

INCITS M1/04-0676

**InterNational Committee for Information Technology Standards
INCITS Secretariat, Information Technology Industry Council (ITI)
1250 Eye St. NW, Suite 200, Washington, DC 20005
Telephone 202-737-8888; Fax 202-63-4922
email: ncits@itic.org**

**Title: Biometric Performance Testing and Reporting Part 1:
Framework**

Source: Project Editor

Date: 5 Oct. 2004 Version: 3.0

Revision	Date	M1 document #	Comments
1.0	11 August 2004	M1/04-0493	First Working Draft for review
2.0	19 Sept. 2004	M1/04-0566	2 nd Working Draft For review
3.0	5 Oct., 2004	M1/04-0xxx	3 rd Working Draft for review

Project editor contact information:

R. Michael McCabe
mccabe@nist.gov
Telephone: (301) 975-2932
FAX: (301) 975-5287

Introduction

This multipart standard develops a common set of methodologies and procedures to be followed for conducting technical performance testing and evaluations. Included are guidelines that address issues regarding required test sizes, performance statistics, error reporting, and presentation of performance results. These procedures can be incorporated in an "end-to-end" system approach or from an individual technical component perspective.

Three protocols for biometric testing and evaluation are addressed by this standard—technology, scenario, and operation. Associated with each biometric application and evaluation protocol are specific concepts, issues, and aspects that must be considered prior to test design and evaluation stages.

TABLE OF CONTENTS

1	SCOPE	5
2	CONFORMANCE	5
3	NORMATIVE REFERENCES	5
3.1	Standards under development	5
4	TERMS AND DEFINITIONS	5
5	BIOMETRIC SYSTEMS	7
5.1	Identification and Verification Systems	8
5.2	Positive and Negative Recognition Claims	8
6	EVALUATION PROTOCOLS	9
6.1	Technology Evaluation	9
6.2	Scenario Evaluation	9
6.3	Operational Evaluation	10
7	TESTING CHARACTERISTICS	10
7.1	Physical Environment	10
7.2	User Population	10
7.3	Habituated vs Non-habituated	11
7.4	Attended vs Unattended	11
7.5	Privacy	11
8	EVALUATION TEST DATA	11
8.1	Online / Offline Testing	11
8.2	Test Size	12
9	GENERALIZED TEST PROCEDURE	12

10	DATA ANALYSIS & REPORTING	13
	ANNEX A.....	15

1 Scope

This standard addresses testing the accuracy of identification and verification devices, algorithms, and systems. This standard does NOT address related performance issues such as throughput, turnaround-time, cost of ownership, life-time cycle costs, user implementations, environmental impact, cost/benefit breakpoints, etc.

This framework part of the multi-part Biometric Performance Testing and Reporting standard is intended to describe the remaining parts of the standard and show their relationships and common aspects. An overview of the primary testing protocols, biometric applications, and performance metrics is presented. It also provides guidance on data analysis techniques, recording of results, and performance reporting measures available.

2 Conformance

Any performance issues identified shall be resolved at a later date.

3 Normative References

3.1 Standards under development

INCITS 1602-D Part 2 Technology Testing and Reporting

INCITS 1602-D Part 3 Scenario Testing and Reporting

INCITS 1602-D Part 4 Operational Testing and Reporting

Biometric Vocabulary Harmonization, under development by ISO/IEC JTC1 SC37 WG1

4 Terms and Definitions

Terms will be defined by the Biometric Vocabulary Harmonization WG1 and those used in each of the remaining parts of this standard in addition to the following terms:

application:

a hardware/software system implemented to satisfy a [broad] set of requirements. In this context, an application incorporates a biometric system to satisfy a subset of requirements related to the verification or identification of an end user's identity so that the end user's identifier can be used to facilitate the end user's interaction with the system.

biometric sample:

raw data representing a biometric characteristic, which is captured and processed by the biometric system or the digital representation of a biometric characteristic used internally by a biometric system.

biometric system:

an automated system capable of capturing, extraction, matching and returning a decision (match / non-match).

biometrics:

the automated use of physiological or behavioural characteristics to determine or verify identity.

capture:

the method of taking a biometric sample from the end user.

