6.2.6 If a threshold were set for a match score value of 35, then all pairs returning a match score of 35 or higher would be declared a “match.” In such a case, all 100,000 genuine pairs would be declared “match” (a true match rate or TMR of 100%) but so, too, would more than 1600 imposter pairs which return a match score at or above 35 (1666 imposter pairs, to be precise). Given that 100,000 imposter pairs were presented to the algorithm and 1666 were declared as “match”, the false match rate (FMR) for a threshold of 35 would be 1666/100000 or 1.666% (1.666 in 100).

6.2.7 The data in Table 1 allows one to calculate the changing values of TMR and FMR as the threshold varies. Table 2 documents this. One imposter pair returns a match score of 65, so at this threshold value, the FMR would be 0.001% or 1-in-100,000. This is the FMR NIST uses when reporting on many types of data, including mugshot data. At this FMR, the TMR would be 99.334%.

Match score	FMR	TMR
0	100.000%	100.000%
5	89.166%	100.000%
10	69.166%	100.000%
15	49.166%	100.000%
20	31.666%	100.000%
25	16.666%	100.000%
30	6.666%	100.000%
35	1.666%	100.000%
40	0.666%	99.999%
45	0.166%	99.994%
50	0.066%	99.984%
55	0.016%	99.934%
60	0.006%	99.834%
65	0.001%	99.334%
70	0.000%	98.334%
75	0.000%	93.334%
80	0.000%	83.334%
85	0.000%	68.334%
90	0.000%	50.834%
95	0.000%	30.834%
100	0.000%	10.834%

Table 2 – FMR and TMR when threshold is set to different match scores. As the number of false matches decreases, too does the number of true matches.

6.2.8 Table 2 shows that one can achieve higher TMRs by lowering the threshold, but this will result in higher FMRs as well. At a threshold of 60, the TMR and FMR values would increase to 99.834% and 0.006%, respectively.

6.3 ROC Curve: Receiver Operating Characteristic curve.

6.3.1 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.

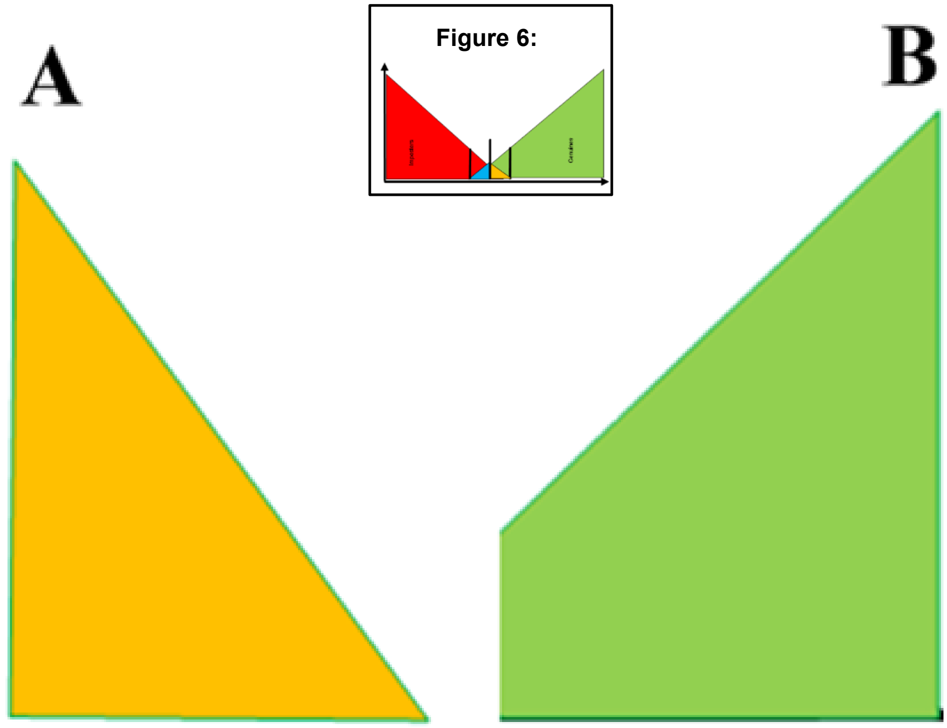


Figure 10: ROC Curve Score Components. A: Excerpt from Figure 6, this is the False Match area. This area represents imposter pairs with a score above the threshold which differentiates between a genuine and an imposter B: Excerpt from Figure 6, This is the True Match area. This area represents genuine pairs with a score where there are no imposter pairs present.   Notice that the left edge of the area is the highest score of a false accept.

