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 1.0 2020.12.11

Understanding and Testing for Face Recognition Systems Operation Assurance

This document provides guidelines and techniques to help administrators of automated face recognition systems (FRS) test them in an operational setting to provide assurance of face recognition accuracy, system integrity, configuration, and data storage.

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 types of testing including, but not necessarily limited to, system setup, system tuning, workflow management and improvement, and proof of concept pilots.

Related Work

The overarching goal of *ISO/IEC 19795-6:2012 — Information technology — Biometric performance testing and reporting — Part 6: Testing methodologies for operational evaluation* [1] "to measure or monitor operational biometric system performance" parses into sub-goals of:

- Determining if performance meets expectations or may be improved through system tuning or reconfiguring;
- Predicting expected performance for increases in number of enrollments and/or systems;
- Obtaining information that affects system performance (e.g., changes in target population and environmental parameters); and/or
- Obtaining performance data from a pilot implementation or to benchmark future systems.

This document serves as a qualitative introduction to a subset of the topics and concepts that are covered in greater technical detail in ISO/IEC 19795-6:2012.

Glossary

FR = face recognition

FRS = face recognition systems

CMC = cumulative match characteristic

ROC = receiver operating characteristics

Exact binary copy = two images that are byte-identical, i.e., they yield the same MD5 checksum. Note: An image encoded as a PNG is **not** an exact binary copy of the same image encoded as a JPG.

Mate = a separately captured image of the same subject

Retrieval rank = the order with which a particular image in the gallery was retrieved respective to a probe and all other images in the gallery

Operational Testing Techniques

The operational testing techniques provided in this document can be performed manually, automatically, or semi-automatically, depending on operational factors and the administrator's technical knowledge.

Note: System administrators familiar with performance tests run by the National Institute of Standards and Technology (NIST) Information Technology Laboratory (ITL) like the Face Recognition Vendor Tests (FRVT) and the Multiple Biometrics Grand Challenge (MBGC) should be aware that biometric test results may not be applicable to other datasets and operational systems with different processing constraints. These test results should be reviewed within the context in which they are obtained and might not generalize to operational data.

Tier-based System Evaluation Guide

A tier-based approach to operational system evaluation is proposed which provides a tradeoff between the difficulty of implementing the testing strategy and the level of system assurance that can be achieved through such testing.

Fig. 1 illustrates the tier-based testing paradigm.

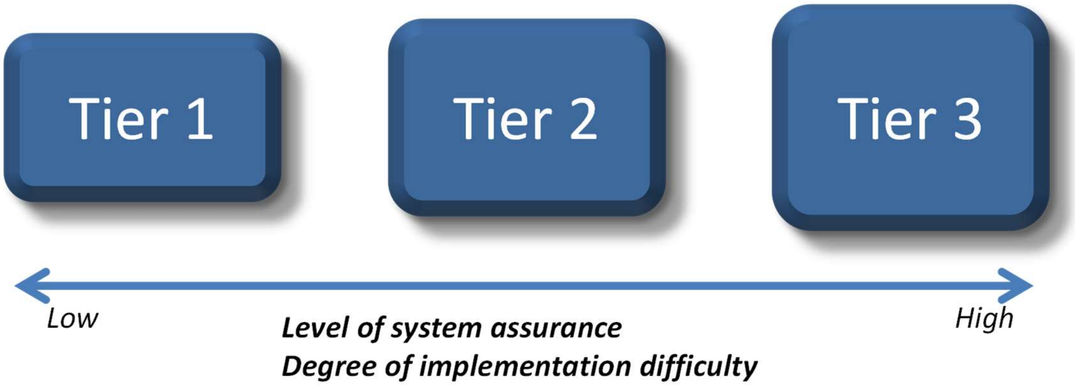
Tier 1
Tier 2
Tier 3
Low
High
Level of system assurance
Degree of implementation difficulty

Fig. 1 Recommendations for performing an operational evaluation of a face recognition system are tiered, which allows administrators a tradeoff between the degree of testing difficulty and the level of system assurance.

Tier 1 – Basic:

The simplest level of testing is Tier 1 testing, or self-identification, which is a basic check to ensure there are not egregious errors in the system. Tier 1 testing can be performed by the system administrator but may also be performed by system users in certain circumstances.

Tier 1 testing is performed as follows:

- Query the face recognition system with images whose exact binary copies have already been enrolled in the database.
- Ensure the Rank-1 match candidate is the same image as the probe.
- Continue to perform the first two steps with additional images as often as availability of computing resources and human effort allow.

