INTERNATIONAL STANDARD

ISO/IEC 19795-1

First edition
2006-04-01

Information technology — Biometric performance testing and reporting —

Part 1: Principles and framework

Technologies de l'information — Essais et rapports de performance biométriques —

Partie 1: Principes et canevas

Reference number
ISO/IEC 19795-1:2006(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 19795-1 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

ISO/IEC 19795 consists of the following parts, under the general title *Information technology — Biometric performance testing and reporting*:

— *Part 1: Principles and framework*

— *Part 2: Testing methodologies for technology and scenario evaluation*

The following parts are under preparation:

— *Part 3: Modality-specific testing* [Technical Report]

— *Part 4: Performance and interoperability testing of data interchange formats*

— *Part 5: Performance of biometric access control systems*

Introduction

This part of ISO/IEC 19795 is concerned solely with the scientific “technical performance testing” of biometric systems and devices. Technical performance testing seeks to determine error and throughput rates, with the goal of understanding and predicting the real-world error and throughput performance of biometric systems. The error rates include both false positive and false negative decisions, as well as failure-to-enrol and failure-to-acquire rates across the test population. Throughput rates refer to the number of users processed per unit time based both on computational speed and human–machine interaction. These measures are generally applicable to all biometric systems and devices. Technical performance tests that are device-specific — for example, fingerprint scanner image quality — are not considered in this part of ISO/IEC 19795.

It is acknowledged that technical performance testing is only one form of biometric testing. Other types of testing not considered in this part of ISO/IEC 19795 include

— reliability, availability and maintainability;

— security, including vulnerability;

— conformance;

— safety;

— human factors, including user acceptance;

— cost/benefit;

— privacy regulation compliance.

Methods and philosophies for these other types of test are currently being considered internationally by a broad range of groups.

The purpose of this part of ISO/IEC 19795 is to present the requirements and best scientific practices for conducting technical performance testing. This is necessary because even a short review of the technical literature on biometric device testing over the last two decades or more reveals a wide variety of conflicting and contradictory testing protocols [1-11]. Even single organizations have produced multiple tests, each using a different test method. Test protocols have varied not only because test goals and available data are different from one test to the next, but also because no standard has existed for protocol creation.

Biometric technical performance testing can be of three types: technology, scenario or operational evaluation. Each type of test requires a different protocol and produces different types of results. Even for tests of a single type, the wide variety of biometric devices, sensors, vendor instructions, data acquisition methods, target applications and populations makes it impossible to present precise uniform testing protocols. Other parts of ISO/IEC 19795 will provide specific advice and requirements for the development and use of such different test protocols. This part of ISO/IEC 19795 addresses specific philosophies and principles that can be applied over a broad range of test conditions.

This part of ISO/IEC 19795 has been developed from the UK Biometrics Working Group’s Best Practices in Testing and Reporting Performance of Biometric Devices [12] which itself drew from two primary source documents developed by the US National Institute of Standards and Technology (NIST) [13, 14], a variety of evaluation reports [7-10], and comments from the Biometrics Consortium Working Group on Interoperability, Performance and Assurance.

Information technology — Biometric performance testing and reporting —

Part 1: Principles and framework

1 Scope

This part of ISO/IEC 19795

— establishes general principles for testing the performance of biometric systems in terms of error rates and throughput rates for purposes including prediction of performance, comparison of performance, and verifying compliance with specified performance requirements;

— specifies performance metrics for biometric systems;

— specifies requirements on test methods, recording of data and reporting of results; and

— provides a framework for developing and describing test protocols, to help avoid bias due to inappropriate data collection or analytic procedures, to help achieve the best estimate of field performance for the expended effort, and to improve understanding of the limits of applicability of the test results.

This part of ISO/IEC 19795 is applicable to empirical performance testing of biometric systems and algorithms through analysis of the matching scores and decisions output by the system, without detailed knowledge of the system's algorithms or of the underlying distribution of biometric characteristics in the population of interest.

Not within the scope of this part of ISO/IEC 19795 is the measurement of error and throughput rates for people deliberately trying to circumvent correct recognition by the biometric system (i.e. active impostors).

2 Conformance

To conform to this part of ISO/IEC 19795, a biometric performance test shall be planned, executed and reported in accordance with the mandatory requirements contained herein.

3 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 17025, *General requirements for the competence of testing and calibration laboratories*

4 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

4.1 Biometric data

4.1.1
sample
user's biometric measures as output by the data capture subsystem

EXAMPLE Fingerprint image, face image and iris image are samples.

NOTE In more complex systems, the sample may consist of multiple presented characteristics (e.g., 10-print fingerprint record, face image captured from different angles, left and right iris image pair).

4.1.2
features
digital representation of the information extracted from a sample (by the signal processing subsystem) that will be used to construct or compare against enrolment templates

EXAMPLE Minutiae coordinates and principal component coefficients are features.

4.1.3
template
model
user's stored reference measure based on features extracted from enrolment samples

NOTE The reference measure is often a template comprising the biometric features for an ideal sample presented by the user. More generally, the stored reference will be a model representing the potential range of biometric features for that user. In this part of ISO/IEC 19795, we normally use "template" to include "model".

4.1.4
matching score
similarity score
measure of the similarity between features derived from a sample and a stored template, or a measure of how well these features fit a user's reference model

NOTE 1 A match or non-match decision may be made according to whether this score exceeds a decision threshold.

NOTE 2 As features derived from a presented sample become closer to the stored template, similarity scores will increase.

4.1.5
verification decision
determination of the probable validity of a user's claim to identity in the system

4.1.6
candidate list
set of potential enrolled identifiers for a subject produced by an identification attempt (or by a pre-selection algorithm)

4.1.7
identification decision
determination of a candidate list for the user's probable identity in the system

4.2 User interaction with a biometric system

4.2.1
presentation
submission of a single biometric sample on the part of a user

4.2.2
attempt
submission of one (or a sequence of) biometric samples to the system

NOTE An attempt results in an enrolment template, a matching score (or scores), or possibly a failure–to-acquire.

4.2.3
transaction
sequence of attempts on the part of a user for the purposes of an enrolment, verification or identification

NOTE There are three types of transaction: enrolment sequence, resulting in an enrolment or a failure-to-enrol; a verification sequence resulting in a verification decision; or identification sequence, resulting in an identification decision.

4.2.4
genuine attempt
single good-faith attempt by a user to match their own stored template

4.2.5
zero-effort impostor attempt
attempt in which an individual submits his/her own biometric characteristics as if he/she were attempting successful verification against his/her own template, but the comparison is made against the template of another user

4.2.6
active impostor attempt
attempt in which an individual tries to match the stored template of a different individual by presenting a simulated or reproduced biometric sample, or by intentionally modifying his/her own biometric characteristics

NOTE Error rates for active impostor attempts will vary from those for zero-effort impostor attempts. Defining the methods and skill used in active impostor attempts is outside the scope of this part of ISO/IEC 19795.

4.2.7
presentation effects
broad category of variables affecting the way in which the users' inherent biometric characteristics are displayed to the sensor

EXAMPLE In facial recognition, this could include pose angle and illumination; in fingerprinting, finger rotation and skin moisture. In many cases, the distinction between changes in the fundamental biometric characteristic and the presentation effects may not be clear (e.g. facial expression in facial recognition or pitch change in speaker verification systems).

4.2.8
channel effects
changes imposed on the presented signal in the transduction and transmission process due to the sampling, noise and frequency response characteristics of the sensor and transmission channel

4.3 Personnel involved in the evaluation

4.3.1
user
person presenting a biometric sample to the system

4.3.2
test subject
user whose biometric data is intended to be enrolled or compared as part of the evaluation

4.3.3
crew
set of test subjects gathered for an evaluation

4.3.4
target population
set of users of the application for which performance is being evaluated

4.3.5
administrator
person performing the testing or enrolment

4.3.6
operator
individual with function in the actual system

EXAMPLE Staff conducting enrolments or overseeing verification or identification transactions.

4.3.7
observer
test staff member recording test data or monitoring the crew

4.3.8
experimenter
person responsible for defining, designing and analysing the test

4.3.9
test organization
functional entity under whose auspices the test is conducted

4.4 Types of evaluation

4.4.1
technology evaluation
offline evaluation of one or more algorithms for the same biometric modality using a pre-existing or specially-collected corpus of samples

4.4.2
scenario evaluation
evaluation in which the end-to-end system performance is determined in a prototype or simulated application

4.4.3
operational evaluation
evaluation in which the performance of a complete biometric system is determined in a specific application environment with a specific target population

4.4.4
online
pertaining to execution of enrolment and matching at the time of image or signal submission

NOTE Online testing has the advantage that the biometric sample can be immediately discarded, saving the need for storage and for the system to operate in a manner different from usual. However, it is recommended that images or signals are collected if possible.

4.4.5
offline
pertaining to execution of enrolment and matching separately from image or signal submission

NOTE 1 Collecting a corpus of images or signals for offline enrolment and calculation of matching scores allows greater control over which attempts and template images are to be used in any transaction.

NOTE 2 Technology testing will always involve data storage for later, offline processing. However, with scenario and operational testing, online transactions might be simpler for the tester – the system is operating in its usual manner and, although recommended, storage of images or signals is not absolutely necessary.

4.5 Biometric applications

4.5.1
verification
application in which the user makes a positive claim to an identity, features derived from the submitted sample biometric measure are compared to the enrolled template for the claimed identity, and an accept or reject decision regarding the identity claim is returned

NOTE The claimed identity might be in the form of a name, personal identification number (PIN), swipe card, or other unique identifier provided to the system.

4.5.2
identification
application in which a search of the enrolled database is performed, and a candidate list of 0, 1 or more identifiers is returned

4.5.3
closed-set identification
identification for which all potential users are enrolled in the system

4.5.4
open-set identification
identification for which some potential users are not enrolled in the system

4.6 Performance measures

4.6.1
failure-to-enrol rate
FTE
proportion of the population for whom the system fails to complete the enrolment process

NOTE The observed failure-to-enrol rate is measured on test crew enrolments. The predicted/expected failure-to-enrol rate will apply to the entire target population.

4.6.2
failure-to-acquire rate
FTA
proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality

NOTE The observed failure-to-acquire rate is distinct from the predicted/expected failure-to-acquire rate (the former may be used to estimate the latter).

4.6.3
false non-match rate
FNMR
proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample

NOTE The measured/observed false non-match rate is distinct from the predicted/expected false non-match rate (the former may be used to estimate the latter).

4.6.4
false match rate
FMR
proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template

NOTE The measured/observed false match rate is distinct from the predicted/expected false match rate (the former may be used to estimate the latter).

4.6.5
false reject rate
FRR
proportion of verification transactions with truthful claims of identity that are incorrectly denied

4.6.6
false accept rate
FAR
proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed

4.6.7
(true-positive) identification rate
identification rate
proportion of identification transactions by users enrolled in the system in which the user's correct identifier is among those returned

NOTE 1 This identification rate is dependent on (a) the size of the enrolment database, and (b) a decision threshold for matching scores and/or the number of matching identifiers returned.

4.6.8
false-negative identification-error rate
FNIR
proportion of identification transactions by users enrolled in the system in which the user's correct identifier is not among those returned

NOTE False-negative identification-error rate = 1 – true-positive identification rate.

4.6.9
false-positive identification-error rate
FPIR
proportion of identification transactions by users not enrolled in the system, where an identifier is returned

NOTE 1 The false-positive identification-error rate is dependent on (a) the size of the enrolment database, and (b) a decision threshold for matching scores and/or the number of matching identifiers returned.