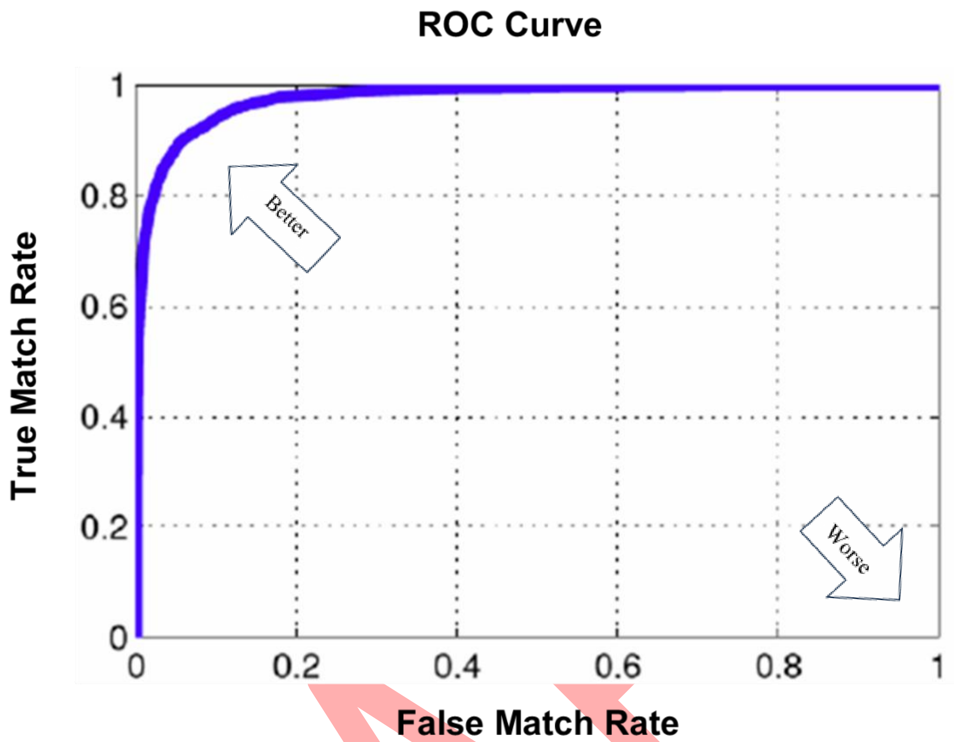


Figure 11: Simple ROC Curve: A simplified Receiver Operating Characteristic (ROC) curve showing the tradeoff between True Match Rate (TMR) and False Match Rate (FMR). Curves closer to the top-left corner indicate more accurate systems.

