ROC

08.07.26

PROUDLY AMERICAN MADE

Introduction to Face Recognition Technology Evaluation (FRTE) Testing

Neal Gieselman: Senior Principal Engineer
Dr Richard VorderBruegge: ret. FBI

Speakers

- Neal Gieselman:
  - Senior Principal Engineer
  - Master of Science Management of Technology
  - FISWG Systems and Capture Group
  - ASTM E30.12 Member
  - IAI Facial Identification Committee

- Dr Richard VorderBruegge:
  - ret. FBI
  - Owner: ForensicVB, LLC
  - FISWG Chair 2009-2015

About FISWG

In 2009, the Federal Bureau of Investigations (FBI) began the Facial Identification Scientific Working Group (FISWG) and was overseen by the FBI for the first few years. The FISWG mission is to develop consensus standards, guidelines, and best practices for the discipline of image-based comparisons of human facial features, provides recommendations for research development activities, and advances the state of the science in an ethical manner.

In 2014, the National Institute of Standards and Technology (NIST) and the Department of Justice (DOJ) created the Organization of Scientific Area Committees (OSAC) for Forensic Science. FISWG is still an active SWG reporting to the OSAC Facial Identification group.

FISWG has been a leader in gathering and disseminating accurate information regarding the proper application of FI, methodologies, and Facial Recognition technologies. FISWG has over 50 different agencies and private industry from across the world in the FI-related discipline meeting twice annually.

Introduction

- 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.
- 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.

Introduction

- 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.
- This document references National Institute Standards and Technology (NIST) as an authoritative testing agency; other agencies may do similar testing.
- The full FISWG document can be found at www.fiswg.org

Introduction

- 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:

- Understand commonly used terms from these tests
- Gain practical knowledge of how these tests are reported
- Understand how these tests can add value for current or future FRS procurements
- Enhance understanding of an operational FRS strengths and limitations
- Generate realistic expectations about what can be derived from these tests

FRTE Document Sections

- Definitions
- Simplistic Test Scenario
- Ideal, Excellent, and Operational Scoring
- Manual Resolve: Human Examiner
- Basic Accuracy Charts: ROC, DET, CMC
- Overall Testing Summary

Definitions

- Imposter pair: a pair of images which are not of the same identity (non-mate)
- Genuine pair: a pair of images which are of the same identity (mate)
- Ground truth: where all imposter pairs and genuine pairs have been verified to be correct

Simplistic Test Scenario

- 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
- A facial collection of 1000 image pairs of genuines and a collection of 1000 image pairs of imposters. The images are passport quality frontal poses.
- 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.

Ideal Scoring

- 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 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

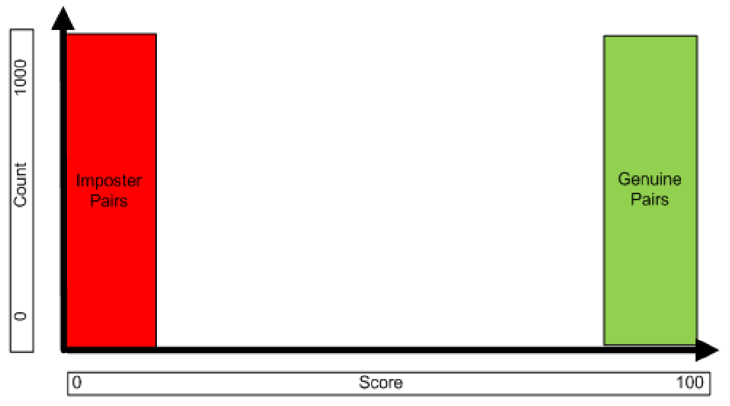


Figure 1: Ideal Scoring

Excellent Scoring

- 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 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.

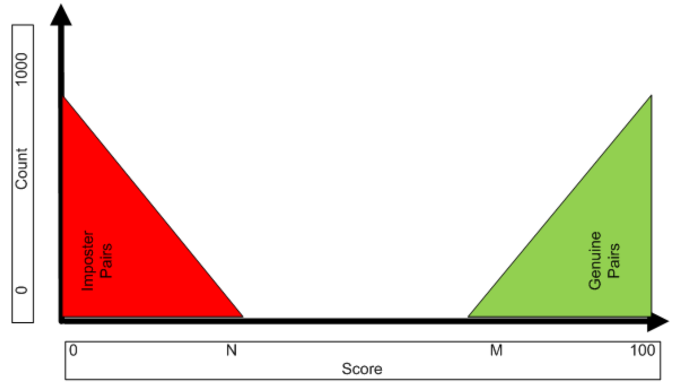


Figure 2: Excellent Scoring

Operational Scoring

- 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 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 valid genuine or imposter.