NOTE 2 With closed-set identification false-positive identification is not possible, as all users are enrolled.

4.6.10
pre-selection algorithm
algorithm to reduce the number of templates that need to be matched in an identification search of the enrolment database

4.6.11
pre-selection error
⟨pre-selection algorithm⟩ error that occurs when the corresponding enrolment template is not in the pre-selected subset of candidates when a sample from the same biometric characteristic on the same user is given

NOTE In binning pre-selection, pre-selection errors happen when the enrolment template and a subsequent sample from the same biometric characteristic on the same user are placed in different partitions.

4.6.12
penetration rate
⟨pre-selection algorithm⟩ measure of the average number of pre-selected templates as a fraction of the total number of templates

4.6.13
identification rank
smallest value k for which a user's correct identifier is in the top k identifiers returned by an identification system

NOTE Identification rank is dependent on the size of the enrolment database, and should be quoted "rank k out of n".

4.7 Data presentation curves

4.7.1
detection error trade-off curve
DET curve
modified ROC curve which plots error rates on both axes (false positives on the x-axis and false negatives on the y-axis)

NOTE An example set of DET curves is shown in 10.6.2, Figure 3.

4.7.2
receiver operating characteristic curve
ROC curve
plot of the rate of false positives (i.e. impostor attempts accepted) on the x-axis against the corresponding rate of true positives (i.e. genuine attempts accepted) on the y-axis plotted parametrically as a function of the decision threshold

NOTE An example set of ROC curves is shown in 10.6.3, Figure 4.

4.7.3
cumulative match characteristic curve
CMC curve
graphical presentation of results of an identification task test, plotting rank values on the x-axis and the probability of correct identification at or below that rank on the y-axis

NOTE An example set of CMC curves is shown in 10.6.4, Figure 5.

4.8 Statistical terms

4.8.1
variance
V
measure of the spread of a statistical distribution

NOTE 1 If $\mathrm{E}(X)$ represents the distribution mean of a random variable X, then $V(X)=\mathrm{E}((X-\mu)^2)$, where $\mu = \mathrm{E}[X]$.

NOTE 2 The variance, if known, shows how close an estimated result is likely to be to its true value.

4.8.2
confidence interval
a lower estimate L and an upper estimate U for a parameter x such that the probability of the true value of x being between L and U is the stated value (e.g. 95 %)

EXAMPLE If $[L,U]$ is a (95 %) confidence interval for parameter x, then probability $\left(x \in \left[L,U\right]\right) = 95\,\%$.

NOTE The smaller the test size, the wider the confidence interval.

5 General biometric system

5.1 Conceptual diagram of general biometric system

Given the variety of applications and technologies, it might seem difficult to draw any generalizations about biometric systems. All such systems, however, have many elements in common. Biometric samples are acquired from a subject by a sensor. The sensor output is sent to a processor which extracts the distinctive but repeatable measures of the sample (the features), discarding all other components. The resulting features can be stored in the database as a template, or compared to a specific template, many templates or all templates already in a database to determine if there is a match. A decision regarding the identity claim is made based upon the similarity between the sample features and those of the template or templates compared.

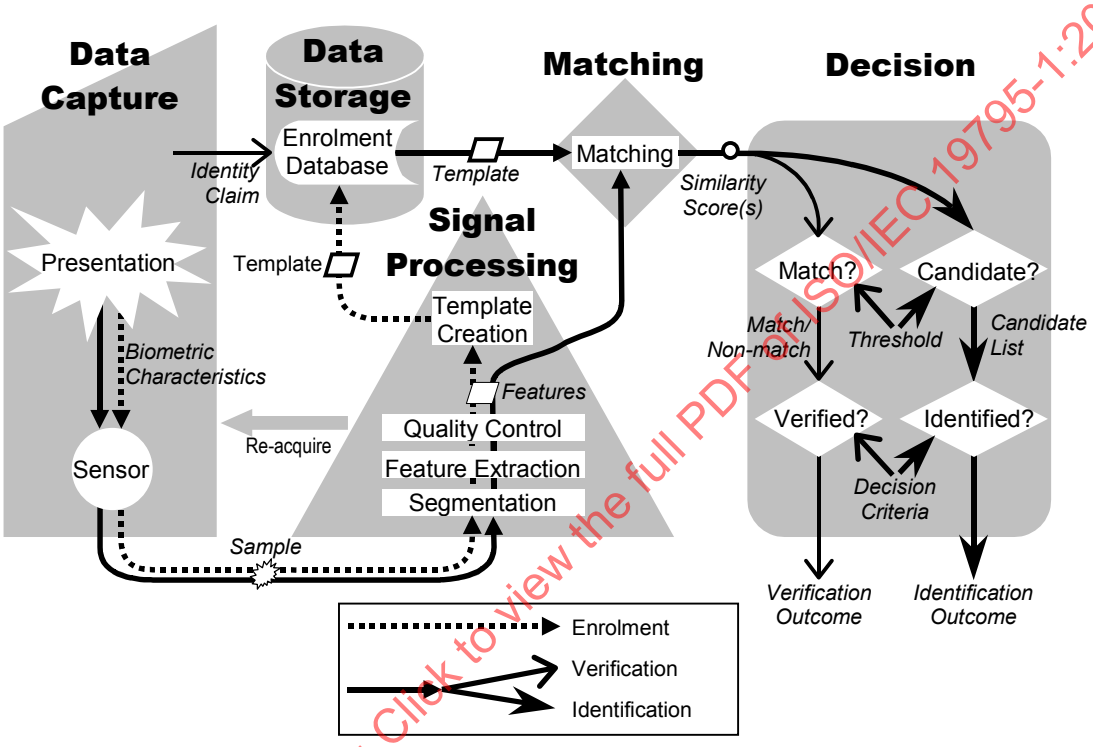


Figure 1 — Components of general biometric system

Figure 1 illustrates the information flow within a general biometric system consisting of *data capture*, *signal processing*, *storage*, *matching,* and *decision subsystems*. This diagram illustrates both enrolment, and the operation of verification and identification systems. The following subclauses describe each of these subsystems in more detail. It should be noted that, in any real biometric system, these conceptual components may not exist or may not directly correspond to the physical components, for example quality control could also take place before segmentation or before feature extraction.

5.2 Conceptual components of a general biometric system

5.2.1 Data capture subsystem

The *data capture subsystem* collects an image or signal of a subject's *biometric characteristics* that they have *presented* to the *biometric sensor*, and outputs this image/signal as a *biometric sample*.

5.2.2 Transmission subsystem (Not portrayed in diagram)

The *transmission subsystem* (not always present or visibly present in a biometric system) will transmit *samples, features,* and/or *templates* between different subsystems. *Samples, features* or *templates* may be transmitted using standard biometric data interchange formats. The biometric *sample* may be compressed and/or encrypted before transmission, and expanded and/or decrypted before use. A biometric *sample* may be altered in transmission due to noise in the transmission channel as well as losses in the compression/expansion process. It is advisable that cryptographic techniques be used to protect the authenticity, integrity, and confidentiality of stored and transmitted biometric data.

5.2.3 Signal processing subsystem

The *signal processing subsystem* extracts the distinguishing *features* from a biometric *sample*. This may involve locating the signal of the subject's *biometric characteristics* within the received *sample* (a process known as *segmentation*), *feature extraction*, and *quality control* to ensure that the extracted features are likely to be distinguishing and repeatable. Should *quality control* reject the received *sample/s*, control may return to the *data capture subsystem* to collect a further *sample/s*.

In the case of enrolment, the *signal processing subsystem* creates a *template* from the extracted biometric *features*. Often the enrolment process requires *features* from several presentations of the individual's *biometric characteristics*. Sometimes the *template* comprises just the *features*.

5.2.4 Data storage subsystem

Templates are stored within an *enrolment database* held in the *data storage subsystem*. Each *template* is associated with details of the enrolled subject. It should be noted that prior to being stored in the *enrolment database*, *templates* may be re-formatted into a biometric data interchange format. *Templates* may be stored within a biometric capture device, on a portable medium such as a smart card, locally such as on a personal computer or local server, or in a central database.

5.2.5 Matching subsystem

In the *matching subsystem*, the *features* are compared against one or more *templates* and *similarity scores* are passed to the *decision subsystem*. The *similarity scores* indicate the degree of fit between the *features* and *template/s* compared. In some cases, the *features* may take the same form as the stored *template*. For verification, a single specific claim of subject enrolment would lead to a single *similarity score*. For identification, many or all *templates* may be compared with the *features*, and output a *similarity score* for each comparison.

5.2.6 Decision subsystem

The *decision subsystem* uses the *similarity scores* generated from one or more attempts to provide the decision *outcome* for a verification or identification transaction.

In the case of verification, the *features* are considered to match a compared *template* when the *similarity score* exceeds a specified *threshold*. A claim about the subject's enrolment can then be verified on the basis of the *decision policy*, which may allow or require multiple attempts.

In the case of identification, the enrolee identifier or *template* is a potential *candidate* for the subject when the *similarity score* exceeds a specified *threshold*, and/or when the *similarity score* is among the highest k values generated for a specified value k. The *decision policy* may allow or require multiple attempts before making an identification decision.

NOTE Conceptually, it is possible to treat multi-biometric systems in the same manner as uni-biometric systems, by treating the combined biometric *samples/templates/scores* as if they were a single *sample/template/score* and allowing the *decision subsystem* to operate score fusion or decision fusion as and if appropriate.

5.2.7 Administration subsystem (Not portrayed in diagram)

The *administration subsystem* governs the overall policy, implementation and usage of the biometric system, in accordance with the relevant legal, jurisdictional and societal constraints and requirements. Illustrative examples include:

— providing feedback to the subject during and/or after data capture;

— requesting additional information from the subject;

— storage and format of the biometric *templates* and/or biometric interchange data;

— provide final arbitration on output from decision and/or scores;

— set *threshold* values;

— set biometric system acquisition settings;

— control the operational environment and non-biometric data storage;

— provide appropriate safeguards for end-user privacy;

— interact with the application that utilizes the biometric system.

5.2.8 Interface (Not portrayed in diagram)

The biometric system may or may not interface to an external application or system via an Application Programming Interface, Hardware Interface or a Protocol Interface.

5.3 Functions of general biometric system

5.3.1 Enrolment

In enrolment, a transaction by a subject is processed by the system in order to generate and store an enrolment template for that individual.

Enrolment typically involves:

— sample acquisition,

— segmentation and feature extraction,

— quality checks, (which may reject the sample/features as being unsuitable for creating a template, and require acquisition of further samples),

— template creation (which may require features from multiple samples), possible conversion into a biometric data interchange format and storage,

— test verification or identification attempts to ensure that the resulting enrolment is usable,

— and should the initial enrolment be deemed unsatisfactory, repeat enrolment attempts may be allowed (dependent on the enrolment policy).

5.3.2 Verification

In verification, a transaction by a subject is processed by the system in order to verify a positive specific claim about the subject's enrolment (e.g. "I am enrolled as subject X"). Verification will either accept or reject the

claim. The verification decision outcome is considered to be erroneous if either a false claim is accepted (false accept) or a true claim is rejected (false reject). Note that some biometric systems will allow a single end-user to enrol more than one instance of a biometric characteristic (for example, an iris system may allow end-users to enrol both iris images, while a fingerprint system may have end-users enrol two or more fingers as backup, in case one finger gets damaged).

Verification typically involves:

— sample acquisition,

— segmentation and feature extraction,

— quality checks, (which may reject the sample/features as being unsuitable for comparison, and require acquisition of further samples),

— comparison of the sample features against the template for the claimed identity producing a similarity score,

— judgement on whether the sample features match the template based on whether the similarity score exceeds a threshold, and

— a verification decision based on the match result of one or more attempts as dictated by the decision policy.