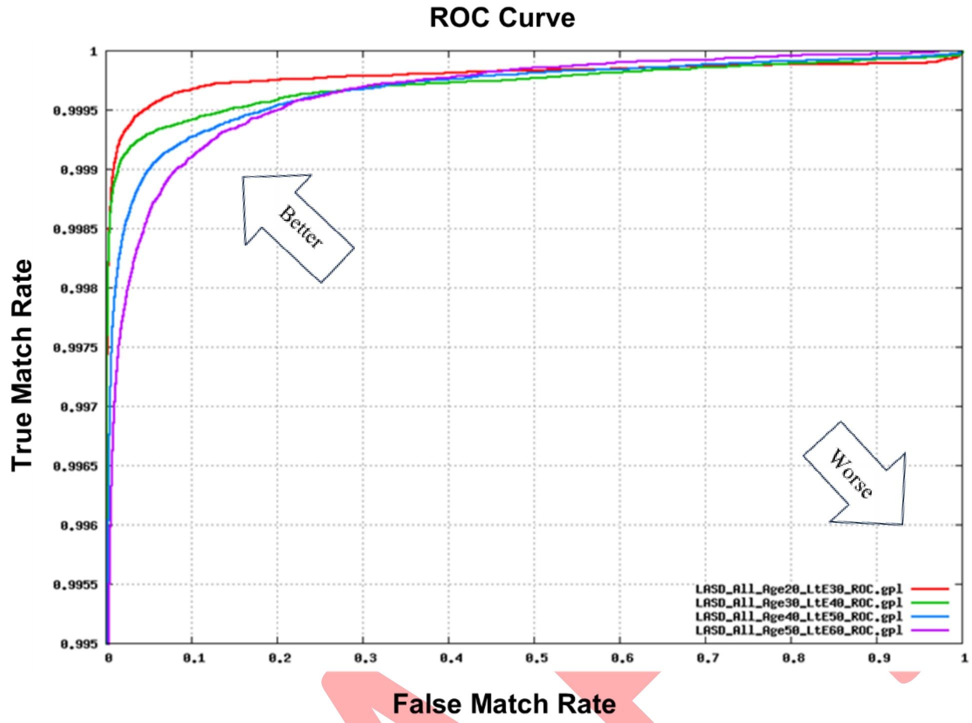


Figure 12: Various ROC Curves: Illustrates ROC curves for multiple algorithms, each showing the relationship between True Match Rate (TMR) and False Match Rate (FMR) across varying threshold settings. Curves that trace closer to the top-left border indicate higher accuracy and better discrimination between genuine and imposter pairs, while those nearer the diagonal represent weaker performance.

6.3.2 A ROC curve demonstrates several things:

6.3.2.1 It shows the tradeoff between True Match Rate and False Match Rate: any increase in True Match Rate will be accompanied by an increase in False Match Rate.

6.3.2.2 The closer the curve follows the left-hand border and then the top border of the ROC chart, the more accurate the algorithm.

6.3.2.3 The closer the curve comes to the 45-degree diagonal of the ROC chart, the less accurate the algorithm.

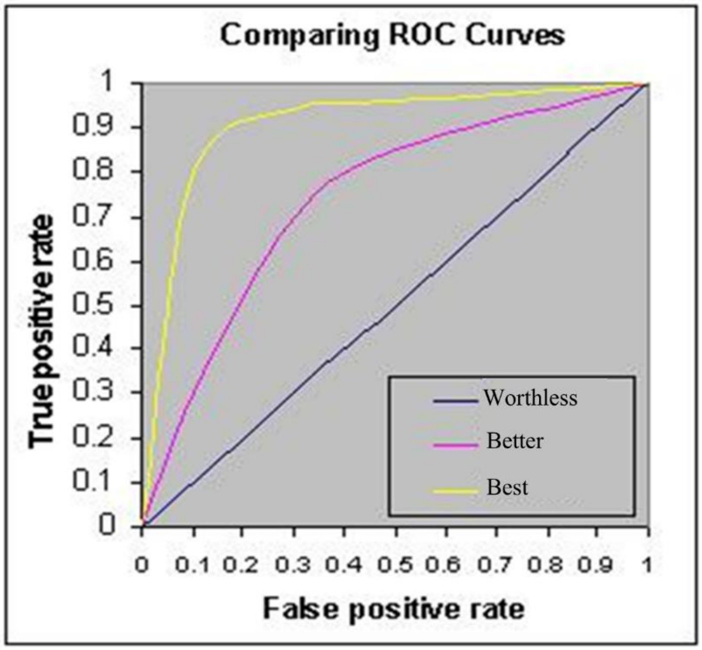


Figure 13: Comparative ROC Curves: Comparison of ROC curves for multiple algorithms. The curve following the upper-left boundary demonstrates superior performance; curves closer to the diagonal indicate reduced discrimination capability.

6.4 DET Curve: Detection Error Tradeoff curve.

6.4.1 This type of plot compares the False Non-Match Rate (FNMR) versus the False Match Rate (FMR). It is useful to note that the FNMR is equal to 1 minus the TMR (FNMR = 1 – TMR). These values are then plotted with an algorithm that varies the threshold score and creates the percentages at each score.