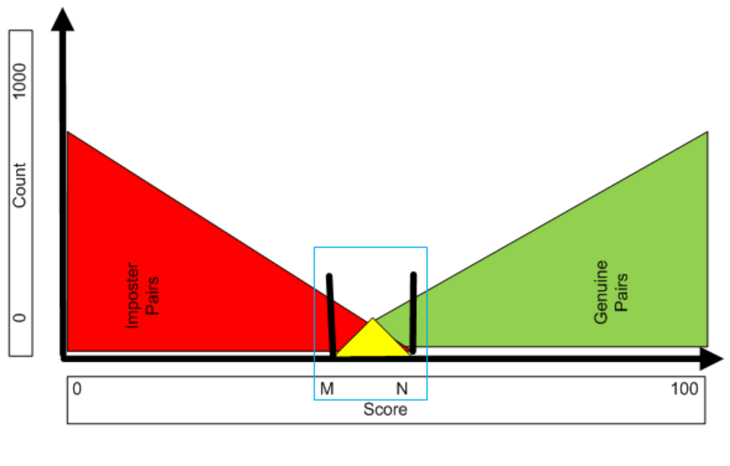


Figure 3: Operational Scoring

Manual Resolve: Human Examiner

- 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.

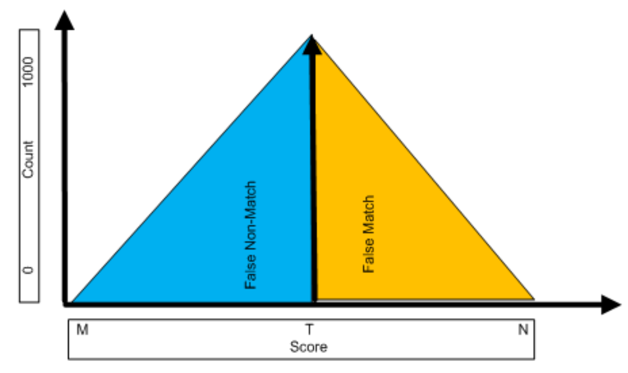


Figure 4: Manual Resolve Area

True Match and True Non-Match

- 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.

- The examples shown 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)

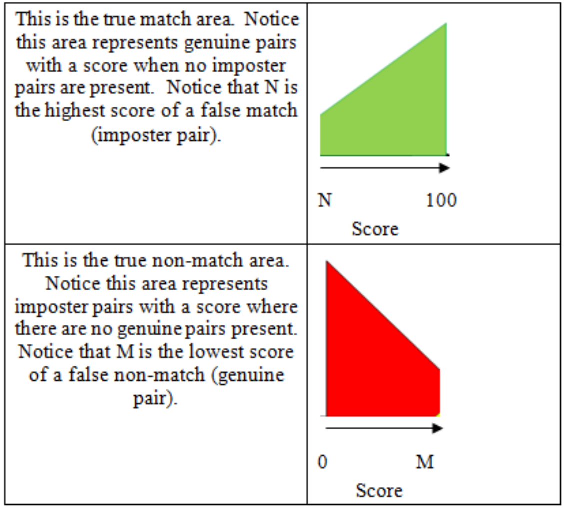


Figure 5: True Match and Non-match Areas

Visualization of Five Scores

- Assume that Figure 17 is the result of testing showing imposter and genuine scores

- The genuine scores are represented in the green area
- The imposter scores are represented in the red area
- The manual resolve “False positives” are represented in the orange area (between T and N)
- The manual resolve “False negatives” are represented in the blue area (between M and T)