Cummulative Match Characteristic curve (CMC):

curve drawn on a graph with ranks plotted on its lateral axis and cumulative match rate (CMR) on its vertical axis, representing the accuracy of an identification algorithm, device, or system

Detection Error Trade-off (DET) curve:

a DET curve plots error rates on both axes, giving uniform treatment to both types of error. The graph can then be plotted using logarithmic axes. This spreads out the plot and distinguishes different well performing systems more clearly than an ROC curve. DET curves can be used to plot matching error rates (False Non-match Rate against False Match Rate) as well as decision error rates (False Reject Rate against False Accept Rate).

enrolment/enrolment

the process of collecting biometric samples from a person and the subsequent preparation and storage of biometric reference templates representing that person's identity.

failure to acquire (FTA):

failure of a biometric system to capture and extract biometric data (comparison data).

failure to enroll (FTE):

failure of the biometric system to form a proper enrolment template for an end-user. The failure may be due to failure to capture the biometric sample or failure to extract template data (of sufficient quality).

false acceptance:

when a biometric system accepts a false claim on identity of an individual (i.e. made by an Impostor or a Defrauder). In a verification type system, false acceptance occurs as a result of a false match. Depending on the application, an identification type system false acceptance may occur as a result of a false match, or a false non-match.

false rejection:

when a biometric system rejects a genuine claim on identity of an individual. In both a verification type system and an identification type system, a false rejection occurs as a result of a false non-match.

identification/identify:

the one-to-many process of comparing a submitted biometric sample against some or all of the biometric templates stored in a database to determine an individual's identity.

Receiver Operating Characteristics (ROC) curve:

Receiver Operating Characteristic curves are an accepted method for showing the performance of pattern matching algorithms over a range of decision criteria. Such curves plot the “false alarm probability” (i.e. false acceptance rate) on the x-axis, and “correct detection probability” (i.e. 1- false rejection rate) on the y-axis.

template:

a data set which represents information extracted from a biometric sample or sample set and is used during biometric authentication as the basis for comparison with an incoming biometric sample.

verification:

the process of comparing a submitted biometric sample against the biometric reference template of a single enrollee whose identity is being claimed, to determine whether it matches the enrollee's template.

5 Biometric Systems

To determine if a biometric sample acquired from one subject matches another sample stored in a database, a partially or fully automated system of components is required to capture, process, and compare that biometric sample with one or more of those stored in a database. This begins with the acquisition of a raw biometric sample containing some physiological or behavioural characteristics of the subject. The hardware component used to capture the biometric will have associated with it certain criteria and procedures to be followed to ensure a high quality capture of the sample. An automated quality assessment apparatus may be used to indicate that the sample should be recaptured in order to acquire a better rendition of the raw input data. A feature extractor then processes this raw biometric sample to generate a more compact and salient representation of characteristics of the sample. This set of features extracted from the biometric sample is then compared against one or more templates stored in a database. These templates are a result of an initial enrolment process performed, whereby each new user establishes and registers their biometric feature set. Determining the degree of similarity between the features derived from a biometric sample and those contained in one or more stored templates is the fundamental core comparison function of a biometric system. This determination is usually followed by a decision process. During the

comparison process, a matching score is generated that is a measure of the similarity between features derived from the biometric sample and those contained in a stored template. A match or non-match decision can be based on whether the matching score exceeds a predetermined decision threshold cut-off. Scores exceeding the cut-off may be considered a match.

5.1 Identification and Verification Systems

Biometric systems can be divided into two general functional classes for identification and verification.

The identification system attempts to recognize a person without any claim to specific identity. A biometric sample from an unknown subject is presented to a biometric identification system that compares the new feature set against a database of biometric templates of known subjects. The resultant scores produced from these comparisons are used to determine and report the identity of the subject belonging to the unknown biometric sample. This is known as a "one-to-many" search or comparison – a single unidentified biometric template is compared to many identified ones.

The verification system verifies the claimed identity of a person. The biometric sample of a subject is presented to the biometric verifications system together with a claimed identity for that biometric sample. Based on the result of the comparison, the system either accepts or rejects the claim to that particular identity. This is known as a "one-to-one" comparison.

Both identification and verification are usually open-set tasks, referring to the possibility that a user may or may not be known to the system. For this reason two fundamental performance measures of a biometric system are needed: true match (1-false non-matches) performance to state correct acceptance of known persons, and true rejection (1-false match) performance to quantify correct rejection of persons not known to the system. These quantities can be traded-off against each other by means of a system threshold, and this effect establishes the detection error trade-off characteristic as the fundamental biometric performance statement.

A few identification applications involve closed-set recognition where it is known a priori that each user is known to the system. This renders false acceptance performance irrelevant.