Tier 1 testing is performed to ensure:

- Gallery images are properly enrolled, and their corresponding templates are both valid and accessible.
- The Face Recognition (FR) system's network access is not interrupted for any distributed resources.
- Software running on local and network resources is not exhibiting any failures.

Tier 1 testing provides minimal assurance regarding the recognition accuracy of the face recognition system. It provides a basic level of assurance of the system integrity, configuration, and data storage.

Tier 2 – Intermediate:

Tier 2 testing provides intermediate retrieval accuracy statistics on the face recognition system using test subjects whose mates are known to be in the gallery. Depending on operational factors and the administrator's technical knowledge, Tier 2 testing may be performed by the system administrator and/or the FR algorithm vendor/integrator.

Tier 2 testing must be undertaken using the operating or vendor recommended default threshold in order that the results are truly representative of expected operational performance.

Testing should be undertaken with operationally representative images. If the query images are from a number of different sources (e.g., passport, CCTV, social media) it may be necessary to categorize the images into different query sets based on these sources. In the FRVT 1:N tests, CMC curves are computed with threshold set to zero as per FRVT 1:N report (NIST IR 8271). However, it may be desired to adjust the threshold based on operational scenarios in order to determine the retrieval rate for lower quality images which replicate operational constraints.

Tier 2 testing is performed as follows:

- Query system with images whose mates are known to be in the system. The query image should be from a separate encounter than at least one gallery mate. Query images should be representative of operational distributions. For example, a database of 10,000 images comprised of 80% white males and 20% black females, the system administrator might query images of 80 white males and images of 20 black females.

- Record the top rank in which the probe's mate was retrieved. If an image from the same encounter as the probe image is contained in the gallery (e.g., the exact binary copy), then do not consider the same encounter candidate list in the retrieval results.

- Continue to perform the first two steps with additional images as often as availability of computing resources and human effort allow.

- Use recorded rank retrievals from all retrieval tests to generate Cumulative Match Characteristic (CMC) accuracies.

When testing on galleries that continually increase in size, the recorded accuracies for this test generally will decrease over time. However, for a properly functioning recognition system no major negative changes in retrieval accuracy should occur for major system updates or high frequency fixed-interval tests.

How to generate CMC scores:

The CMC scores list what percentage of image queries had their mate returned at a particular retrieval rank or better. The CMC scores typically would store accuracies up to the first *N* ranks. The value of *N* would be determined based on how many retrievals are typically examined in the system's operational use. Thus, for an application where analysts examine the top 20 matches, *N*=20.

The CMC scores contain *N* values, which correspond to the Rank-1 identification rate, the Rank-2 identification rate, and all the way up to the Rank-*N* identification rate. These identification rates are interpreted as follows. The Rank-1 identification rate is what percentage of queries had their mate at the first retrieval rank (i.e., the closest match). The Rank-2 identification rate is what percentage of queries had the mate returned at the second retrieval rank, or better. Similarly, the Rank-3 identification rate is the percentage of queries that had their mate retrieved at the third rank, or better.

It is important to note the "or better" portion of the above description. If an image is matched at Rank-1, then it is also included in the Rank-2 (and beyond) CMC scores. Thus, when plotting the CMC scores respective to the retrieval ranks, the graph will not decrease.

When computing CMC scores in operational testing, the user would submit an image query and record the highest rank pertaining to a match. This process would be continued K times, where (as discussed above) K is determined based on available resources.

As an example, suppose K = 50 (i.e., 50 different images submitted to the system). For *N* = 7 ranks, the test may yield the results in Table 1:

Table 1 Example of Candidates for Rank Order 1 - 7

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number at Rank	32	7	1	3	1	0	1

The remaining 5 of the 50 submissions do not return within the designated limit of 7 candidates, thus are not reported in the table. Given these results, the CMC scores are as shown in Table 2:

Table 2 Example of CMC Scores for Candidates for Rank Order 1 - 7

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**
CMC score	32/50 =0.64	39/50 =0.78	40/50 =0.8	43/50 =0.86	44/50 =0.88	44/50 =0.88	45/50 =0.9

Thus, when performing Tier 2 testing, these CMC scores would be logged each time the test was performed. With a fixed set of query images, one should not expect major deviations from these CMC scores.