EXAMPLE In a verification system allowing up to three attempts to be matched to an enrolled template, a false rejection will result with any combination of failures-to-acquire and false non-matches over three attempts. A false acceptance will result if a sample is acquired and falsely matched to the enrolled template for the claimed identity on any of three attempts.

5.3.3 Identification

In identification, a transaction by a subject is processed by the system in order to find an identifier of the subject's enrolment. Identification provides a candidate list of identifiers that may be empty or contain only one identifier. Identification is considered correct when the subject is enrolled, and an identifier for their enrolment is in the candidate list. The identification is considered to be erroneous if either an enrolled subject's identifier is not in the resulting candidate list (false-negative identification error), or if a transaction by a non-enrolled subject produces a non-empty candidate list (false-positive identification error).

Identification typically involves:

— sample acquisition,

— segmentation and feature extraction,

— quality checks, (which may reject the sample/features as being unsuitable for comparison, and require acquisition of further samples),

— comparison against some or all templates in the enrolment database, producing a similarity score for each comparison,

— judgement on whether each matched template is a potential candidate identifier for the user, based on whether the similarity score exceeds a threshold and/or is among the highest k scores returned, producing a candidate list,

— an identification decision based on the candidate lists from one or more attempts, as dictated by the decision policy.

NOTE 1 In fully automated systems, the assigned identifier might be that corresponding to the template giving the highest similarity score (providing that this exceeds the specified threshold). When there is a human operator, the system may show a candidate list of the top *r* matches, for a final decision by the operator. Determining the ultimate performance metrics for systems using examination of possible matches by human operators is beyond the scope of this document.

NOTE 2 If all subjects using an identification system are known to be enrolled in the system, there can never be false-positive identification errors. In this degenerate case, known as closed-set identification, performance evaluation interest is normally focussed on how the rate of correct identification is related to the size of the candidate list returned.

5.4 Enrolment, verification & identification transactions

Each of the above biometric functions will depend on a user transaction. The transaction will consist of one or more attempts as allowed or required by the corresponding decision policy. For example, the decision policy may allow three attempts to verify; then the transaction may consist of one attempt, two attempts if the first attempt is rejected, or three attempts if the first two attempts are rejected.

Each attempt may consist of one or more presentations dependent on sensor operation, policy on sample quality, and any settings limiting the number of presentations or time permitted per attempt. For example an enrolment attempt may require the biometric sample to be submitted more than one time. Biometric verification systems often process a sequence of samples in a single attempt, for example: (a) collecting samples over some fixed period to find the best matching sample; (b) collecting samples until either a match is obtained or the system times out; or (c) collecting samples until one of sufficient quality is obtained, or the system times out.

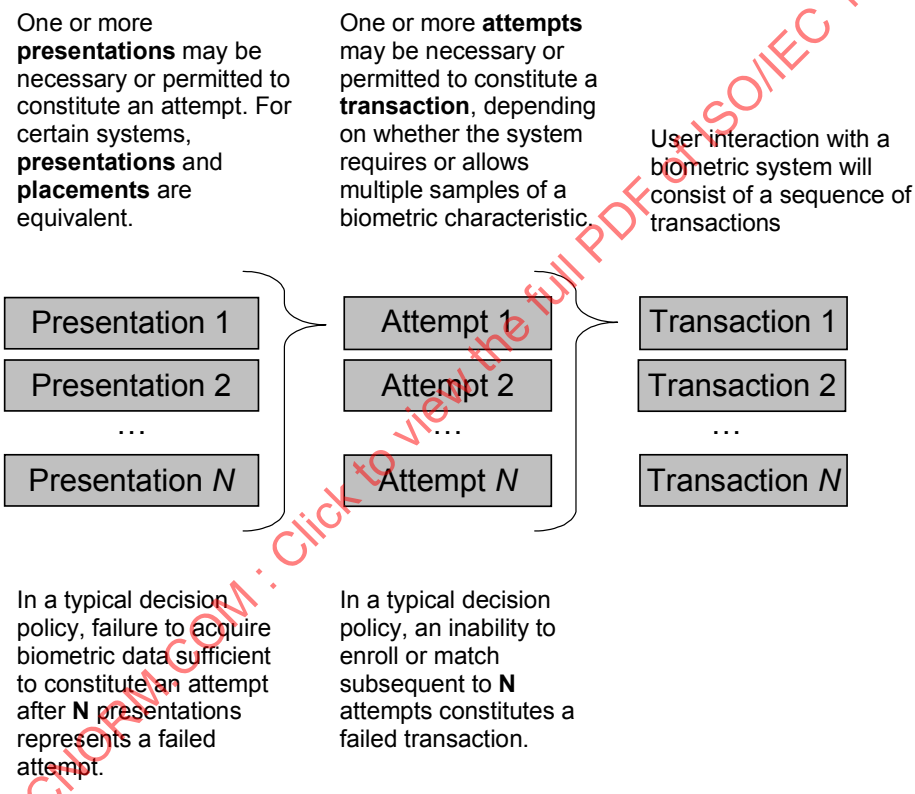


Figure 2 — Presentations, attempts and transactions

5.5 Performance measure

5.5.1 Error rates

Verification and identification decision errors are due to matching errors (i.e. false match and false non-match errors) or sample acquisition errors (i.e. failures-to-enrol and failures-to-acquire). How these fundamental errors combine to form decision errors will depend upon the number of comparisons required; whether there is a positive or negative claim of identity; and the decision policy, e.g. whether the system allows multiple attempts.

NOTE Though biometric performance has traditionally been stated in terms of the decision error rates, i.e. false accept rate and false reject rate, conflicting definitions are implicit in the literature. Literature on large-scale identification systems often refers to a “false rejection” occurring when a submitted sample is incorrectly matched to a template enrolled by another user. In access control literature, a “false acceptance” is said to have occurred when a submitted sample is incorrectly matched to a template enrolled by another user. False match rate and false non-match rate are not generally synonymous with false accept rate and false reject rate. False match or non-match rates are calculated over the number of comparisons, but false accept or reject rates are calculated over transactions and refer to the acceptance or rejection of the stated hypothesis, whether positive or negative. Furthermore, false accept or reject rates are inclusive of failures-to-acquire.

5.5.2 Throughput rates

Throughput rates show the number of users that can be processed per unit time based both on computational speed and human-machine interaction. These measures are generally applicable to all biometric systems and devices. Attaining adequate throughput rates is critical to the success of any biometric system. Throughput rates for verification systems, such as those for access control, are usually controlled by the speed of user interaction with the system in the process of submitting a good quality biometric sample. Throughput rates for identification systems, such as enrolment in a social service program, can be heavily impacted by the computer processing time required to compare the enrolment sample to the database of stored templates. So, depending upon the type of system, it may be appropriate to measure the interaction times of users with the system and also the processing rate of the computational hardware. Actual benchmark measurement of computer processing speed is covered in texts such as [12] and is considered outside the scope of this document. Measurement of the speed of the human-machine interaction will require a precise definition of the action with indicates the initiation of the interaction and the action that terminates the interaction. This definition should be determined prior to the start of the test and noted in the test report. The test report should also include a brief listing of the actions of the user included in the human-machine interaction.

5.5.3 Types of performance evaluation

Testing a biometric system will involve the collection of input images or signals, which are used for template generation at enrolment and for calculation of matching scores for verification or identification attempts. The images/signals collected can either be used immediately for an online enrolment, verification or identification attempt, or may be stored and used later for offline enrolment, verification or identification.

a) In a technology evaluation, testing of all algorithms is carried out on a standardized corpus, ideally collected by a “universal” sensor (i.e. a sensor that collects samples equally suitable for all algorithms tested). Nonetheless, performance against this corpus will depend on both the environment and the population in which it is collected. Although example data may be distributed for developmental or tuning purposes prior to the test, the actual testing needs to be done on data that has not previously been seen by algorithm developers. Testing is carried out using offline processing of the data. Because the corpus is fixed, the results of technology tests are repeatable.

b) In a scenario evaluation, testing is carried out on a complete system in an environment that models a real-world target application of interest. Each tested system will have its own acquisition sensor and so will receive slightly different data. Consequently, if multiple systems are being compared, care will be required that data collection across all tested systems is in the same environment with the same population. Depending on the data storage capabilities of each device, testing might be a combination of offline and online comparisons. Test results will be repeatable only to the extent that the modelled scenario can be carefully controlled.

c) In an operational evaluation, depending on the data storage capabilities of the operational system, offline testing might not be possible. In general, operational test results will not be repeatable because of unknown and undocumented differences between operational environments. Furthermore, “ground truth” (i.e. who was actually presenting a “good faith” biometric measure) can be difficult to ascertain, particularly if an operational evaluation is performed under unsupervised conditions without an administrator, operator or observer present.

Annex A summarizes the different characteristics of the different types of evaluations.

6 Planning the evaluation

6.1 General

As the first step in an evaluation, the experimenter shall determine:

a) the systems/application/environment to be evaluated;

b) the aspects of performance to be measured, and

c) how a dataset for evaluating performance will be created (i.e. which is the appropriate evaluation type: technology, scenario, or operational?).

These decisions form the basis for developing an appropriate test protocol, specifying appropriate environmental controls, test subject selection and test size.

NOTE The choice of evaluation type might be determined by, for example, the availability of a corpus of test samples for a technology evaluation, or of an installed system for an operational evaluation. There might also be circumstances in which all three types of testing would be carried out sequentially, perhaps gradually narrowing down modality options and systems under consideration for the eventual deployment of a biometric identification system.

6.2 Use of other parts of ISO/IEC 19795

Different systems and applications may require differences in test methodology, for example due to:

a) differences in environments;

b) differences in user populations (e.g. differences in user habituation);

c) differences in biometric modalities (e.g. due to different modalities being affected by different environmental conditions, and the differences between testing of predominantly behavioural and predominantly physiological biometrics);

d) differences in the performance metrics of interest (e.g. overall performance for verification, open-set identification, and closed-set identification are measured differently);

e) differences in the data available (e.g. identification systems which use pre-selection will not provide similarity scores for all sample features to template comparisons; however, the missing data cannot be treated as unknown: the sample is likely to have a poor similarity score when compared to any template that was not pre-selected); and

f) additional problems in establishing the ground truth for identification systems (where users do not present a claimed identity).

This part of ISO/IEC 19795 provides the basic principles for conducting and reporting a performance evaluation. Further parts of ISO/IEC 19795 may provide more specific guidance and requirements for particular types of evaluation, biometric modalities, target applications, or evaluation purposes.