5.2 Positive and Negative Recognition Claims

In addition to the identification or verification aspect of a system, consideration must be given to whether the system performance to be evaluated is based on positive or negative recognition claims. A subject who claims not to be enrolled in a biometric system is essentially making a negative claim. For example, a subject registering as a new enrollee in a system designed to provide public benefits, must not be a doubly registered recipient for those public benefits. On the other hand, a subject who must prove that their current presented biometric matches their previously enrolled biometric is making a positive claim for recognition. Such would be the case of that subject who had previously

enrolled in the system and is now making a claim to receive those benefits. The false reject and false alarm statistic have different meanings for each of these situations and must be viewed accordingly.

6 Evaluation Protocols

The approach used to test a system, to select the data, and to measure the performance is determined by the evaluation protocols selected. This standard addresses the technology, scenario, and operational protocols. Ordered by complexity, technology testing is probably the simplest, with the other two leading to greater levels of complexity. Technology testing may be viewed as being at the component level, scenario testing as system testing with live data capture in a simulated environment, and operational testing as system testing in the actual environment. Annex A presents a summary comparison of several features across the technology, scenario, and operational testing evaluations.

6.1 Technology Evaluation

This type of evaluation tests the technology itself. It is conducted using stored sample data, and without a live human volunteer corps interacting with sensors. Its primary use is in the assessment of underlying power or utility associated with a technology or biometric modality. It is suited therefore to the comparison of one or more, possibly competing, algorithms from a single technology. It can be used also to evaluate modular systems, to compare sample sets using a single technology, or to iteratively improve systems.

Technology tests are characterized by their strict repeatability and the possibility to employ very large legacy populations. They are disadvantaged by their inability to assess the live aspects of real biometric deployments.

6.2 Scenario Evaluation

The scenario evaluation is intended to test a biometric system in a manner that is as representative of the final real-world application as possible while still maintaining control of some of the vital performance related variables. It may be viewed as a pilot study for an operational evaluation. Information acquired from a scenario evaluation can be used for the benefit of an operational evaluation. Scenario testing is usually conducted using mature implementations of technologies that already known to be capable of meeting the overall performance requirements. It is usually conducted online with an attending human volunteer corps with the intention of evaluating complete systems, including the sensor, in a realistic environment.

Online scenario testing is suited to prediction of deployed performance and to identifying factors that determine it. It is capable of identifying ergonomic factors associated with human users, and to measuring overall throughput. It is only repeatable to the extent that the modelled scenario and the human population can be controlled. An online test may be supplemented with offline analysis if the captured sample data is retained.

6.3 Operational Evaluation

An operational evaluation tests a live system deployed in its native environment on its intended application. The primary goal of an operational evaluation is to determine if the performance of a complete biometric system meets the requirements of that specific application environment on the specific target population. It differs from a scenario test in that the population and environment are no longer controlled. In particular the presence of an impostor is not usually known in an operational test, and thus false acceptance performance cannot be directly quantified.

A well-defined program of operational evaluation can be used as an effective quality control feature for a specific population and environment. For example by tracking the performance a particular device and comparing it to other "well-performing" devices, decisions can be made regarding the replacement of a device with unsatisfactory performance.

Operational evaluations are difficult to repeat because of unknown and undocumented differences between operational environments. This is further complicated by the difficulty of obtaining the "ground truth" information, that is, who was actually presenting a good faith biometric measure.

7 Testing Characteristics

All kinds of evaluation will depend upon the environment, the population used, habituation, attended operation, and privacy.

7.1 Physical Environment

When testing biometric systems or components, environmental factors can have severe influences on performance. Accuracy of enrolment and verification processes can be affected by environmental conditions such as light, sound, dust, weather, humidity, and cleanliness of the capture station. The degree to which a process is affected will depend on the device itself. For example, iris and face recognition are dependent on the amount of light on the subject, while fingerprint capture can be affected by temperature, dirt, humidity, and cleanliness of the platen. Biometric devices may also need to be tested for their robustness under various environmental conditions and in particular for their dependence of security on environmental factors. Additionally, the accuracy of a sensor may be adversely affected by the aging process. Test evaluations should take factors like these into consideration and may require that test evaluations to be conducted under controlled environments that match the intended environments.

7.2 User Population

User population characteristics should also be taken into account regarding biometric system evaluations. The population used to test a technology will influence the accuracy performance. As an example, for a security application in a hospital, a population consisting of "white-collar" workers will provide better and more realistic fingerprint

images than a population of brick masons. Biometric devices depend on user population characteristics. Therefore, tests should be carried out on population samples that match the characteristics of the proposed user community. Consequently, any performance results obtained must be related not only to the environment but also the population chosen.