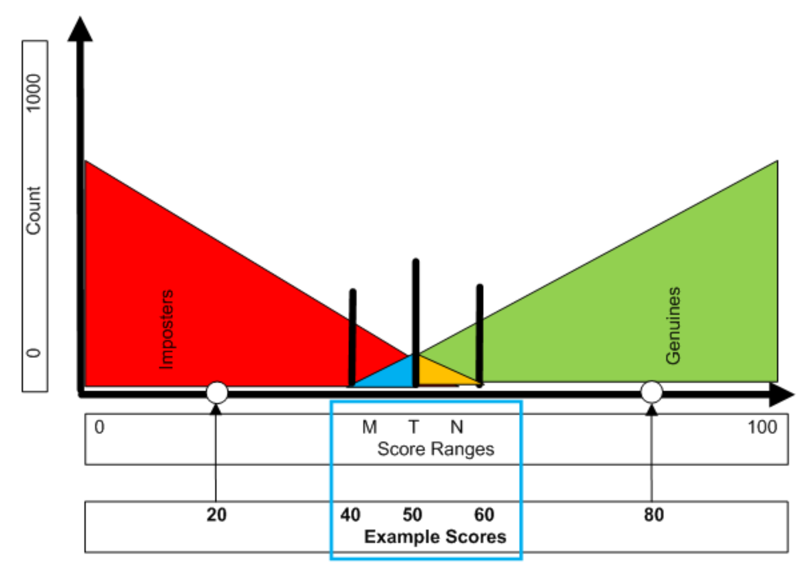


Figure 17: Operational Scoring

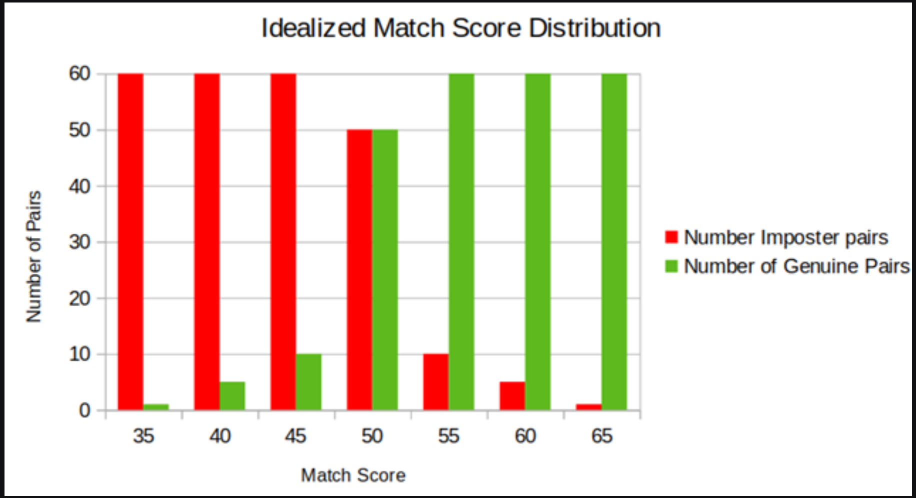
Visualization of Five Scores

- A description of the five example scores:
  from left to right: 20, 40, 50, 60, 80

20: A very low score suggesting an imposter. No genuine pair scored this low but a genuine pair "could" produce this score given exceptional or untested image quality issues.
40: This test score has a very low chance of being a genuine pair because this score is at the low end of the manual resolve score range.
50: This test score had a 50/50 chance of being a genuine or an imposter pair because it is in the middle of the manual resolve score range.
60: This test score has a very low chance of being an imposter pair because this score is at the high end of the manual resolve score range.
80: A very high score suggesting a genuine. No imposter pair scored this high but an imposter pair "could" produce this score given exceptional or untested image quality issues, a sibling, a look a-like, or a confirmed twin.

“Interpretation” of *Similarity* scores

- NIST tests can help organizations define their preferred threshold scores, but agencies should also understand how specific *similarity* scores correspond to imposter and genuine pairs



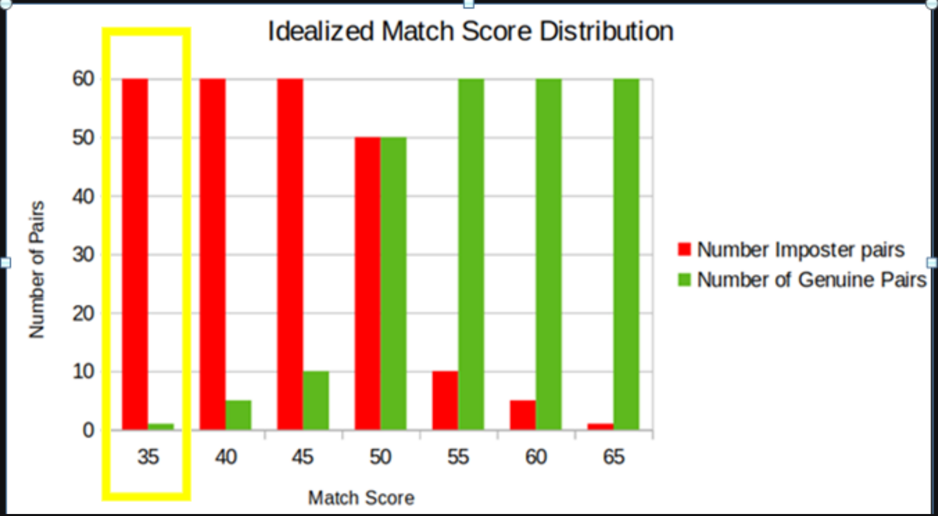
Similarity Score	Number of Imposter pairs	Number of Genuine Pairs
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000