6.3 Determine information about the system

The experimenter shall determine the following information about the system or systems to be tested in order to plan appropriate data collection procedures.

a) Does the system log transaction information? If not, then this information will need to be recorded manually by the test subject, operator or test observer.

b) Does the system save sample images or features for each transaction? This will be necessary if matching scores are to be generated offline.

c) Does the system return matching scores or just accept or reject decisions? In the latter case, data may have to be collected at a variety of security settings to generate a DET curve (see 7.2.3). If matching scores are returned, what information is available regarding parameters and scale?

d) Is the vendor's software developer's kit (SDK) available? Offline generation of genuine and impostor matching scores will require use of software modules from the SDK for:

   1) generating enrolment templates from enrolment samples;

   2) extracting sample features from the test samples; and

   3) generating the matching scores between sample features and templates. Matching scores produced by the offline codes should be equal to those produced by the live system. This may involve adjustment of parameters.

e) Are system modifications required for testing? Will required modifications alter system performance characteristics?

f) Does the system generate independent templates? The correct procedures for collecting or generating impostor transactions are different if templates are dependent (see 7.6.2.6 and 7.6.3.2).

g) Does the system use algorithms that adapt the template after successful verification? In such cases consideration shall be given as to how much template adaptation should occur prior to measuring performance; and also whether impostor testing might adversely affect the templates (see 7.4.4 and 7.6.1.4).

h) What are the recommended image quality and matching decision thresholds for the target application? These settings will affect the quality of presented samples, and error rates.

i) Are the expected approximate error rates known? This information can help in determining whether the test size is appropriate (see B.1).

j) What are the factors that will influence performance for this type of system? These will need to be controlled (see 6.4).

k) Does performance depend on the size of the enrolled database? This is the case for most identification systems, but also for some verification systems, such as those executing cohort enrolment, or those embedding a one-to-many search within the verification process.

NOTE In scenario and operational testing, any adjustments to the devices and their environment for optimal performance (including quality and decision thresholds) will need to take place prior to data collection. Stricter quality control can result in fewer false matches and false non-matches, but a higher failure-to-acquire rate. The decision threshold also needs to be set appropriately if matching results are presented to the user — positive or negative feedback will affect user behaviour. Vendors may be able to advise on the optimal environment and trade-off between settings.

6.4 Controlling factors that influence performance

6.4.1 Biometric system performance figures can be highly application-, environment- and population-dependent. Annex C provides a list of user, application, and environmental and system factors that have been found to affect the performance of one or more types of biometric system. How these factors are to be controlled shall be decided in advance of data collection.

6.4.2 Factors influencing the measured performance shall be explicitly or implicitly divided into one of four classes for control:

a) factors incorporated into the structure of the experiment (as independent variables) so that the effect they may have can be observed;

b) factors controlled to become part of the experimental conditions (unchanging throughout the evaluation);

c) factors randomized out of the experiment; and

d) factors judged to have negligible effect, which will be ignored. Without this final category the experiment would become unnecessarily complex.

This may involve some preliminary testing of systems to determine which factors are most significant and which may be safely ignored. In determining which factors to control, there may be a conflict between the needs for internal validity (i.e. differences in performance are due only to the independent variables recorded in the study) and external validity (i.e. the results truly represent performance on the target application).

EXAMPLE Suppose we are comparing the performance of two systems and are concerned over whether the skill or personality of the enrolment supervisor affects performance. Possible ways to control this factor are: a) to design the experiment to measure the performance differential between supervisors as well as between systems; b) to use only one supervisor, or to carefully script the supervisor/subject interaction to be as consistent as possible throughout the experiment; c) to allocate enrolment attempts randomly among all supervisors, thereby avoiding any systematic bias; or d) if there is prior evidence that differences between enrolment supervisors are small compared to the differences between systems, the experiment might ignore this factor.

6.4.3 For technology testing, a generic application and population may be envisioned, ensuring that the tests are neither too hard nor too easy for the systems being evaluated.

6.4.4 For scenario testing, a real-world application and population should be specified and modelled in order that the biometric system can be tested on representative users in a realistic environment.

6.4.5 In operational testing, the environment and the population are determined *in situ* with little control over them by the experimenter.

6.4.6 Of particular importance when planning the test is the time interval between enrolment and the collection of verification/identification data. Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as “template ageing”. This refers to the increase in error rates caused by time-related changes in the biometric pattern, its presentation, and the sensor. Collection of genuine transaction data shall therefore be separated in time from enrolment by an interval commensurate with the target application. If this interval is not known, then separation in time should be as long as is practicable. A rule-of-thumb is to separate the samples at least by the general time of healing of that body part.

EXAMPLE For fingerprints, two to three weeks should be sufficient. Eye structures possibly heal faster, allowing image separation of only a few days. Considering a haircut to be an injury to a body structure, facial images should perhaps be separated by one or two months.

NOTE Specific testing designed to test either user habituation (improving matching scores) or template ageing (degrading scores) requires multiple samples over time. Unless template ageing and habituation occur on different known time scales, there will be no way to de-convolve their counteracting effects.

6.5 Test subject selection

6.5.1 Both the enrolment and transaction functions require input signals or images. These samples should come originally from a test population or crew. If it is necessary to use artificially generated samples or features (including those created by modifying real data) such use shall be reported and justified, and the method of generation and assumptions regarding appropriateness shall be described. Results for synthetic and non-synthetic data shall be reported separately, and results for mixed synthetic & non-synthetic data shall report details of the mixture.

NOTE The use of artificially generated images would improve the internal validity of technology evaluations, as all the independent variables affecting performance are controlled. However, external validity is likely to be reduced. The corpus is also likely to be biased in respect of systems that model the biometric images in a similar way to that used in their generation.

6.5.2 The test crew shall not include people whose biometric characteristics have previously been used to develop or tune the biometric system being tested.

6.5.3 The crew should be demographically similar to that of the target application for which performance will be predicted from test results. This will be the case if test subjects are randomly selected from the potential users for the target application. In other cases volunteers will have to be relied on.

6.5.4 Recruiting the crew from volunteers could bias the tests. People with unusual features, the regularly employed or the physically challenged, for instance, could be under-represented in the sample population. Those with the strongest objections to the use of the biometric technology are unlikely to volunteer. It may be necessary to select unevenly from volunteers in order that the test crew is as representative as possible, and does not under-represent known problem cases. Current understanding of the demographic factors affecting biometric system performance is so poor that target population approximation will always be a major problem limiting the predictive value of tests.

6.5.5 Enrolment and testing is normally carried out in different sessions, separated by days, weeks, months or years, depending on the target application. A test crew with stable membership over this period is difficult to find, and it should be expected that some test subjects will drop out between enrolment and testing.

6.5.6 For overt target applications, the test crew should be appropriately instructed and motivated so that their behaviour follows that of the target application. If test subjects become bored with routine testing, they may be tempted to experiment or be less careful. Such possibilities shall be avoided.

6.5.7 For covert target applications, test subjects should ideally behave as if they were unaware of sample capture at the instant it happens. This might be achieved by passively capturing data over an extended period, and by using RFID tags to establish test subjects' correct identifier without needing their input.

6.5.8 When possible, test subjects should be fully informed about the required data collection procedure, aware of how the raw data will be used and disseminated and told how many sessions will be required and the durations of those sessions. Regardless of the use of the data, the identities of the crew should never be released. A consent form acknowledging that each test subject understands these issues should be signed, and shall be maintained in confidence by the experimenter.

NOTE For some types of testing, e.g. operational testing of covert identification system, informing test subjects may be impractical, or may alter their behaviour thereby invalidating the collected results.

6.6 Test size

6.6.1 General

The size of an evaluation, in terms of the number of test subjects and the number of attempts made (and, if applicable, the number of fingers, hands or eyes used per person) will affect how accurately error rates are measured. The larger the test, the more accurate the results are likely to be. Rules such as the Rule of 3 and Rule of 30 (described in B.1) may be used to provide lower bounds to the number of attempts needed for a given level of accuracy. However, these rules are over-optimistic, as they assume that error rates are due to a single source of variability, which is not generally the case with biometrics. Ten enrolment-test sample pairs from each of 100 people is not statistically equivalent to a single enrolment-test sample pair from each of 1000 people, and will not deliver the same level of certainty in the results.

NOTE As the test size increases, the variance of estimates decrease, but the scaling factor depends on the source of variability. For example, users may have differing error rates [16], giving a component of variance that scales as 1 / (number of test subjects) instead of 1 / (number of attempts). This effect is discussed in more detail in Annex B.

6.6.2 Collecting multiple transactions per user

6.6.2.1 The evaluation may collect multiple transactions from each test subject. Circumstances in which several transactions should be collected from each user include:

a) testing the effects of ageing, habituation, and other systematic variations;

b) testing of systems using template updating;

c) testing the extent to which different users have different individual error rates; or

d) when the transaction is not fully defined prior to testing, e.g. to determine how varying the number of attempts per transaction alters performance.

NOTE If the cost and effort of obtaining and enrolling the crew did not have to be considered, the ideal test might have many test subjects, each making a single transaction — this would provide independence between transactions. However, in real life, it is significantly easier to get existing enrolees to return than to find and enrol new test subjects. Furthermore, whenever an attempt is made, several additional attempts can be collected at the same time with marginally more effort. Such multiple transactions will show some correlations; nevertheless, it is often the case that using multiple transactions from fewer test subjects will produce a smaller uncertainty in test results than a trial of equal cost using a single transaction from slightly more test subjects.

6.6.2.2 The number and frequency of test transactions collected per test subject should be in keeping with the target application. The test plan may vary this provided that the altered pattern of transactions does not significantly affect error rates.

NOTE User behaviour can vary with each successive attempt due to increased familiarity with the device or feedback of their authentication results. For example, the first attempt a user makes might have a higher failure rate than any following attempts. As a result, the observed false non-match rate will depend on the pattern of attempts per user, as defined by the test protocol. Generally, error rates will be measured averaged not only over the target population, but also over the types of attempt a user might reasonably make. Averaging over multiple attempts can help in this case. However, there is the possibility that altering the number and pattern of attempts per user might affect user behaviour enough to significantly affect the measured error rates.

EXAMPLE Use of multiple transactions would be inappropriate in tests where the user should be unfamiliar with the device or biometric application.

6.6.3 Recommendations on test size

The number of people tested is more significant than the total number of attempts in determining test accuracy.

a) The crew shall be as large as practicable. The measure of practicality is likely to be the expense of crew recruitment and tracking.

b) Sufficient samples shall be collected per test subject so that the total number of attempts exceeds that required by the Rule of 3 or Rule of 30 as appropriate. If it is possible to collect these multiple samples on different days, or from different fingers, eyes or hands (and the additional samples are still representative of normal use[1)], doing so can help reduce the dependencies between samples by the same person.

c) Once data has been collected and analysed, the uncertainty in the performance measures shall be estimated, determining whether the test was large enough.

NOTE The law of diminishing returns applies — a point will be reached where errors due to bias in the environment used, or in test subject selection, will exceed those due to size of the crew and number of tests.

6.7 Multiple tests

6.7.1 The cost of data collection is so high that it can be desirable to conduct multiple tests with one data collection effort. Technology evaluation allows for this. In the case of biometric devices for which image standards exist (fingerprint [18], face [19], iris [20] and voice [21]), a single corpus may be collected for offline testing of pattern-matching algorithms from multiple vendors. In effect this decouples the data collection and signal processing subsystems. This is not problem-free however, as these subsystems are usually not completely independent. For instance, the quality control module, which may require the data collection subsystem to reacquire the image, is part of the signal processing subsystem. Furthermore, even if image

1) For example, use of the little finger is probably not representative of normal use of a fingerprint system, and the resulting error rates will be different [17]. Similarly, an inverted left hand would not be representative in a right-handed hand geometry system.

standards exist, image quality is affected by the vendor-specific user interfaces that guide the data collection process. Consequently, offline technical evaluation of algorithms using a standardized corpus may not give a good indication of total system performance, and can also be biased in favour of some systems and against others.

6.7.2 Scenario evaluations of multiple systems can also be conducted simultaneously by having a test crew use several different devices or scenarios in each session. However, this approach will require care. One possible problem is that the test subjects will become habituated as they move from device to device. To equalize this effect over all devices, the order of their presentation to each test subject should be randomized. A further potential problem occurs where ideal behaviour for one device conflicts with that for another. For example, some devices work best with a moving image, while others require a stationary image. Such conflicts may result in lower-quality test images for one or more of the devices under test.