Figure 4

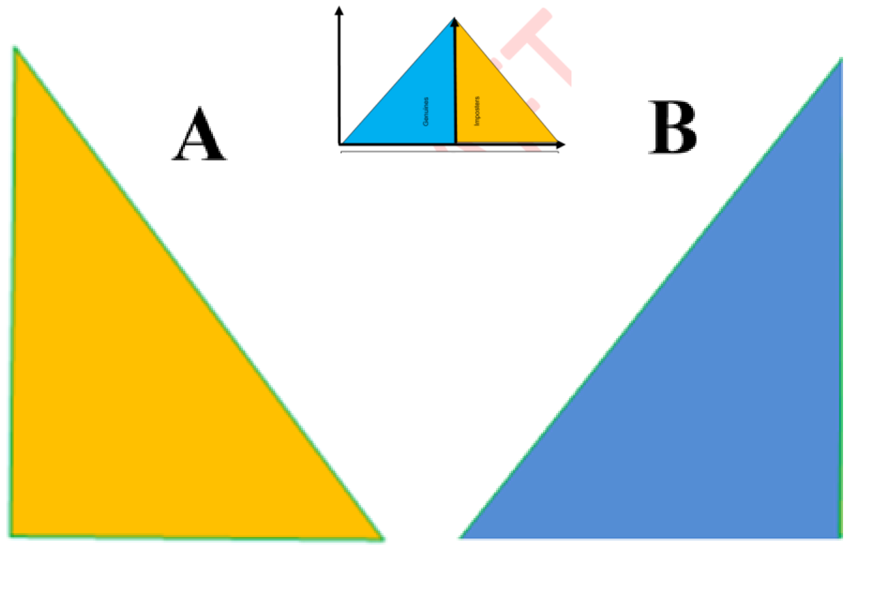


Figure 14: DET Curve Scoring Regions. A: Excerpts from Figure 4, this is the False Accept area. This area represents imposter pairs with a score above the threshold score which differentiates between a genuine and an imposter. B: Excerpt from Figure 4, this is the False Reject area. This area represents genuine pairs with a score below the threshold score which differentiates between a genuine and an imposter.

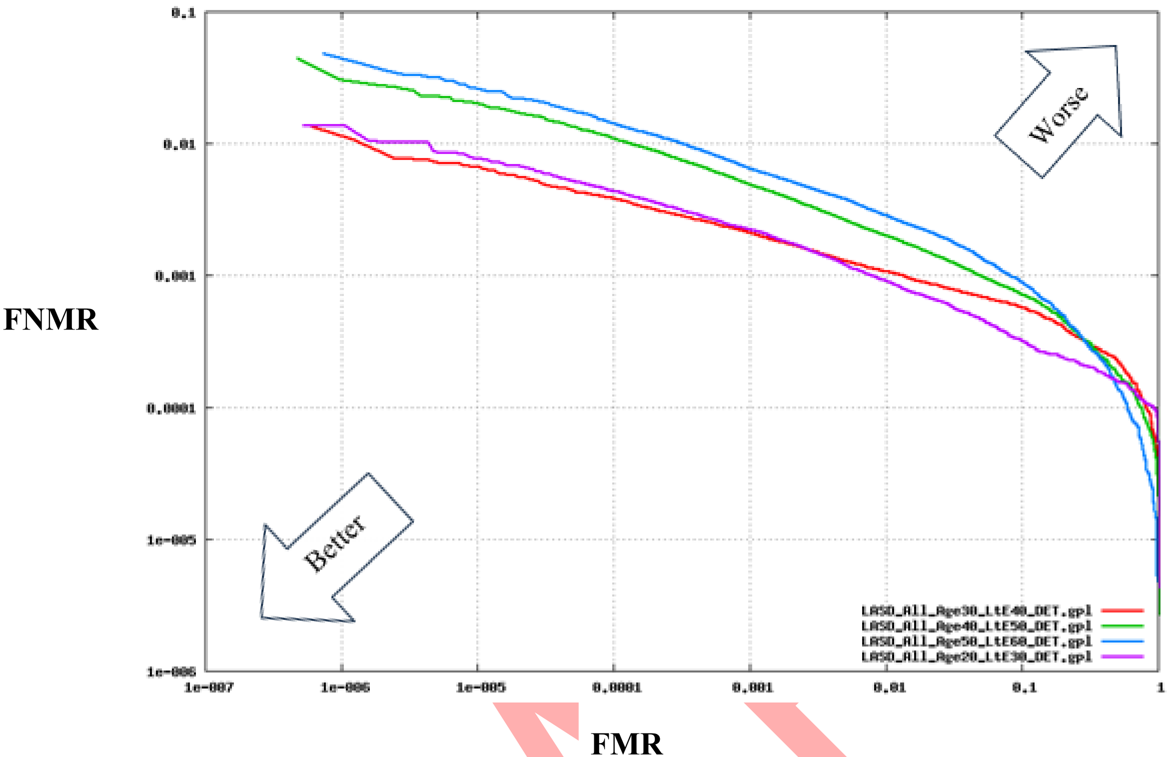


Figure 15: Various DET Curves: Detection Error Tradeoff (DET) curves for multiple algorithms. Curves that remain lower and flatter indicate better performance with fewer errors across threshold settings.