®2026 ROC | All Rights Reserved

ROC

"Interpretation" of *Similarity* scores

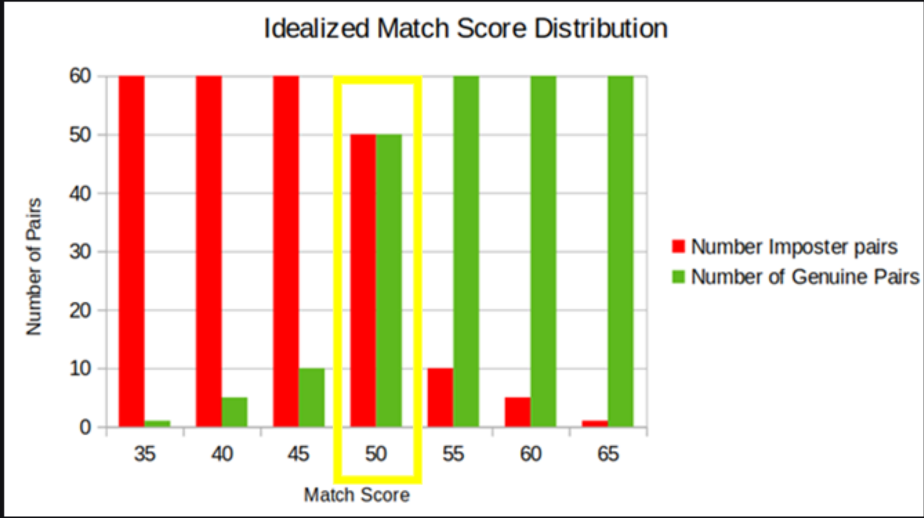
- A *similarity* score of 35 corresponds to 1000 imposter pairs and 1 genuine pair:
  - 99.9% chance this represents an imposter (1000/1001)
  - 0.1% chance this is a genuine pair (1/1001)



Similarity Score	Number of Imposter pairs	Number of Genuine Pairs
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000

“Interpretation” of *Similarity* scores

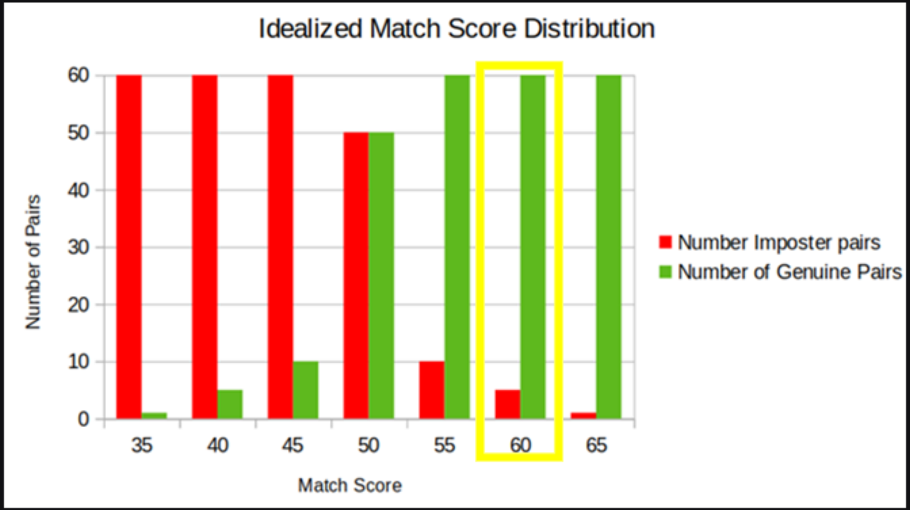
- A *similarity* score of 50 corresponds to 50 imposter pairs and 50 genuine pairs:
  - 50% chance this represents an imposter (50/100)
  - 50% chance this is a genuine pair (50/100)



Similarity Score	Number of Imposter pairs	Number of Genuine Pairs
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000

“Interpretation” of *Similarity* scores

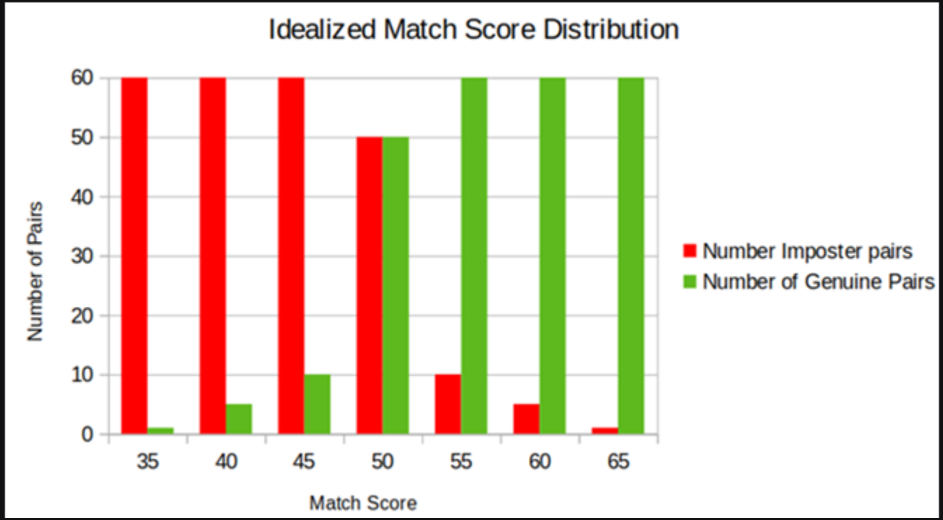
- A *similarity* score of 60 corresponds to 5 imposter pairs and 500 genuine pairs:
  - 1% chance this represents an imposter (5/505)
  - 99% chance this is a genuine pair (500/505)



Similarity Score	Number of Imposter pairs	Number of Genuine Pairs
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000

“Interpretation” of *Similarity* scores

- **CAVEAT!** *Similarity* score distributions are derived from test data. If the operational data is of different quality from the test data, the values discussed above are likely to be different.