7.3 Habituated vs Non-habituated

The results obtained from habituated users (individuals presenting their biometric one or more times a day to gain entrance into a room) are different to those of non-habituated users (those presenting a biometric for entitlement benefits). The degree to which the users are habituated can have a significant effect on the results of the evaluation. Habituated users become accustomed to and better cooperate with the biometric acquisition process. This would be especially significant in scenario testing.

7.4 Attended vs Unattended

Whether a biometric device is located in an attended or unattended location will have an effect on the evaluation results. Devices located at attended locations provide the advantage of requesting an additional presentation of a biometric by the subject if the first sample acquired is of poor quality. At unattended locations the biometric system is forced to use what it is given regardless of the quality of the sample. However, automatic quality control procedures may be used to solicit another sample. This attribute coupled with the habituated aspect will impact results of a test, especially that of a scenario evaluation.

7.5 Privacy

Privacy must be a fundamental security function to be considered in an evaluation of a biometric system. The question of whether the use of biometrics and biometric products improve or degrade one's privacy is much debated. There must be adequate safeguards to protect the individual's unique information contained in their biometric identifier. Mechanisms must be in place to prevent disclosure of identity biometrics from third parties. Furthermore, individuals presenting biometric samples must be confident that disclosure of invasive information such as that which could be derived regarding health information and infirmities is protected. Finally, individuals presenting their biometrics must not have any fear of regeneration of their biometrics for other people or uses.

8 Evaluation Test Data

8.1 Online / Offline Testing

The direct enrolment or calculation of a matching score for a potential user performed at the time the image or signal is acquired is termed as "online" processing. It incorporates the interaction of a human with a biometric sensor. There are two apparent advantages of this mode of processing. First, a poorly acquired biometric image sample can be recaptured immediately thus reducing the failure to acquire or enrol concerns. Secondly,

once acquired the biometric image can be discarded saving the need for storage and for the system to operate in a non-standard manner.

The use of archived imagery, rather than the interaction of a human and biometric sensor, for enrolment or calculation of a matching score is termed as "offline" processing. This imagery may be the result of saved images from an "online" collection of biometrics. Offline processing provides greater control over the manner in which attempts and feature sets extracted from biometric sample images can be used in transactions. Repeatable evaluation of arbitrary biometrics can be performed using common test procedures, interfaces and metrics. Technology enhancement (such as a recalculation of a biometric sample's feature set) can also be evaluated in an offline testing mode. Additional benefits of offline testing include repeatability, interchange format testing, and the use of large populations.

Technology evaluations always rely on offline testing. For scenario and operational evaluations online transactions might be a better choice as they are simpler for the user as the system is operating in its usual manner and no special handling of images is necessary. But regardless of the test type evaluation, offline testing can be more appropriate than online testing for specific situations. For example, if the data to be used in the test is itself operational data, or is otherwise closely representative of data that will be acquired in deployment, then the test can be considered an operational test.

8.2 Test Size

Determination of test size will depend on both the unknown correlations and the anticipated error rates. As neither of these can generally be known in advance, it is recommended to use the largest test population that can be reasonably managed. The size of an evaluation in terms of subjects and attempts made will affect the accuracy of measured error rates biometric accuracy determination requires the use of large-scale realistic samples for testing. Without a large database of samples to test with, it is difficult to attain the required confidence to determine the biometric accuracy and error bounds on performance evaluations.

9 Generalized Test Procedure

The performance of a biometric system depends on its ability to discriminate accurately between samples. Its effectiveness can be defined primarily in terms of false match and non-match rates and failure to enrol and acquire rates. In the case where ground truth (the unique identity for each sample) is known, the preferred and most efficient technique to determine these measures is to perform a cross-comparison of all user samples with all enrolled samples. A similarity score is determined for each comparisons and this expresses the similarity between features derived from a presented sample and a stored template, or a measure of how well these features fit a user's reference model. A match/non-match decision may be made according to whether this score exceeds a decision threshold. These similarity scores are each recorded in the appropriate paired-entry cell of a similarity matrix with the y -axis addressing the user samples and the x -axis

the enrolled templates. The match and non-match distributions of the scores can be constructed from this matrix of scores and the required performance measures can be calculated and reported. In cases where ground truth is unknown, the reader is referred to the test methodology most appropriate to their system needs.