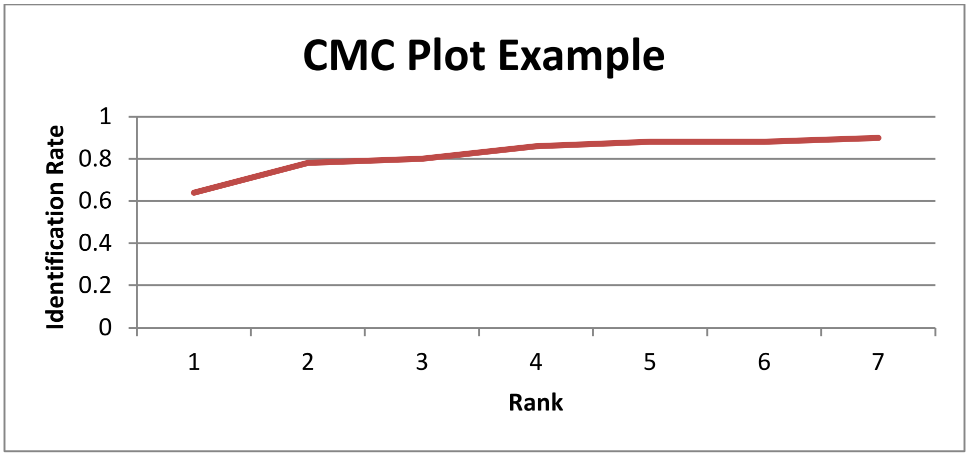

Fig. 2 Example CMC Plot

Fig. 2 is a plot of the CMC curve of the same example previously discussed. One can visualize certain aspects of CMC curves, and how they can be interpreted. One key point is that the Identification Rate is never decreasing with increasing Rank. When reading this plot, it can be interpreted as follows: at Rank=3, the rate is 0.8. Thus, in the test example, 80% of the time the subjects are matched within the top three ranks. Similarly, at Rank=7, the identification rate is 0.9. Thus, in the test example, 90% of the subjects are matched within the top seven ranks. Finally, because only results of the top seven ranks are recorded, the CMC curve never reaches 100% identification rate. Measuring results up to a higher rank, such as 50, may or may not allow one to record an identification rate of 1.0.
For more detail about creating a CMC curve, readers should refer to the "Relating ROC and CMC Curves" [3] and ""Face Recognition Vendor Test (FRVT) Part 2: Identification" [5].

<u>Tier 3 – Advanced:</u>

Tier 3 testing provides advanced recognition accuracy statistics on the face recognition systems. This section introduces and briefly discusses the Receiver Operating Characteristics (ROC) measurement and how to interpret this result. For more in depth discussions on how to generate and interpret advanced results such as ROC, readers are referred to other documents that provide these details [3][4].

Generally, these accuracy measurements cannot be computed by the FR system administrator. Instead, these measurements will often be made available through software provided by the FR algorithm vendor/integrator. Additionally, these results can be computed in non-operational scenarios and offline environments using sufficient and representative ground truth operational data. For example, the National Institute of

Standards and Technology routinely measures the accuracy of FR vendor algorithms on various image matching scenarios [4]. By understanding how to read advanced accuracy measurements, FR system administrators will have a better idea of which algorithms will best suit their respective organization’s needs.