Similarity Score	Number of Imposter pairs	Number of Genuine Pairs
35	1000	1
40	500	5
45	100	10
50	50	50
55	10	100
60	5	500
65	1	1000

Summary

- 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

- The examples shown 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*.**

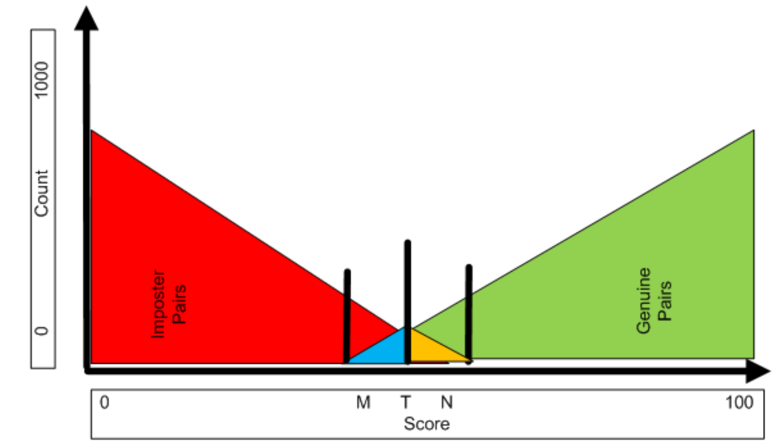


Figure 6: Operational Scoring

Basic Accuracy Charts

- 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.

- 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

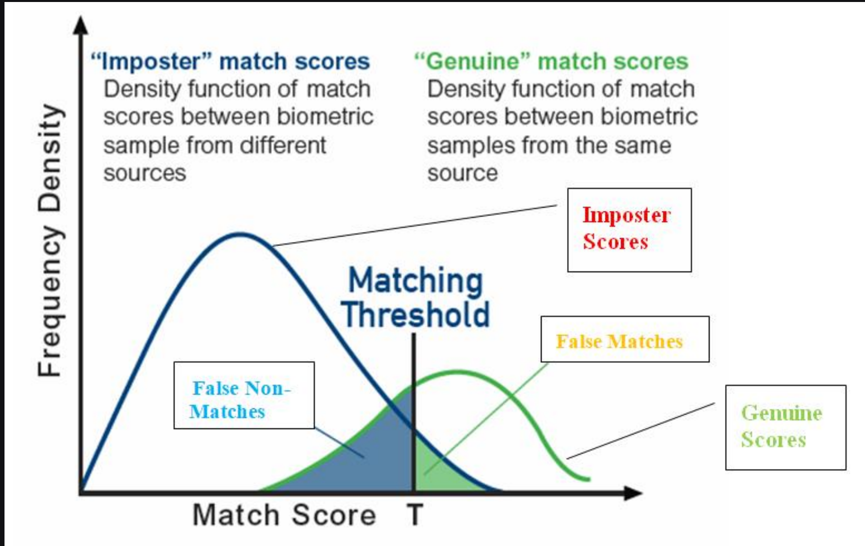


Figure 7: Similarity Score Distributions

ROC Curve

- 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.
- These areas are then plotted with an algorithm that varies the threshold score and creates the percentages at each score
- An algorithm that is wrong as often as it is right will be a diagonal from 0,0 to 1,1 (blue line)
- A better algorithm is shaded pink
- The best algorithm shown is shaded yellow