6.4.2 A DET curve demonstrates several things:

6.4.2.1 It shows the tradeoff between the False Match rate (FMR) and the False Non-Match Rate (FNMR) for all threshold values.

6.4.2.2 The closer the curve follows the bottom border and the flatter the curve is the more desirable result (see also Figure 18).

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

6.4.2.4 See section NIST References and Definitions for more examples and definitions on DET curves.

6.5 CMC Curve: Cumulative Match Characteristic curve.

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

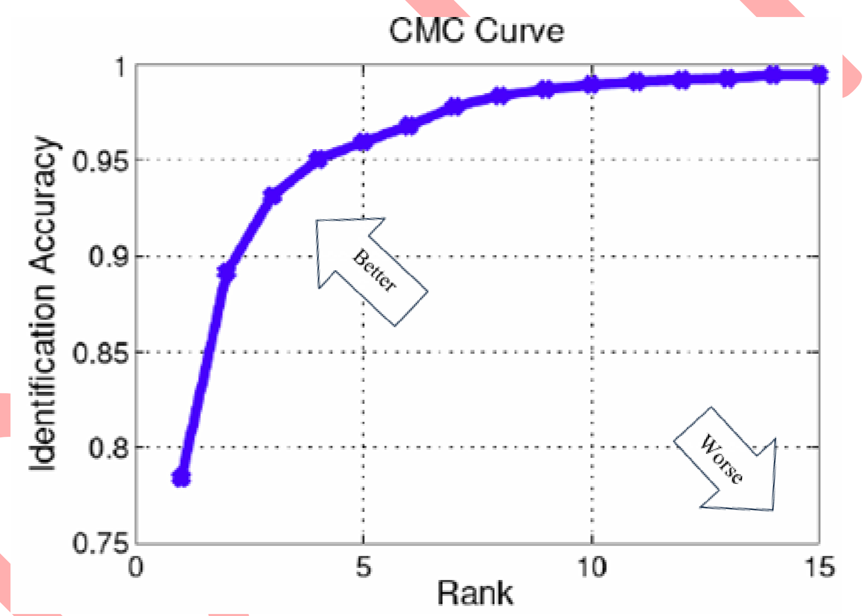


Figure 16: Simple CMC Curve: Cumulative Match Characteristic (CMC) curve illustrating how often the correct identity appears within the top candidate ranks. Steeper curves represent systems that achieve accurate identifications at lower ranks.