7 Data collection

7.1 Avoidance of data collection errors

7.1.1 Collected biometric image samples or features are properly referred to as a corpus. The information about those images and the users who produced them is referred to as the metadata. Both the corpus and the metadata can be corrupted by human error during the collection process. Error rates in the collection process may easily exceed those of the biometric device. For this reason, extreme care shall be taken during data collection to avoid both corpus (misacquired image) and metadata (mislabelled image) errors.

7.1.2 Typical corpus errors are:

a) test subjects using the system incorrectly (and outside the limits allowed by the experimental controls), e.g. mistakenly using a fingerprint scanner upside down; and

b) cases where a blank or corrupt image is acquired if the user enters a PIN, but moves on before a proper image is captured.

7.1.3 Possible causes of metadata errors include:

a) test subjects being issued with the wrong PIN;

b) typing errors in PIN entry; and

c) using the wrong body part, e.g. using a middle finger when the index finger is required.

7.1.4 Data collection software minimizing the amount of data requiring keyboard entry, multiple collection personnel to double-check entered data and built-in data redundancy shall be used. Supervisors shall be familiar with the correct operation of the system and the possible errors to guard against. To avoid a variable interpretation of what constitutes a misacquired sample, objective criteria shall be set in advance. Any unusual circumstance surrounding the collection effort, and the transactions affected, shall be documented by the collection personnel.

7.1.5 Even with precautions, some data collection errors are likely to be made, which will add uncertainty to the measured test results. After-the-fact correction of corpus or metadata errors should be based on redundancies built into the collection system, and should not be solely reliant on the outputs of the tested biometric algorithm. In this respect, systems that can save sample images and/or transaction logs offer more scope for error correction than systems where all the details have to be recorded manually.

7.1.6 Collection personnel shall not manually discard nor use an automated mechanism to discard collected samples unless the samples conform to some formal, pre-determined, documented and reported exclusion criteria. The number of samples so excluded shall be reported.

EXAMPLE Exclude fingerprint sample if inked area is less than 0,25 cm^2.

7.2 Data and details collected

7.2.1 The data that can be collected automatically will depend on the biometric system implementation. For the purposes of evaluation, ideally the systems should automatically log all enrolment, verification or identification attempts, including details of claimed identity and matching and quality scores, and if possible, also save sample images or features. This brings the following advantages:

a) enrolment templates and matching scores may be generated offline, provided that the vendor SDK is available — this will allow for a full cross-comparison of sample features and templates, giving a higher number of impostor scores;

b) the collected images may be reused to evaluate algorithm improvements or (provided the images are in a suitable format) to evaluate other algorithms in a technology evaluation;

c) potential corpus or metadata errors may be checked by visually inspecting the images or through examining the transaction log; and

d) the amount of data that should to be recorded by hand, and the potential for transcription errors, is minimized.

7.2.2 Many biometric systems do not provide this ideal functionality in their normal mode of operation. With vendor co-operation, it may be possible to incorporate this functionality into an otherwise standard system, but care should be taken that system performance is not affected. For example, the time taken in logging images may slow the system and affect user behaviour. If sample images or features cannot be saved, enrolment, genuine and impostor transactions shall be conducted online, and results recorded manually if necessary. This shall require closer supervision by the test personnel to ensure that all results are logged correctly.

7.2.3 Some systems do not return matching scores, but just a match or non-match decision at the current security setting. To plot a detection-error trade-off (DET) graph in such cases, genuine and impostor attempt data shall be collected or generated at a number of security settings. The vendor may advise on the appropriate range of security settings. The selected values for the security setting (which could be “low”, “medium” and “high”) will parameterize the DET curve in place of the decision threshold. In the case of online testing, for correct estimation of error rates each user shall make transactions at each chosen security setting.

NOTE Test protocols could be constructed such that a user’s genuine attempts are made at progressively more lenient settings, stopping when a match is obtained; and a user’s impostor attempts are made at progressively stricter settings, stopping once a non-match is obtained. Such protocols will not conform to this part of ISO/IEC 19795, as the effects of multiple attempts will be confounded with those of changing the decision threshold.

7.2.4 The data collection plan may need to include a mechanism whereby a subject may request their samples and biographical information to be expunged from the system. Otherwise such redaction may be time-consuming and error-prone.

7.3 Enrolments

7.3.1 Enrolment transactions

7.3.1.1 Each test subject shall enrol only once, though an enrolment may generate more than one template (for example a template for each fingerprint, or multiple face poses). Multiple attempts at enrolment may be allowed to achieve one good enrolment. Care shall be taken to prevent accidental multiple enrolments.

7.3.1.2 Practice tests may be performed at the time of enrolment to ensure that the enrolment samples are of sufficient quality to produce a later match, and to familiarize test subjects with the system. Scores resulting from such practice tests should not be recorded as part of the genuine comparison record (unless performance immediately after enrolling is being measured).

7.3.1.3 If available, enrolment samples should be recorded.

7.3.2 Enrolment conditions

7.3.2.1 Enrolment conditions should model the target application enrolment. The taxonomy [22] of the enrolment environment will determine the applicability of the test results. Vendor recommendations should be followed and the details of the environment should be noted. The noise environment requires special care. Noise can be acoustic, in the case of speaker verification, or optical, in the case of eye, face, finger or hand imaging systems. Lighting "noise" is of concern in all systems using optical imaging, particularly any light falling directly on the sensor and uncontrolled reflections from the body part being imaged. Lighting conditions should reflect the proposed system environment as closely as possible. It is especially important to note that test results in one noise environment will not be translatable to other environments.

7.3.2.2 Every enrolment shall be carried out under the same general conditions. Many data collection efforts have been ruined because of changes in the protocols or equipment during the extended course of collection[2)]. The goal should be to control presentation and transmission channel effects so that such effects are either uniform across all test subjects, or randomly varying across test subjects.

7.3.2.3 As the tests progress, an enrolment supervisor may gain additional working knowledge of the system, which could affect the way later enrolments are carried out. To guard against this, the enrolment process and criteria for supervisor intervention shall be determined in advance, and adequate supervisor training shall be provided.

7.3.3 Enrolment failures and presentation errors

7.3.3.1 The biometric system may prevent acceptance of some enrolment attempts. Quality control modules for some systems requiring multiple images for enrolment will not accept images that vary highly between presentations; other quality control modules will reject single poor-quality images. If these modules allow for tuning of the acceptance criteria, vendor advice should be followed. Multiple enrolment attempts may be allowed, with a predetermined maximum number of attempts or maximum elapsed time. All quality scores and enrolment samples should be recorded. Advice or remedial action to be taken with users who fail an enrolment attempt shall be predetermined as part of the test plan.

7.3.3.2 The proportion of test subjects failing to enrol at the chosen criteria shall be recorded and reported. If possible the reasons for enrolment failure should also be recorded and reported (e.g. those without the biometric characteristics, cases where a sample could not be acquired, or failures/exceptions of the enrolment algorithm, or those unable to successfully verify in practice attempts).

7.3.3.3 Not all quality control is automatic. Intervention by the experimenter may be required if the enrolment measure presented was inappropriate according to some predetermined criteria. For instance, enrolling users may present the wrong finger, hand or eye, recite the wrong enrolment phrase or sign the wrong name. This data should be removed, but a record of such occurrences shall be kept.

7.3.3.4 Data editing to remove inappropriate biometric presentations may have to be based on removal of outliers, but the effect of this on resulting performance measures shall be fully noted. Enrolment data shall not be removed simply because the enrolled template is an outlier.

7.4 Genuine transactions

7.4.1 Genuine transaction data shall be collected in an environment, including noise, that closely approximates the target application. This test environment shall be consistent throughout the collection process. The motivation of test subjects, and their level of training and familiarity with the system, should also mirror that of the target application.

2) A famous example is the "Great Divide" in the KING speech corpus [23]. About halfway through the collection, for a reason nobody now remembers, the recording equipment had to be temporarily disassembled. It was later reassembled according to the original wiring diagram; nonetheless the frequency response characteristics were slightly altered, creating a divide in the data and complicating the scientific analysis of algorithms based on the data.

NOTE In the case of technology evaluation, the target application refers to the envisaged application being neither too hard nor too easy for the capabilities of the algorithms to be tested.

7.4.2 The collection process should ensure that presentation and channel effects are either uniform across all users or randomly varying across users. If the effects are held uniform across users, then the same presentation and channel controls in place during enrolment should be in place for the collection of the test data. Systematic variation of presentation and channel effects between enrolment and test data will lead to results distorted by these factors. If the presentation and channel effects are allowed to vary randomly across test subjects, there shall be no correlation in these effects between enrolment and test sessions across all users.

7.4.3 In the ideal case, between enrolment and the collection of test data, test subjects should use the system with the same frequency as the target application. However, this may not be a cost-effective use of the test crew. It may be better to forego any interim use, but allow re-familiarization attempts immediately prior to test data collection.

7.4.4 For systems that may adapt the template after successful verification, some interim use between enrolment and collection of genuine attempt and transaction data may be appropriate. The amount of such use should be determined prior to data collection, and should be reported with results.

7.4.5 The sampling plan shall ensure that the data collected are not dominated by a small group of excessively frequent, but unrepresentative users.

7.4.6 Great care shall be taken to prevent data entry errors and to document any unusual circumstances surrounding the collection. Keystroke entry on the part of both test subjects and test administrators should be minimized. Data could be corrupted by impostors or genuine users who intentionally misuse the system. Every effort shall be made by test personnel to discourage these activities, however, data shall not be removed from the corpus unless external validation of the misuse of the system is available.

7.4.7 Users are sometimes unable to give a usable sample to the system as determined by either the test administrator or the quality control module. Test personnel should record information on failure-to-acquire attempts where these would otherwise not be logged. The failure-to-acquire rate measures the proportion of such attempts, and is quality threshold dependent. As with enrolment, quality thresholds should be set in accordance with vendor advice.

NOTE Quality threshold (and decision threshold) settings may influence the performance of users — stricter thresholds encourage more careful presentation of the biometric pattern, looser thresholds allow more sloppiness. The corpus itself may therefore not be as threshold-independent as assumed.

7.4.8 Test data shall be added to the corpus regardless of whether or not it matches an enrolled template. Some vendor software does not record a measure from an enrolled user unless it matches the enrolled template. Data collection under such conditions would be severely biased in the direction of underestimating false non-match error rates. If this is the case, non-match errors shall be recorded by hand. Data shall be excluded only for predetermined causes independent of comparison scores.

7.4.9 All attempts, including failures-to-acquire, shall be recorded. In addition to recording the raw image data if practical, details shall be kept of the quality measures for each sample if available and, in the case of online testing, the matching score or scores.

7.5 Identification transactions of users enrolled in the system

7.5.1 Identification transactions for enrolled users shall be collected and recorded in the same general way as genuine verification transactions. The recorded outcome will consist of the list of candidate identifiers. If matching scores or quality scores are produced by the system these should also be recorded.

7.5.2 Identification transactions may be generated offline, including by simulating the identification process as a set of verification transactions against each template in the database. In the general case, however, identification may use pre-selection to limit the number of templates compared by the matching algorithm.

NOTE In order to determine performance of the pre-selection algorithm, the number of pre-selected templates for each identification attempt should be recorded (see Annex D).

7.6 Impostor transactions

7.6.1 General

7.6.1.1 Impostor transactions shall be generated online or offline.

a) Online impostor transactions involve test subjects submitting samples to be compared against other people's enrolment templates.

b) Offline impostor transactions are generated by comparing against enrolment templates, features extracted from samples collected either in genuine transactions or from a distinct set of transactions by unenrolled test subjects. Offline computation allows a full cross-comparison approach in which every sample feature is compared against every non-self template.