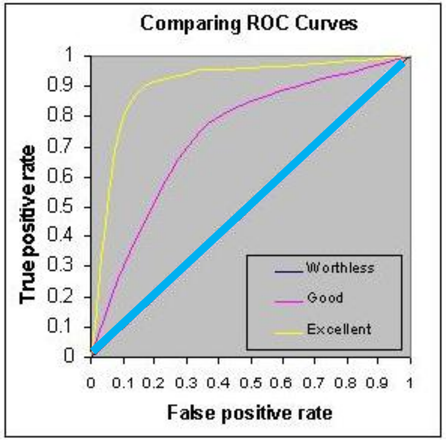


Figure 10: ROC Curves

ROC Curve

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

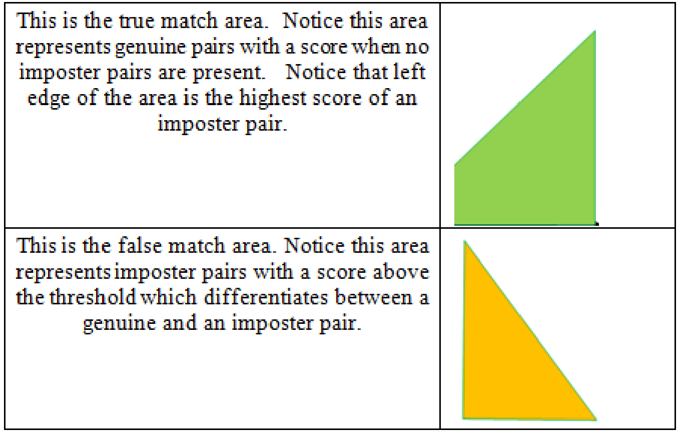


Figure 8: ROC Curve Score Components

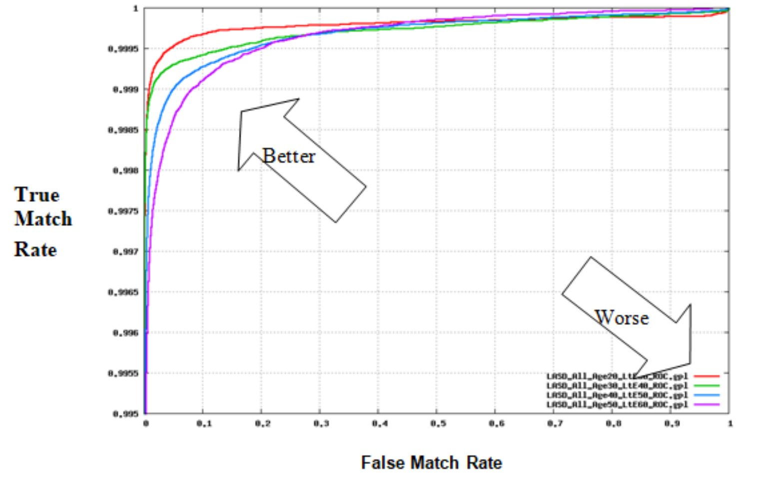


Figure 9: Multiple ROC Curves

ROC Curve

- Figure 10 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.

- 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.
  - ***The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.***
  - ***The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.***

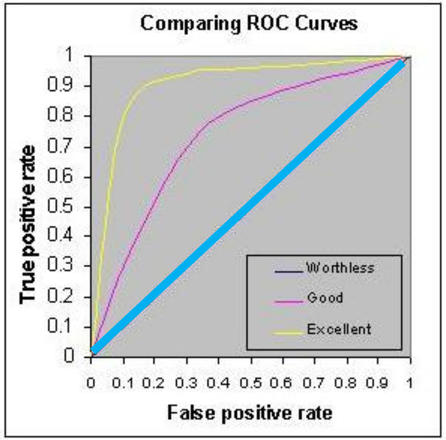


Figure 10: ROC Curves

Detection Error Tradeoff (DET) Curve

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

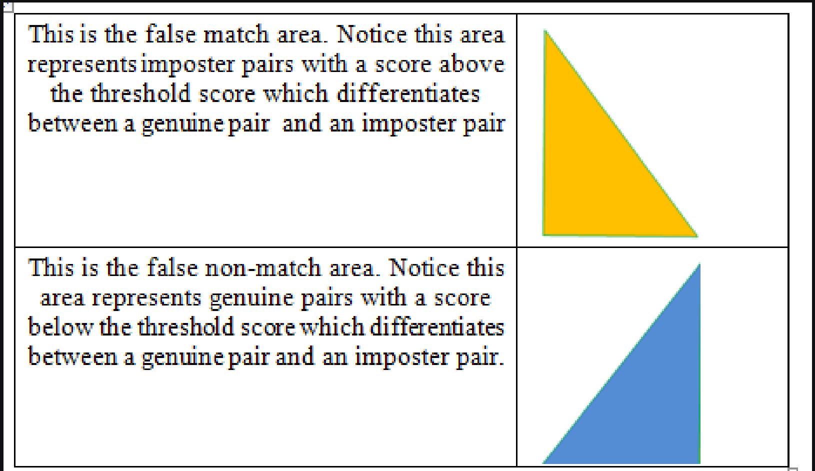
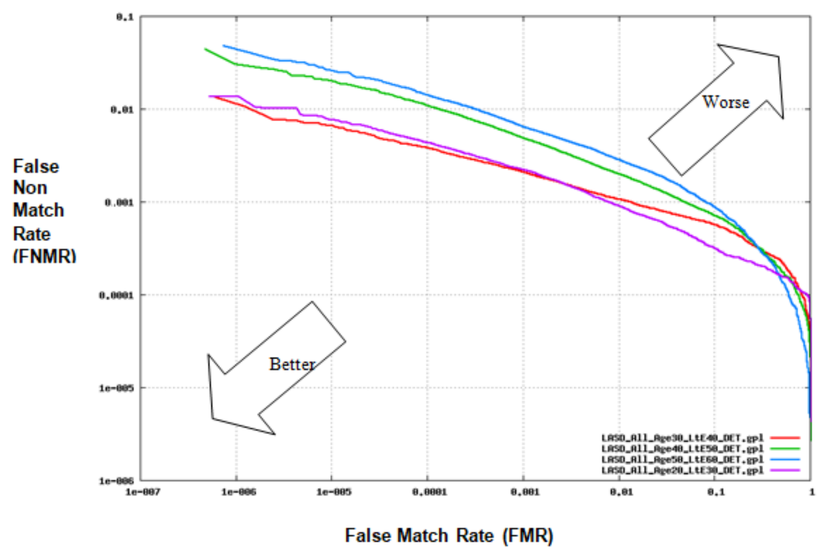
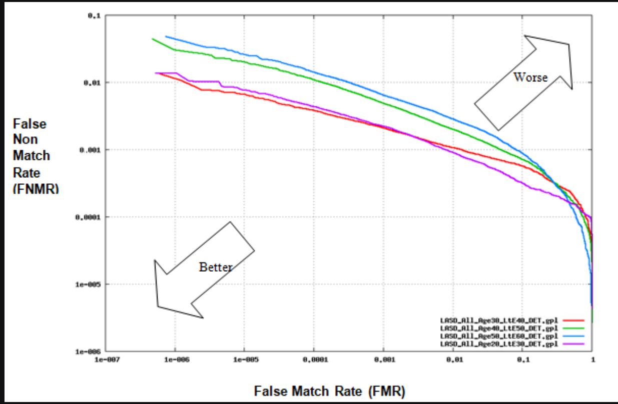