6.5.2 In Figure 16 above:

6.5.2.1 The chance of a true mate in the candidate results at rank 1 is ~77%.

6.5.2.2 The chance of a true mate in the candidate results at rank 5 is ~96%.

6.5.2.3 The chance of a true mate in the candidate results at rank 10 is ~99%.

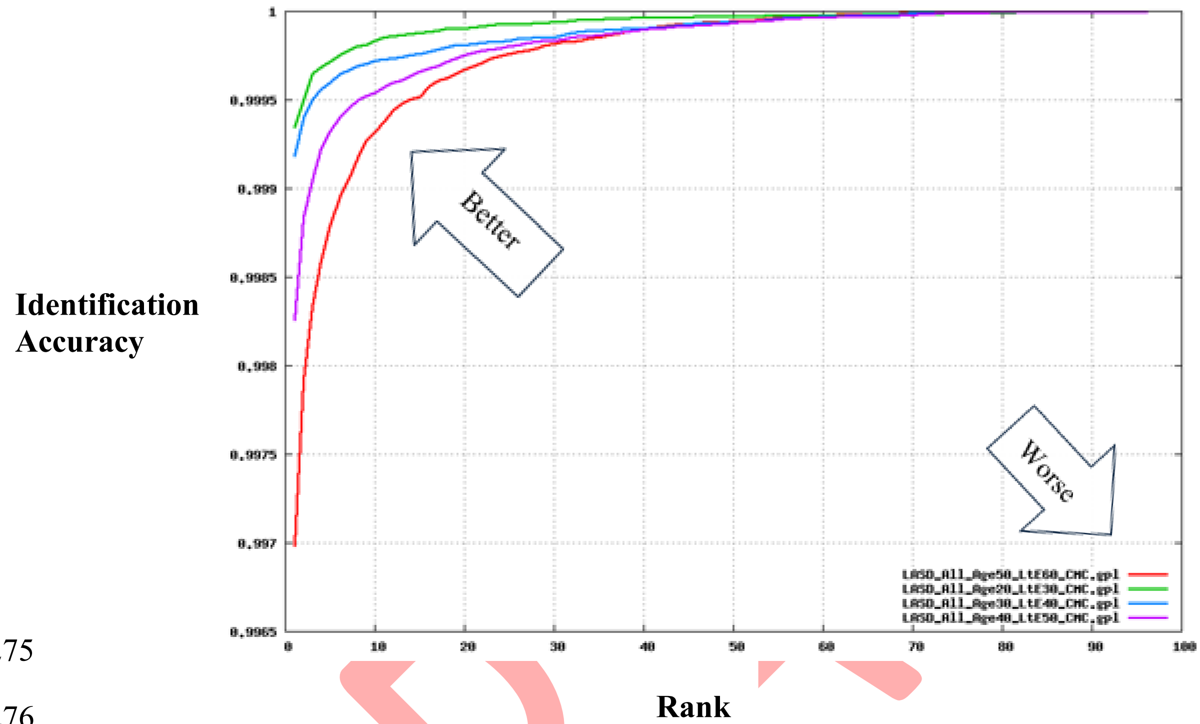


Figure 17: Various CMC Curves: Comparative CMC curves showing relative ranking performance for different algorithms or data sets. Curves approaching 100% at lower ranks indicate stronger 1:N identification performance.

6.6 To generate the DET, ROC, and CMC charts used by FRTE types of results:

6.6.1 Biometrics samples are compared against each other.

6.6.1.1 Each probe biometric sample is compared against all gallery samples.

6.6.1.2 Genuine and imposter scores are generated.

6.6.1.3 The resulting scores are sorted and ranked.

6.6.1.4 True Match Rate (TMR) and False Non-match Rate (FNMR) are computed at multiple thresholds.

6.6.2 Create and plot the ROC curve.

6.6.2.1 False Match Rate (FMR) and False Non-match Rate (FNMR) are computed at multiple thresholds.

6.6.3 Create and plot the DET curve.

6.6.4 Create and plot the CMC curve.

7. Boundaries

7.1 Agencies should read and interpret NIST FRTE tests with a focus and understanding on how these tests inform agencies deploying an FRS and the limitations of the tests.

7.2 Strengths of NIST tests:

7.2.1 Well defined and well-planned tests which are regularly updated.

7.2.2 Structured and non-biased test of the state of facial biometric technology.

7.2.3 Careful selection of data used for the test.

7.2.4 Testers and reviewers are widely recognized for their expertise.

7.2.5 Results documented extensively and placed for public review.

7.2.6 Historical test results show the progress for facial biometric technology over the last twenty years.

7.3 Limitations of NIST tests:

7.3.1 Cannot ensure that the algorithms tested are commercially available.

7.3.2 The vendors that participate in NIST tests are experts in their proprietary technology and can therefore customize and tune their algorithm to the test data sets.

7.3.3 Cannot predict whether a vendor tested will be bought and taken private and therefore never tested again.

7.3.4 NIST is somewhat limited in the gallery sized they can effectively test. If an agency has an operational gallery or (say) 50M, then the NIST tests done to date may not scale this high.

8. Example NIST Report

8.1 An example of a NIST report that presents DET and CMC curves is **NIST Ongoing Face Recognition Technology Evaluation (FRTE) Part 1: Verification:** **https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf**