7.6.1.2 The type of evaluation will often determine whether online or offline impostor transactions will be used.

a) In technology evaluation, impostor transactions are always analysed offline. However, occasionally there may be a corpus of impostor attempts, to be analysed instead of or in addition to the set of cross-comparison impostor transactions.

b) For scenario evaluation: the most appropriate method will probably depend on whether the system is able to save samples from genuine transactions. If so, cross-comparison will generate many more impostor attempts than could be achieved through online use of test subjects.

c) For operational evaluations, development of impostor scores may not be straightforward. If the operational system saves sample images or extracted features, impostor scores can be computed offline. If, as is likely, this data is not saved, impostor scores can be obtained through online testing. Because of the non-stationary statistical nature of the data across users, it is preferable to use many impostor test subjects, each challenging few randomly chosen non-self templates, than to use a few test subjects challenging many non-self templates. In some cases, the use of inter-template comparisons for impostor transactions may be appropriate.

7.6.1.3 Impostor transactions shall not be based on intra-individual comparisons. With some biometric modalities, the user may be able to present different biometric instances — e.g. any of up to ten fingers, left or right eye, etc. To improve the independence of different samples from a single test subject, an evaluation could allow enrolment of more than one finger, hand or eye as different (sub-) identities. However, within-individual comparisons are *not* equivalent to between-individual comparisons, and shall not be included in the set of impostor transactions.

EXAMPLE Different fingerprints from the same person will have a similar number of fingerprint ridges, and are more likely to match than fingerprints from different people.

7.6.1.4 For systems that may adapt the template after successful verification, this facility should be disabled during impostor transactions. If this is not possible, collection of impostor transactions should be delayed until all genuine test transactions are collected.

7.6.2 Online collection of impostor transactions

7.6.2.1 Online impostor transactions are collected by having each test subject make zero-effort impostor attempts against each of a pre-determined number of non-self templates randomly selected from all previous enrolments (sometimes from all previous enrolments within the same demographic group). The random selection shall be independent between users.

NOTE The use of background databases of biometric samples or templates acquired from different (possibly unknown) environments and populations is not considered best practice.

7.6.2.2 Resulting impostor scores shall be recorded, together with the true identity of both the impostor and the impersonated template. As it is likely that these impostor transactions are taking place alongside genuine transactions, care shall be taken that results are attributed to the correct set of scores.

7.6.2.3 Impostor attempts shall be made under the same conditions as the genuine attempts.

7.6.2.4 If a test subject is aware that an impostor comparison is being made, changes in presentation behaviour may result in unrepresentative results, particularly with biometric systems that are based on predominantly behavioural characteristics. Therefore, to avoid even subconscious changes in presentation, test subjects should ideally not be told whether the current comparison is a genuine or impostor transaction.

7.6.2.5 Impostor transactions may be collected before all test subjects have enrolled. Though the first enrolled templates will have a higher probability of being targeted for an impostor comparison, this will not bias the calculation of impostor error rates if, as is usually the case, test subjects are enrolled in an order that has no regard to the quality of their biometric measures.

7.6.2.6 For systems that have dependent templates, test subjects making impostor attempts shall not be enrolled in the database when the attempt is made (except in the case of closed-set identification). This can involve selecting a subset of test subjects who will not be enrolled in the system and so can be used as impostors.

7.6.3 Offline generation of impostor transactions

7.6.3.1 General

7.6.3.1.1 Offline impostor comparisons are made in the same basic way as online comparisons, by either:

— randomly selecting with replacement both samples and templates for the non-self comparisons;

— randomly selecting, for each genuine sample, a number of non-self templates from all those enrolled for comparison with the sample features (random selection of templates being independent for each sample); or

— performing a full cross-comparison, in which each sample feature is compared with every non-self template.

7.6.3.1.2 Offline development of matching scores should be carried out with software modules of the type available from vendors in SDKs. One module will create templates from enrolment samples. A second module will create sample features from test samples. These modules will sometimes be the same piece of code. A third module will return a matching score for any assignment of a sample feature to a template. If processing time is not a problem, features from all genuine samples should be compared to all non-self templates. If there are T templates and N features (from the same test crew), $N(T-1)$ comparisons against non-self templates can be performed. These impostor comparisons will not be statistically independent, but this approach is statistically unbiased and represents a more efficient estimation technique than the use of randomly chosen impostor comparisons [24].

7.6.3.1.3 Many biometric systems collect and process a sequence of samples in a single attempt, for example:

a) collecting samples over some fixed period, and scoring the best matching sample;

b) collecting samples until either a match is obtained or the system times out;

c) collecting samples until one of sufficient quality is obtained, or the system times out; or

d) collecting a second sample when the score from the first sample is very close to the decision threshold.

In such cases, a single sample from a genuine attempt might not be suitable as an impostor sample. In case a), the sample saved will be the one that best matches the genuine template. However, an impostor attempt would be based on the sample best matching the impersonated template. To determine whether it is appropriate to base cross-comparison on a single genuine sample, the following two questions shall be addressed.

— Does the saved sample depend on the template being compared?

— If so, does this materially affect the matching scores generated?

If the answers to both these questions are yes, then either the whole sample sequence shall be saved and used in offline analysis, or impostor scores shall be generated online.

7.6.3.2 Offline generation of impostor transactions when templates are dependent

7.6.3.2.1 For systems with dependent templates, unbiased impostor scores may be generated using a jack-knife approach to create the enrolment templates. The jack-knife approach is to enrol the entire crew with a single test subject omitted. This omitted test subject is then used as an unknown impostor, comparing their sample features to all enrolled templates. This enrolment process is repeated for each crew member, and a full set of impostor scores can be generated.

7.6.3.2.2 A simpler technique may be used, in which the test crew is randomly partitioned into impostors and enrolees. Offline enrolment ignores the data from impostor test subjects, while offline false match scoring ignores data from enrolee test subjects. This is a less efficient use of the data than the jack-knife approach.

7.6.3.3 Offline generation of impostor transactions using inter-template comparisons

Cross-comparison of enrolment templates may sometimes provide impostor scores. This can be useful, for example, in operational evaluations where samples or features of transactions are not saved. Each of N test (or enrolment) templates can be compared to the remaining $(N-1)$ test (or enrolment) templates. Template cross comparison shall not be used unless:

a) enrolment and verification require the same user input (for example both require a single presentation);

b) enrolment and verification use the same algorithms to extract and encode sample features; and

c) quality control for enrolment is the same as for verification attempts.

If these requirements are not fulfilled, template cross-comparison is likely to result in biased estimation of impostor scores [22]. This is true whether the enrolment template is averaged or selected from the best enrolment sample. No methods currently exist for correcting this bias.

7.7 Identification transactions of users not enrolled in the system

7.7.1 Estimation of the false-positive identification-error rate requires identification transactions by test subjects not enrolled in the system. These shall not be test subjects who failed enrolment.

7.7.2 All identification attempts should be recorded, together with the subject's identifier, the resulting lists of candidate identifiers and, if available, matching scores. Identification transactions of enrolled users and of not enrolled users should be made under the same conditions.

7.7.3 Identification transactions may be collected against portions of the enrolment database of various sizes to record how identification performance varies with database size.

7.7.4 If the enrolment and identification samples of enrolled test subjects are stored, then unenrolled subject identification transactions may be generated offline using a jack-knife approach. The entire crew is enrolled with a single test subject omitted. The system then tries to identify the omitted test subject against the remainder of the test crew, and the process is repeated for each test subject in turn. The considerations of 7.6.3.1.3 on the adequacy of the stored data for impostor comparisons also apply in this case.

8 Analyses

8.1 General

8.1.1 If the test crew is representative of the target population, and each test subject has one enrolment template and makes the same number (and pattern) of transactions, the observed error rate proportions will be the best estimates of the true error rates.

8.1.2 When the test crew is not representative of the target population (for example over-representation of known problem cases), or test transactions of individual test subjects are un-representative of those of the test crew as a whole (for example test subjects making more or fewer transactions than average), weighting may be appropriate to redress the imbalance. If error rates are estimated using a weighted proportion, the method of weighting shall be reported. When weighting by class of user is used, the observed class error rates should also be reported.

EXAMPLE If test subjects make differing numbers of verification or identification attempts, errors for each test subject might be weighted in inverse proportion to the number of attempts the test subject makes, as a simple proportion could bias the estimated error rates towards those of the heavy users of the system or toward those requiring multiple attempts for acceptance.

8.1.3 It can be useful to measure error rates on a per person basis, per person-class (e.g. separate error rates for males and females), or per type of biometric instance (e.g. separate error rates for each finger position).

a) Such individual metrics may be of intrinsic interest, indicating the type of person for whom better or worse performance might be achieved.

b) Per person or per person-class metrics are needed when the best estimate is a weighted proportion.

c) The spread of individual error-rates can help in estimating the uncertainty of performance estimates.

8.1.4 If errors in enrolment, sample acquisition, and verification or identification are classified by cause, or by the step in the enrolment, acquisition or matching process, then it may be possible to determine separate error rates for the different causes, or for the different components of the process.

8.2 Fundamental performance metrics

8.2.1 Failure-to-enrol rate

8.2.1.1 The failure-to-enrol rate is the proportion of the population for whom the system fails to complete the enrolment process. The failure-to-enrol rate shall include:

— those unable to present the required biometric characteristic;

— those unable to produce a sample of sufficient quality at enrolment; and

— those who cannot reliably produce a match decision with their newly created template during attempts to confirm the enrolment is usable.

NOTE 1 It is also possible to determine a failure-to-enrol rate for different biometric instances, such as different fingers, for example to report different failure-to-enrol rates for thumbs, index fingers, etc.

NOTE 2 In technology evaluations, analysis is based on a previously collected corpus and there will be no problem in obtaining a sample image. Even so, there may be enrolment failures — for example, when the image sample is of too low a quality for features to be extracted.

8.2.1.2 The failure-to-enrol rate for the target population shall be estimated as the proportion (or weighted proportion) of the test crew who could not be enrolled under the predetermined enrolment policy.

8.2.1.3 The failure-to-enrol rate depends on the enrolment policy that governs the sample quality threshold for enrolment, the decision threshold to confirm the enrolment is usable, and the number of attempts or time allowed for enrolment in an enrolment transaction. The enrolment policy shall be described along with the observed failure-to-enrol rate.

NOTE Setting stricter quality requirements at enrolment will increase the failure-to-enrol rate but improve matching performance.

8.2.1.4 Attempts by users unable to enrol in the system shall not contribute to the failure-to-acquire rate, or matching error rates.

8.2.2 Failure-to-acquire rate

8.2.2.1 The failure-to-acquire rate is the proportion of verification or identification attempts for which the system fails to capture or locate a sample of sufficient quality. The failure-to-acquire rate shall include:

— attempts where the biometric characteristic cannot be presented (e.g. due to temporary illness or injury) or captured;

— attempts for which the segmentation or feature extraction fail; and

— attempts in which the extracted features do not meet the quality control thresholds.

NOTE 1 It is also possible to determine a failure-to-acquire rate for transactions, e.g., measuring the proportion of transactions for which no attempts provided a sample of sufficient quality for matching.

NOTE 2 In technology evaluations, analysis is based on a previously collected corpus and there will be no sample capture failures. A failure-to-acquire rate for the corpus may be known. Additional acquisition problems, for example when the sample is of too low a quality for feature extraction would add to the failure-to-acquire rate.