Figure 11: DET Curve Scoring Regions



Detection Error Tradeoff (DET) Curve

- It shows the tradeoff between the FMR and FNMR for all threshold values
- The closer the curve follows the bottom border and the flatter the curve is the more desirable result
- ***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***



Detection Error Tradeoff (DET) Curve

- Reference: Face Recognition Technology Evaluation (FRTE) Part 2 Identification

Same person under two identities

Twins

Siblings

Look-alike

Identification: seldom needs human review

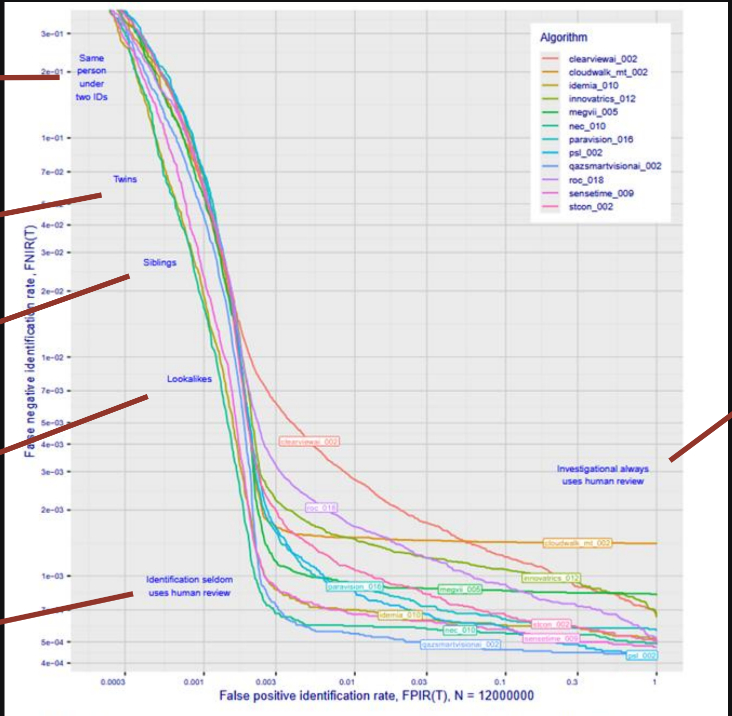


Figure 1: Identification miss rates across the false positive range. N = 12 million individuals are enrolled with one recent image.

Investigational: always needs human review

Cumulative Match Curve (CMC)

- 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.

  - 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%.**

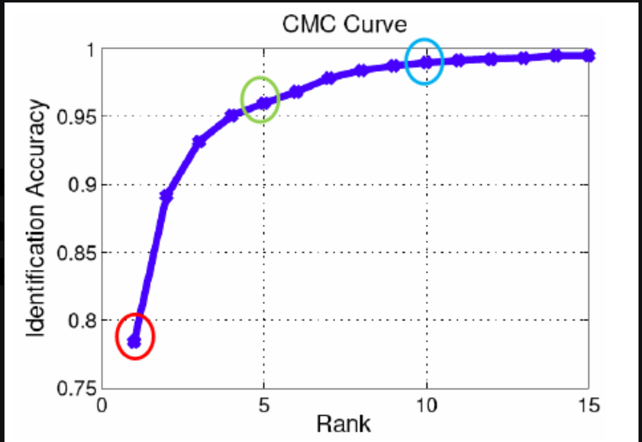


Figure 13: CMC Curve

Cumulative Match Curve (CMC)