8.2 This report used the following sets of data):

8.2.1 Child Exploitation images

8.2.2 Visa images

8.2.3 Mugshot images

8.2.4 Selfie images

8.2.5 Webcam images

8.2.6 Wild images

8.3 Each set of images has the following metric defined:

8.3.1 Number of images

8.3.2 Number of subjects

8.3.3 Number of subjects with known mates

8.3.4 How the images were captured

8.3.5 Basic image description

8.3.6 Age of subjects

8.3.7 Nationality of subjects

8.4 How the tests for each type of image were done:

8.4.1 How the images were compared.

8.4.2 How many templates were produced for each 1:1 verification.

8.4.3 Failure to enroll metrics.

8.4.4 How accuracy was determined and plotted.

8.5 DET Curves:

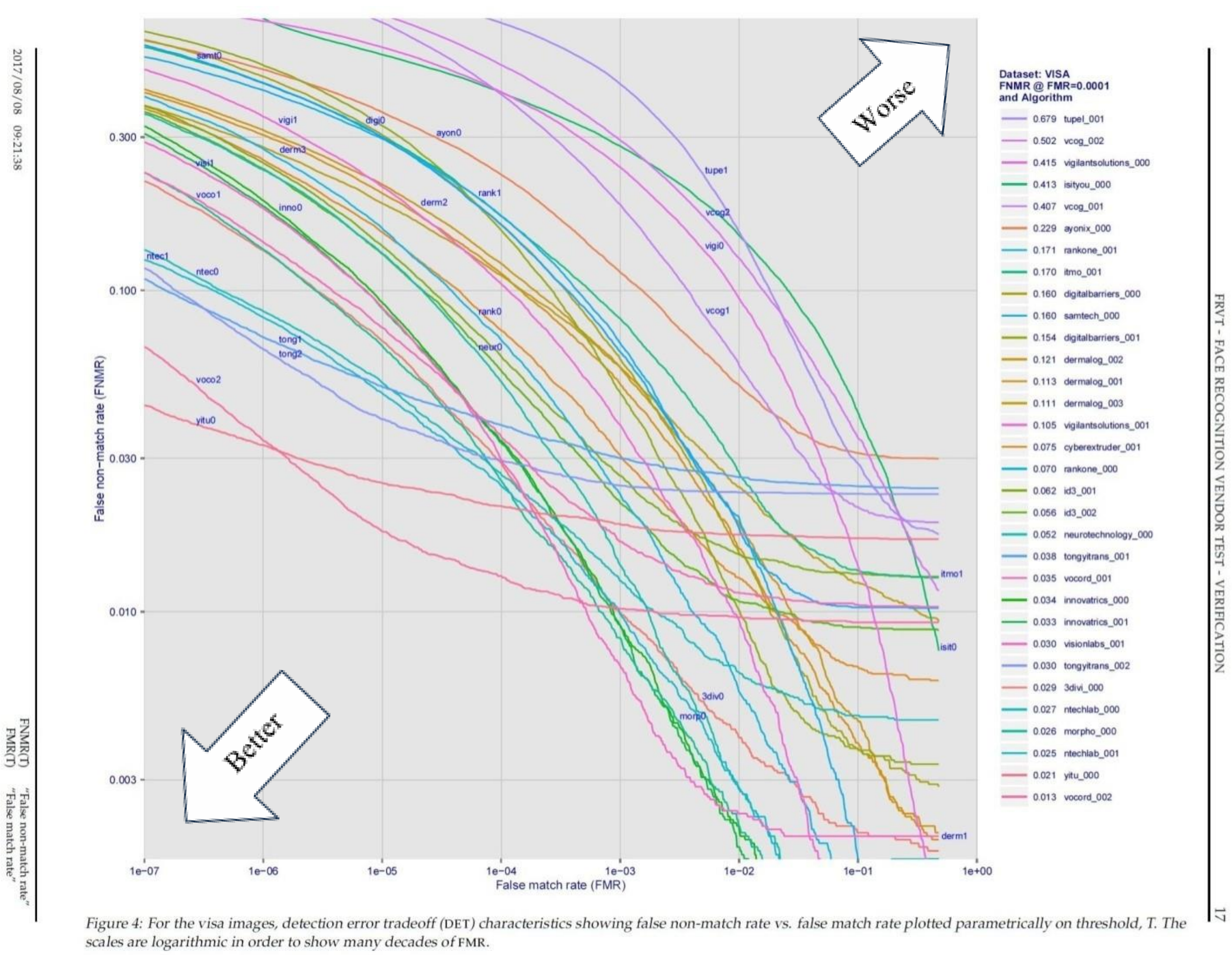


Figure 18: DET Curve Example: Shows DET curves for all algorithms evaluated using the Visa image dataset in the NIST FRTV. Each curve reflects tradeoffs between false match and false non-match rates specific to that dataset.