8.2.2.2 The failure-to-acquire rate shall be estimated as the proportion (or weighted proportion) of recorded genuine attempts (and possibly any online zero-effort impostor attempts) that could not be completed due to failures at presentation (no image captured), segmentation, feature extraction, or quality control.

8.2.2.3 The failure-to-acquire rate will depend on thresholds for sample quality, as well as the allowed duration for sample acquisition or allowed number of presentations. These settings shall be reported along with the observed failure-to-acquire rate.

NOTE Setting stricter quality thresholds for sample acquisition will increase the failure-to-acquire rate, but improve matching performance.

8.2.2.4 Attempts where the raw sample was not acquired or did not meet quality thresholds are not processed by the matching algorithm, and do not generate matching scores. Such failures-to-acquire shall be excluded in calculating the false match and false non-match error rates, but shall be included in calculating the false accept and false reject rates. The failure-to-acquire, false match and false non-match rates shall be calculated at the same quality acceptance threshold settings.

8.2.3 False non-match rate

8.2.3.1 The false non-match rate is the proportion of samples, acquired from genuine attempts, that are falsely declared not to match the template of the same characteristic from the same user supplying the sample.

8.2.3.2 The false non-match rate shall be estimated as the proportion (or weighted proportion) of recorded genuine attempts that were passed to the matching subsystem, for which the similarity score produced was below the matching decision threshold.

8.2.3.3 The false non-match rate depends on the matching decision threshold, and shall be quoted along with the observed false match rate at the same threshold (or plotted against the false match rate at the same threshold in an ROC or DET curve).

8.2.3.4 In evaluations where test subjects have made multiple attempts, it can be useful to show how the false non-match rate varies over the test crew. This may be done by calculating an error rate for each test subject's attempts, and plotting a histogram showing the error rate for each test subject, ordering test subjects in increasing order of their error rates.

8.2.4 False match rate

8.2.4.1 The false match rate is the proportion of samples, acquired from zero-effort impostor attempts, that are falsely declared to match the compared non-self template.

NOTE In a zero-effort impostor attempt, the individual submits their own biometric characteristic as if they were attempting successful verification against their own template. In the case of dynamic signature verification, for example, an impostor would sign their own signature in a zero-effort attempt. In such cases, where impostors may easily imitate aspects of the required biometric, a second impostor measure based on active impostor attempts may be needed. However, defining the methods or level of skill to be used in active impostor attempts is outside the scope of this part of ISO/IEC 19795.

8.2.4.2 The false match rate shall be estimated as the proportion (or weighted proportion) of recorded zero-effort impostor attempts that were passed to the matching subsystem, for which the similarity score produced was greater than or equal to the matching decision threshold.

8.2.4.3 The false match rate depends on the matching decision threshold, and should be quoted along with the observed false non-match rate at the same threshold (or plotted against the false non-match rate at the same threshold in an ROC or DET curve).

8.2.4.4 If a test subject is enrolled, and their template affects the templates of others in the system, or if the matching algorithm modifies itself using this (and other) templates, then impostor attempts using that subject will be biased, and should not be used to estimate the false match rate. Clauses 7.6.2.6 and 7.6.3.2 detail how to deal with such cases.

EXAMPLE Eigenface systems, using all enrolled images for creation of the basis-images, and cohort-based speaker recognition systems are two examples for which templates are dependent.

8.2.4.5 Comparison of genetically identical biometric characteristics (for instance, between a person's index and middle fingers, or across identical twins) yields different score distributions than comparison of genetically different characteristics [25-27]. Consequently, such genetically similar comparisons shall not be considered in deriving the false match rate.

8.2.4.6 In evaluations where there are several impostor transactions per subject, or per template, it can be useful to show how the false match rate varies over test subjects, and over stored templates. This involves calculating the individual false match error rate for impostor attempts by each subject, and for impostor attempts against each template. Histograms may be plotted to show the error rate for each test subject, ordering test subjects in increasing order of their error rates.

EXAMPLE A face recognition system could admit a set of "golden faces" where false matches occur mainly with this set of faces. Histograms showing variation of error rates across subjects could reveal this vulnerability.

8.3 Verification system performance metrics

8.3.1 General

A first order estimation of the false accept and false reject rates for transactions of multiple attempts can be derived from the detection error trade-off curve. However, such estimates cannot take account of correlations in sequential attempts and in the comparisons involving the same user, and consequently can be quite inaccurate. Therefore these performance metrics shall be derived directly, using test transactions with multiple attempts as specified by the decision policy.

8.3.2 False reject rate

8.3.2.1 The false reject rate is the proportion of genuine verification transactions that will be incorrectly denied. A transaction may consist of one or more genuine attempts depending on the decision policy.

8.3.2.2 The false reject rate shall be estimated as the proportion (or weighted proportion) of recorded genuine transactions that were incorrectly denied. This includes transactions denied due to failures-to-acquire as well as those denied due to matching errors.

EXAMPLE If a verification transaction consists of a single attempt, then a failure–to-acquire, or a false non-match will cause a false rejection, and the false reject rate would be given by:

```
FRR = FTA + FNMR * (1 – FTA)
```

where:

FRR	is the false reject rate;
FTA	is the failure-to-acquire rate;
FNMR	is the false non-match rate.

8.3.2.3 The false reject rate will depend on the decision policy, the matching decision threshold, and any threshold for sample quality. The false reject rate shall be reported with these details, alongside the estimated false accept rate at the same values, (or plotted against the false accept rate at the same threshold(s) in an ROC or DET curve).

8.3.3 False accept rate

8.3.3.1 The false accept rate is the expected proportion of zero-effort non-genuine transactions that will be incorrectly accepted. A transaction may consist of one or more non-genuine attempts depending on the decision policy.

8.3.3.2 The false accept rate shall be estimated as the proportion (or weighted proportion) of recorded zero-effort impostor transactions that were incorrectly accepted.

EXAMPLE If a verification transaction consists of a single attempt, then a false acceptance requires that the submitted sample is not rejected by the quality control (i.e. no failure-to-acquire) and that there is a false match. The false accept rate would be given by:

```
FAR = FMR * (1 – FTA)
```

where:

FAR	is the false accept rate;
FMR	is the false match rate;
FTA	is the failure-to-acquire rate.

8.3.3.3 The false accept rate will depend on the decision policy, the matching decision threshold, and any threshold for sample quality. The false accept rate shall be reported with these details, alongside the estimated false reject rate at the same values, (or plotted against the false reject rate at the same threshold(s) in an ROC or DET curve).

8.3.4 Generalized false reject rate and generalized false accept rate

Comparison of systems having different failure-to-enrol rates may require use of generalized false reject and false accept rates which combine enrolment, sample acquisition and matching errors. The method of generalization should be appropriate to the evaluation. A typical generalization is to treat a failure-to-enrol as if the enrolment completed, but all subsequent verification or identification transactions by that enrolee, or against their template, fail. The method of generalization shall be reported.

EXAMPLE 1 We assume a scenario evaluation where unenrolled subjects take no further part in the evaluation, and that a verification transaction consists of a single attempt. In this case a generalized false acceptance occurs if (i) both the subject making the impostor attempt, and the impersonated subject are enrolled, and (ii) the submitted sample is not rejected by the quality control (i.e. no failure-to-acquire) and (iii) that there is a false match. A generalized false rejection occurs if (i) the subject is not enrolled, or (ii) the submitted sample cannot be acquired, or (iii) there is a false non-match. The generalized false accept and false reject rates would be given by:

```
GFAR = FMR * (1 - FTA) * (1 - FTE)^2
GFRR = FTE + (1 - FTE) * FTA + (1 - FTE) * (1 - FTA) * FNMR
```

where

GFAR	is the generalized false accept rate;
GFRR	is the generalized false reject rate;
FMR	is the false match rate;
FNMR	is the false non-match rate;
FTE	is the failure-to-enrol rate;
FTA	is the failure-to-acquire rate.

EXAMPLE 2 In a technology evaluation, enrolment templates are generated from all gallery images that do not cause a failure-to-enrol, and attempt features are generated from all probe images that do not cause a failure-to-acquire. In this case the generalized false accept and false reject rates would be given by:

```
GFAR = FMR * (1 - FTA) * (1 - FTE)
GFRR = FTE + (1 - FTE) * FTA + (1 - FTE) * (1 - FTA) * FNMR
```

8.4 (Open-set) Identification system performance metrics

8.4.1 General

A first order estimation of the (true positive) identification rate for closed-set systems, and the false positive and false negative identification rates for open-set systems, can be derived from the matching error DET curve. However, such estimates cannot take account of correlations in the comparisons involving the same user, and consequently can be quite inaccurate. Therefore, at least for small database sizes, identification transactions should be collected to derive these performance metrics directly. Estimation of the performance of large-scale identification systems (beyond the size of the test) may need to be extrapolated using both the first-order estimation and the identification performance on the smaller database. The model used for extrapolating performance should be reported in such cases.

EXAMPLE The performance of an identification using a single biometric sample against a database of size N might be approximated using the following formulae, providing these formulae have been validated by the identification error rates observed on the test data.

```
FNIR = FTA + (1 - FTA) * FNMR
FPIR = (1 - FTA) * (1 - (1 - FMR)^N)
```

where:

FPIR	is the false-positive identification-error rate;
FNIR	is the false-negative identification-error rate;
FTA	is the failure-to-acquire rate;
FMR	is the false match rate;
FNMR	is the false non-match rate;
N	is the number of templates in the database.

NOTE In the case of identification systems using pre-selection, the above model of performance can be extended using the performance metrics for the pre-selection algorithm (see Annex D).

8.4.2 Identification rate

The (true-positive) identification rate at rank r is the proportion of identification transactions by a user enrolled in the system, for which user's true identifier is included in the candidate list returned. When a single point identification rank is reported, it should be referenced directly to the database size.

EXAMPLE "The identification rate at rank 1 was 95% against a database of 250 entries".

8.4.3 False-negative and false-positive identification-error rates

8.4.3.1 The false-negative identification-error rate is the proportion of identification transactions by users enrolled in the system, for which the user's correct identifier is not included in the candidate list returned.

8.4.3.2 The false-positive identification-error rate is the proportion of identification transactions by users not enrolled in the system, for which a non-empty list of candidate identifiers is returned.

NOTE The false-positive identification-error rate increases with the number of people enrolled in the system.

8.4.3.3 Open-set identification performance can also be plotted as an ROC (plotting the true-positive identification rate against the false-positive identification-error rate) or as a DET (plotting the false-negative identification-error rate against the false-positive identification-error rate) for a fixed database size, and fixed number of identities returned.

NOTE For database size 1, these curves show (one-to-one) verification performance.

8.4.3.4 The overall identification performance of an open-set system, as the enrolment database grows may be shown as a plot of identification rate (at rank 1) against size of enrolment database, for a constant value of the false-positive identification-error rate (requiring the threshold to be adjusted as the database grows). Alternatively a set of DET curves showing the relationship between false-positive and false-negative identification-error rates may be plotted for various database sizes, as in the example shown in Annex E, Figure E.1.

8.5 Closed-set identification

8.5.1 The identification rate at rank r is the probability that a transaction by a user enrolled in the system includes that user's true identifier within the top r matches returned. When a single point identification rank is reported, it should be referenced directly to the database size.

EXAMPLE "The identification rate at rank 1 was 95 % against a database of 250 entries".

8.5.2 The primary measure of closed-set identification performance is normally shown as a cumulative match characteristic curve in which the (true-positive) identification rate at rank *r* is plotted as a function of *r*.

NOTE A suggested algorithm for efficiently generating this data points on the CMC curve is provided in Annex F.