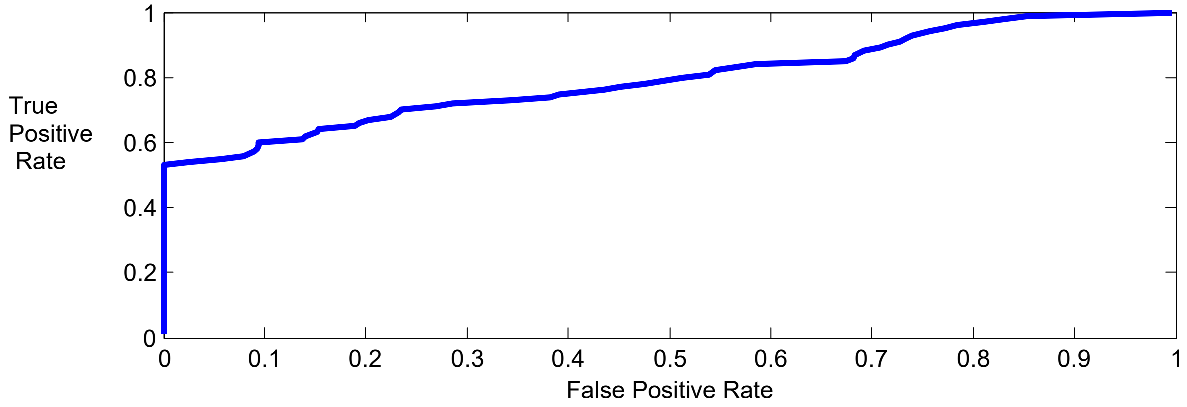

Fig. 3 Example ROC curve

Fig. 3 is an example of an ROC curve. A ROC curve plots on the x-axis the false positive rate (i.e., the percentage of impostor (or different subject) comparisons that exceed the match threshold) versus the true positive rate (i.e., the percentage of genuine (or same subject) comparisons that exceed the match threshold) on the y-axis. Any particular point on the ROC plot corresponds to the measured true positive and false positive rate at a particular match score threshold. The value of visualizing FR algorithm accuracy in the form of an ROC plot is that an organization (with the help of its integrator) can better determine which match threshold should be used as a function of how many false positive identifications will require human resource and effort to process, or the minimum true positive rate acceptable by its’ mission.

Note that while these performance metrics can be computed for most FR systems, they may not be appropriate for certain FRS applications. For example, rank based systems, used primarily in human adjudicated applications may not benefit from such testing. Further, when tracking ROC accuracy over time in systems with highly variable gallery characteristics, it may be the case that results are not stable over time. By contrast, FRS applications that threshold match scores will generally have more use for such ROC analysis, as they can use these measurements to tune such decision thresholds over time.

Additional Considerations and Best Practices:

When to perform testing:

- Operational testing should be performed before and after all major system updates (e.g., software, hardware, network) to ensure success of any such updates and measure changes in biometric matching accuracy. Such testing is imperative due to the volatility of system updates.

- Fixed-interval testing should also be performed. If the system is susceptible to attacks (e.g., cyber-security related), or has experienced recent errors, then fixed-interval testing should occur at a higher frequency.

- The system administrator should set operational and fixed-interval testing schedules based on the availability of computing resources and human effort. There is an expectation that Tier 1 and Tier 2 testing will be performed much more frequently than Tier 3 testing, which might be performed only before and after major system updates.

How to select test images:

- Operational test images should target enrollments that span different periods of time from the near term to the long term. This may help uncover any defects in system performance that result from template corruption, errors in updates, algorithm changes, or patches to extraction and/or matching algorithms.

- For Tier 2 and Tier 3 testing, when selecting testing images, subjects contained in those images who are recidivate (i.e., enrolled frequently in the database), should be avoided, as they could reduce the consistency of the accuracy reports.

- Subject to agency policies around retention of images, deceased subjects are ideal for enrolling into the database as test images, as deceased state mostly ensures no new encounters with test subjects. The MEDS database [2] contains deceased subjects, and is publicly available.

- Data may be collected from an uncontrolled set of test subjects that are reflective of the system's target population or a test crew that is representative of the system's target population, but that has not been enrolled in the system.

Other:

- If available, a non-operational system may be configured to be used in an "evaluation mode" to collect information not available during normal system operation.
- It is important that test data is representative of the operational database in terms of both subjects and image quality.
- Due to the introductory nature in this document, there are advanced configurations and operational concerns not addressed which include:
    - Use of score thresholds when testing

- Use of other accuracy statistics: Detection Error Tradeoff (DET), False Accept Rate (FAR), False Reject Rate (FRR)
- Demographic differentials in FRS algorithms
- Selection and continuity of imagery used for probe and database
- How to address identity ground truth in mate/imposter imagery

- These advanced issues and other considerations when testing (e.g., system setup and tuning) will be covered in future FISWG documents.

References

[1] *ISO/IEC (International Organization for Standardization/International Electrotechnical Commission). 2012. ISO/IEC 19795-6:2012 — Information technology — Biometric performance testing and reporting — Part 6: Testing methodologies for operational evaluation*. http://webstore.ansi.org/RecordDetail.aspx?sku=ISO/IEC+19795-6:2012.

[2] NIST/ITL (National Institute of Standards and Technology/Information Technology Laboratory). 2011. *NIST Special Database 32–Multiple Encounter Dataset (MEDS)*. http://www.nist.gov/itl/iad/ig/sd32.cfm.

[3] Arun Ross "Relating ROC and CMC Curves". https://www.nist.gov/system/files/documents/2020/09/15/12_ross_cmc-roc_ibpc2016.pdf

[4] 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

[5] P. Grother, M. Ngan, K. Hanaoka "Face Recognition Vendor Test (FRVT) Part 2: Identification" https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8271.pdf,

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