8.5.1 Figure 18: This chart shows the DET curve for all algorithms for the Visa data set. Note that this DET curve would be different for each type of facial data sets tested.

8.6 CMC Curves:

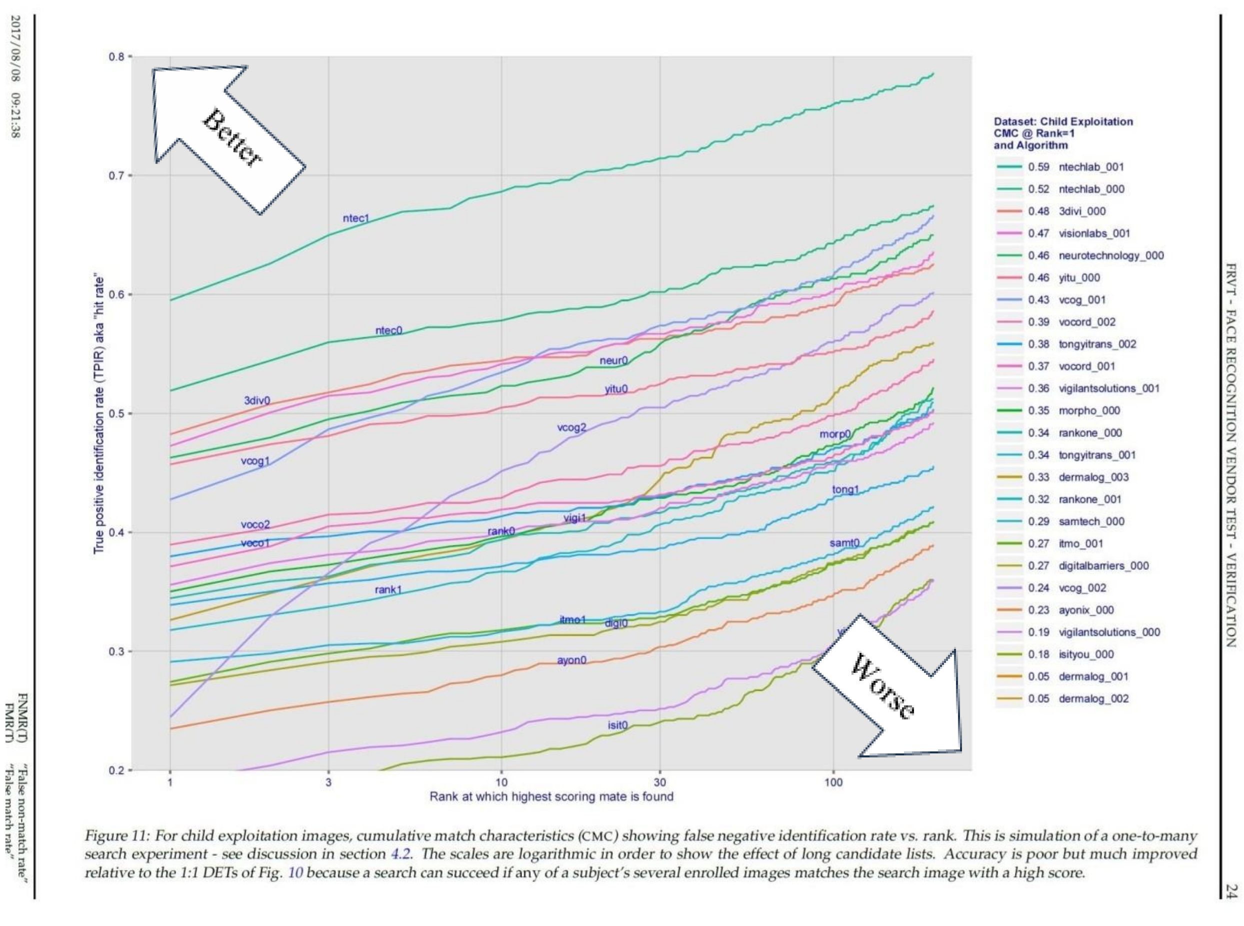


Figure 19: CMC Curve Example: Depicts CMC curves for the Child Exploitation dataset in the NIST FRTV. Variations among curves illustrate differences in identification accuracy across tested algorithms.

8.6.1 Figure 19: This chart shows the CMC curve for all algorithms for the Child Exploitation data set. Note that this CMC curve would be different for each type of facial data sets tested.

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