Regardless of the evaluation protocol being tested, consideration should be given to the originator of both the user feature set and the enrolment templates. Templates generated by different algorithms or devices may result in degraded performance as compared to those generated by the same source. This aspect of testing goes to the core of interoperability. For certain biometrics, the templates generated by different devices may not contain the same data and this will affect performance. Therefore, other than for interoperability testing, whenever possible all tests should rely on the use of the original sample data rather than proprietary templates from biometric system vendors to achieve the greatest accuracy. For technology testing this may be a straightforward and easy to achieve goal. However, for scenario or operational evaluations it may not be possible.

10 Data Analysis & Reporting

For each of the three evaluation protocols, performance statistics including error rates are to be reported. Generally, this analysis and reporting is divided into image acquisition analysis, verification system analysis, and identification system analysis. Although some of the details are treated uniquely by each of the protocols, the same general tools, approach, and presentations are used across the protocols.

The image acquisition analysis is divided into the failure to enrol rate and the failure to acquire rate. Failure to enrol includes subjects unable to present the required biometric feature or those unable to produce one or more images of sufficient quality suitable for enrolment. The failure to acquire category consists of those subjects for whom the system was unable to capture an image of sufficient quality. These two figures are meaningful in technology, scenario, and operational tests.

Verification performance is reported on a Receiver Operator Characteristic (ROC) plot. The purpose of a verification system is to simultaneously perform two tasks. The first is to correctly verify the identity of a person when the claim is legitimate. The second is to reject people who are not who they claim to be. Unfortunately, there is a trade-off between these two tasks, and one cannot simultaneously maximize the performance of both tasks. The performance statistic for verifying the identity is the probability of correct verification. This is the probability that a system will verify the identity of a legitimate claim. The performance statistic for rejecting false claims is the false accept rate. This is the probability that a false claim will be accepted as being true; i.e., someone fools the system and an unauthorized person is granted access. A ROC measures the trade-off between the probability of correct verification and the false alarm rate by plotting the false alarm rate on the x -axis against the true accept rate or the probability of a correct verification on the y -axis.

A Detection Error Trade-off (DET) is also used for this purpose. It is a modified ROC curve that plots the false reject rate (rather than the true match rate) along the y -axis. The DET curve plotted on logarithmic axes has the effect of spreading out the plot and distinguishes different well-performing systems more clearly.

The ROC and DET plots are also used for the identification system analysis. Guidelines and details for each evaluation protocol are presented in the other respective parts of this standard. However, there is a special case for identification analysis such that the enrolled population will always contain a mate for the user. This is known as a closed-set identification. For this situation a Cumulative Match Characteristic curve (CMC) is used to graphically illustrate the results. This curve plots the rank on the x -axis and the probability of identification at that rank or better on the y -axis.

Annex A

(Informative)

Comparison of Technology, Scenario, and Operational Tests

Type of Test	Technology	Scenario	Operational
Subject of testing	Biometric component (matching or extraction algorithm, sensor)	Biometric system	Biometric system
Ground truth	Known, subject to data collection errors and intersections in merged data sets	Known, subject to data collection errors and tester failure to note unwanted subject behavior	Unknown
User behavior controlled by experimenter	Not applicable during testing. May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled (unless user behavior is an independent variable)	Uncontrolled
User has real-time feedback of the result of attempt	No	Yes	Yes
Repeatability of results	Repeatable	Quasi-repeatable (if test scenario and population controlled)	Not repeatable
Control of physical environment	May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled and/or recorded	Not controlled, ideally Recorded
User interaction recorded	Not applicable during testing. Maybe recorded when biometric data recorded	Recorded	Recorded during enrollment. May be recorded during verification/identification
Typical results reported	Comparison of biometric components or versions of components (e.g., matching or extraction algorithms or sensors), Determine critical performance factors	Compare biometric systems, Determine critical performance factors. Measure simulated performance	Measure performance in an operational environment
Typical metrics	Most performance metrics. Not end-to-end throughput. Most error rates. Good for large-scale identification system performance where difficult to assemble large	Predicted end-to-end throughput, FMR, FNMR, FTA, FTE End-to-end throughput.	Operational FRR. Operational FAR?

	test crew.		
Constraints	Appropriate test database, e.g., gathered with one or more sensors, the identity of which may or may not be known.	Operational, instrumented system	Operational, instrumented system; typically only decision rates are available
Human test population	Recorded	Live	Live