- The CMC is used as a measure of 1: N identification system performance. It judges the ranking capabilities of an identification system.
- ***The closer the curve follows the top left corner and the top border of the CMC curve, the more accurate the test.***

The chance of a genuine m
The chance of a genuine n
The chance of a genuine m

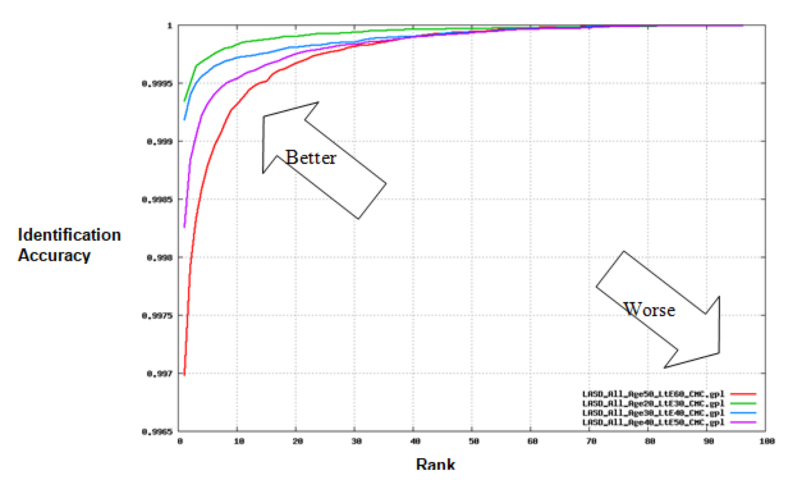


Figure 14: Various CMC Curve

Overall Testing Process

- To generate the DET, ROC, and CMC charts used by FRTE testing the following process steps can be summarized:
  - 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
  - TMR and FNMR are computed at multiple score thresholds
  - Create and plot the ROC Curve
  - FMR and FNMR are computed at multiple score thresholds
  - Create and plot the DET Curve
  - Determine the candidate ranks at which a genuine match occurs
  - Create and plot the CMC Curve

ROC vs. DET vs. CMC

- For an FRS deployed for identification card issuance (i.e., DMV, passport, badging) the ROC curve has more value:
  - Issuing fraudulent id cards (false accepts) is to be minimized
  - False rejects are tolerated and can be manually addressed to correct them
- For an FRS deployed for investigative searching (i.e., law enforcement, national security) the DET curve has more value:
  - Missed identifications (false rejects) are to be minimized
  - Improper identifications (false accepts) are addressed by human examiners
- CMC performance is used to determine how many search candidates are returned to achieve a desired identification rate

Personal Experience

- NIST has been stating this for 15 years:

  “*Operational implementations usually employ a single face recognition algorithm. Given algorithm specific variation, it is incumbent upon the system owner to know their algorithm. While publicly available test data from NIST and elsewhere can inform owners, it will usually be informative to specifically measure accuracy of the operational algorithm on the operational image data, perhaps employing a biometrics testing laboratory to assist.*”
- I would suggest one addition to this:

  “Incorporating Mission specific operational processes and risk assessment“

Personal Experience

- In my last operational deployment I tested an algorithm across three versions
- Same data set and vendor, each release had different behavior and scoring ranges
- Accuracy improved with each, but the behavior of the algorithm changed
- The final deployment did improve operational accuracy, but false rejects were located by the human examiners
- The root causes of these were image quality related, but the expertise and tenacity of the human examiners excelled
- The scoring thresholds used were off by a small amount, but because we had good test results to reference, we knew the impacts of adjusting the search parameters: score threshold, manual review time impacts, etc
- Without the test results, any adjustments would have been guesses

Personal Experience

- FRTE type testing is easy once properly defined and executed, and can be easily repeated if needed adapting to operational or Mission related updates:
  - New algorithms
  - Changes to image quality
  - Legal or policy updates
  - Mission threats
- Assuming total reliance on FRTE test results without proper testing and incorporating Mission related aspects will result in operational failures
- I prefer not to guess when Mission failure is a possible outcome
- If someone says “Don’t worry”, then one should worry
- My response: “Show me the data and we can make a fact based decision”

Summary

- If deployed and utilized properly FRT is a critical asset in supporting public safety and national security
- FSIWG and NIST OSAC have a large collection of documents that can assist in FRT terminology, deployments, examiner training, etc
- FISWG has a multi document series titled “Operational Assurance” which details these steps in great detail in an incremental manner

- **Improper arrests need to be reduced**

Introduction to Face Recognition Technology Evaluations (FRTE) Testing

Questions, comments, concerns?

www.fiswg.org

Neal Gieselman: Senior Principal Engineer
Neal.gieselman@roc.ai
Dr Richard VorderBruegge: ret. FBI
rwvb@ForensicVB.com