8.5.3 One drawback of the CMC is its dependence on the number of people enrolled in the system. For this reason, a graph plotting the identification rate at rank 1 as a function of the number of enrolments should be included with the results.

8.6 Detection error trade-off / Receiver operating characteristic curves

8.6.1 The detection error trade-off (DET) measures shall be developed using the genuine and impostor matching scores from comparisons between single attempt features and single enrolment templates. Each attempt will result in a recorded matching score. Scores developed for genuine attempts will be ordered. Impostor scores will be handled similarly. Outliers should be investigated to determine if labelling errors are indicated. Removal of any scores from the test should be fully documented and will lead to external criticism of the test results.

NOTE Histograms for both genuine and impostor scores can be instructive, but will not be used in the development of the DET curve. Consequently, we make no recommendations regarding the creation of the histograms from the transaction data, although this is a very important area of continuing research interest. The resulting histograms will be taken directly as the best estimates for the genuine and impostor distributions. Under no circumstances should models be substituted for either histogram as an estimate of the underlying distribution.

8.6.2 DET (or ROC) curves are established through the accumulation of the ordered genuine and impostor scores. As the score varies over all possible values, the DET (or ROC) curve is plotted parametrically, each point (x, y) representing the false match and false non-match rates using that score as the decision threshold. The false match rate is the proportion of impostor similarity scores at or above the current value of the score parameter, and the false non-match rate is the proportion of genuine similarity scores below the score parameter. The curves should be plotted with false match rate on the abscissa (x-axis) and false non-match rate on the ordinate (y-axis). Axes depicting error rates may use logarithmic scales.

NOTE A suggested procedure for efficiently deriving the data points on the DET / ROC is provided in Annex F.

8.6.3 DET (or ROC) curves can also be used to plot the relationship between the false accept rate and false reject rate in a similar manner. The false accept rate and false reject rate will depend on the false match rate, false non-match rate, and failure-to-acquire rate in a manner that will depend on the decision policy. Transactions of multiple attempts may require generation of a new transaction score based on the similarity scores of the constituent attempts (e.g., the maximum value of the similarity scores for a best of three attempts decision policy). Similarly DET (or ROC) curves may be used to show the relationship between identification error rates.

8.7 Uncertainty of estimates

8.7.1 Performance estimates will be affected by both systematic errors and random errors. Random errors include those due to the natural variation in test subjects and sample presentation. Systematic errors include those due to bias in the test procedures, for example if certain types of individual are under-represented in the test crew. Neither type of error is perfectly quantifiable, and therefore there will be an uncertainty in the results of a performance evaluation. Nevertheless, the uncertainty in the measured performance shall be estimated. Annex B provides some methods by which uncertainty in performance results may be estimated.

8.7.2 Uncertainties arising from random effects become smaller as the size of the test increases, and can often be estimated from the collected data. It may also be possible to determine the effects of some of the systematic errors. For example, checking whether the error rates for an under-represented category of individuals are consistent with the overall error rates could show whether a properly balanced test crew would give different error rates. Part of the performance trial may be repeated in different environmental conditions to check that the measured error rates are not unduly sensitive to small environmental changes.

9 Record keeping

9.1 In order to assess whether an evaluation has been conducted in accordance with this part of ISO/IEC 19795, record keeping shall be in accordance with the requirements of ISO/IEC 17025. Records shall include:

a) the original sample images (if collected), unless the volume of data renders this impractical;

b) if sample images are not collected, templates for each enrolment, and the feature data for each verification or identification attempt should be stored (if these are available);

c) matching scores and decisions output by the biometric system, if available;

d) the methods used to derive performance measures and uncertainties;

e) the identities of staff responsible for conducting enrolment and supervising the collection of transaction data; and

f) sufficient information to establish an audit trail;

9.2 Sufficient information shall be kept to:

a) enable the evaluation to be repeated under conditions as close as possible to the original, and

b) facilitate, if possible, identification of factors affecting the uncertainty of results.

9.3 Records (whether written or electronic) shall be protected to avoid loss or change of the original test data. If alterations must be made, a copy of the original shall be kept with a note of the alterations.

9.4 Where mistakes occur (in the data collection procedures etc.) records should show both the original erroneous data, and the corrected values.

10 Reporting performance results

10.1 Fundamental metrics

The following fundamental performance metrics are applicable to all biometric systems, and should be reported if available:

a) failure-to-enrol rate;

b) failure-to-acquire rate;

c) false match rate and corresponding false non-match rate (preferably over the range of threshold values); and

d) where appropriate, histograms showing how individual error rates vary between subjects.

10.2 Verification system metrics

Verification system performance shall be reported using the following metrics:

a) failure-to-enrol rate, if available, otherwise a statement shall be provided that the failure-to-enrol rate is unknown;

b) failure-to-acquire rate, if available, otherwise a statement shall be provided that that the failure-to-acquire rate is unknown;

c) false accept rate and corresponding false reject rate (preferably over a range of threshold values);

d) where appropriate, generalised false accept rate and corresponding generalised false reject rate (preferably over a range of threshold values) together with details of the method of generalisation; and

e) where appropriate, histograms showing how individual error rates vary between subjects.

10.3 Identification system metrics

Open-set identification system performance shall be reported using the following metrics:

a) failure-to-enrol rate, if available, otherwise a statement shall be provided that the failure-to-enrol rate is unknown;

b) failure-to-acquire rate, if available; otherwise a statement shall be provided that the failure-to-acquire rate is unknown;

c) false-positive identification-error rate, and corresponding false-negative identification-error rate (preferably over the range of threshold values);

d) size of database;

e) several DET or ROC curves may be shown corresponding to different sizes of the template database, different numbers of identifiers returned etc.; and

f) where appropriate, histograms showing how individual error rates vary between subjects.

10.4 Closed-set identification system metrics

Closed-set identification system performance shall be reported using the following metrics:

a) cumulative match characteristic (CMC) curve;

b) size of database.

10.5 Reporting test details

Performance statements such as the DET curve, failure-to-enrol and failure-to-acquire rates, and binning penetration and error rates, are dependent on test type, application and population. For these measures to be interpreted correctly, the following additional information should be provided:

a) details of the system(s) tested. This should include more than just the biometric component, as factors such as the user interface will influence performance too;

b) the type of evaluation:

- technology evaluation: details of the corpus used;
- scenario evaluation: details of the test scenario;
- operational evaluation: details of the operational application;

c) size of evaluation:

- number of test subjects;
- number of fingers, hands or eyes etc enrolled by each test subject;
- number of visits made by test subject;
- number of transactions per test subject (or test subject finger, etc) at each visit;

d) demographics of the test crew (age, gender, etc.);

e) details of the test environment;

f) time separation between enrolments and test transactions;

g) quality and decision thresholds used during data collection;

h) details of how factors potentially affecting performance were controlled (see Annex C);

i) details of the test procedure, e.g. policies for determining enrolment failures;

j) details of the level of training, familiarization, and habituation of the test crew in the use of the system;

k) details of abnormal cases and data excluded from analysis;

l) estimated uncertainties (and method of estimation);

m) deviations from the guidelines of this part of ISO/IEC 19795 should also be explained. Sometimes it will be necessary to compromise one aspect in order to achieve another – for example, randomising the order of using fingers on a fingerprint device might lead to user confusion and a higher number of labelling errors.

10.6 Graphical presentation of results

10.6.1 General

10.6.1.1 The matching and/or decision performance of a biometric system over a range of decision thresholds should be graphically represented using either ROC or DET curves, but not both.

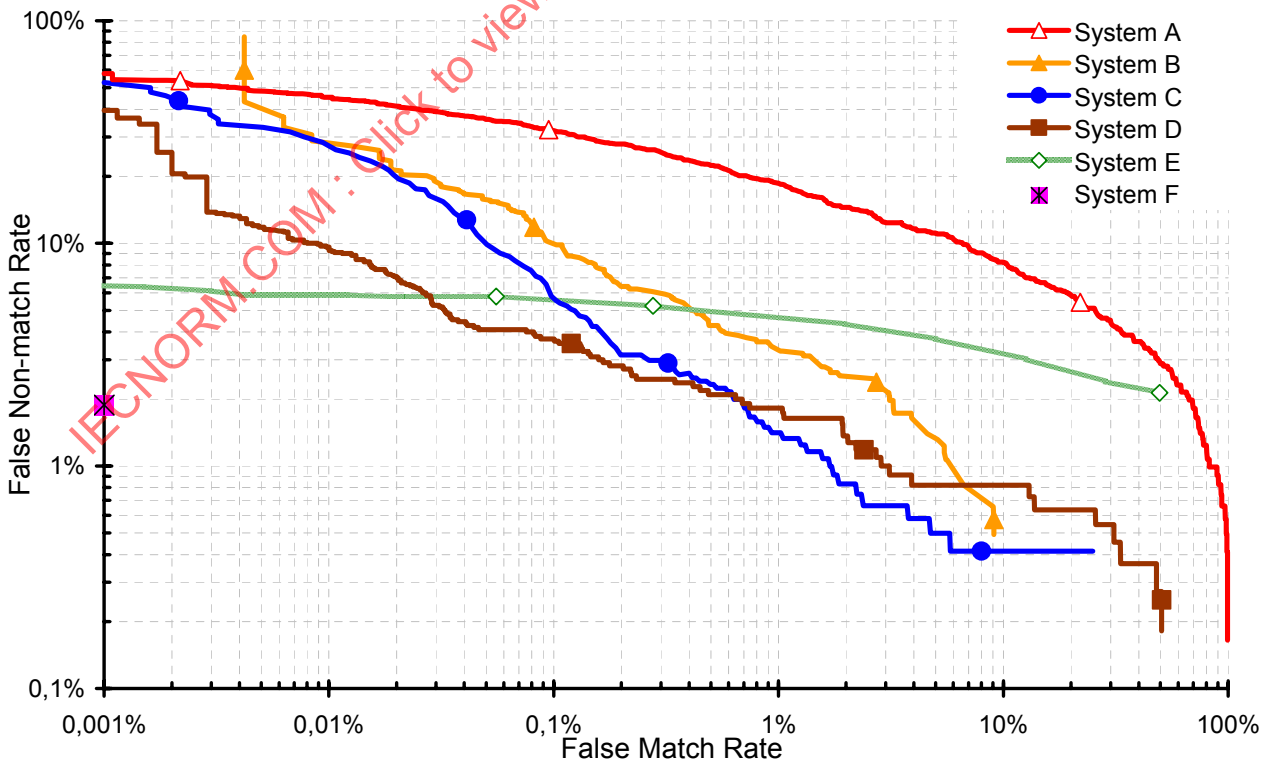
10.6.1.2 Axis scales (minimum and maximum values shown, and the use of logarithmic scales) should be selected for clarity of the presented results, and should be consistent between different graphs in the same report. If necessary to change scaling to maintain clarity, there should be a note to the figure remarking on the change of scales.

10.6.1.3 For comparing the performance of different systems, the decision error DET or ROC (false reject rate versus false accept rate), which shows the combined effect of matching errors, image acquisition errors, binning errors, and enrolment errors, will be more helpful than graphs showing the fundamental error rates.

10.6.2 DET curve

10.6.2.1 DET curves may be used to plot matching error rates (false non-match rate against false match rate), decision error rates (false reject rate against false accept rate), and open-set identification error rates (false negative identification rate against false positive identification rate).

10.6.2.2 Logarithmic axes may be used, to help spread out the plot for greater clarity. In the case of logarithmic plots, an observed error rate of zero errors in N trials may be plotted as a value 0,5/N, or as the scale minimum if greater.



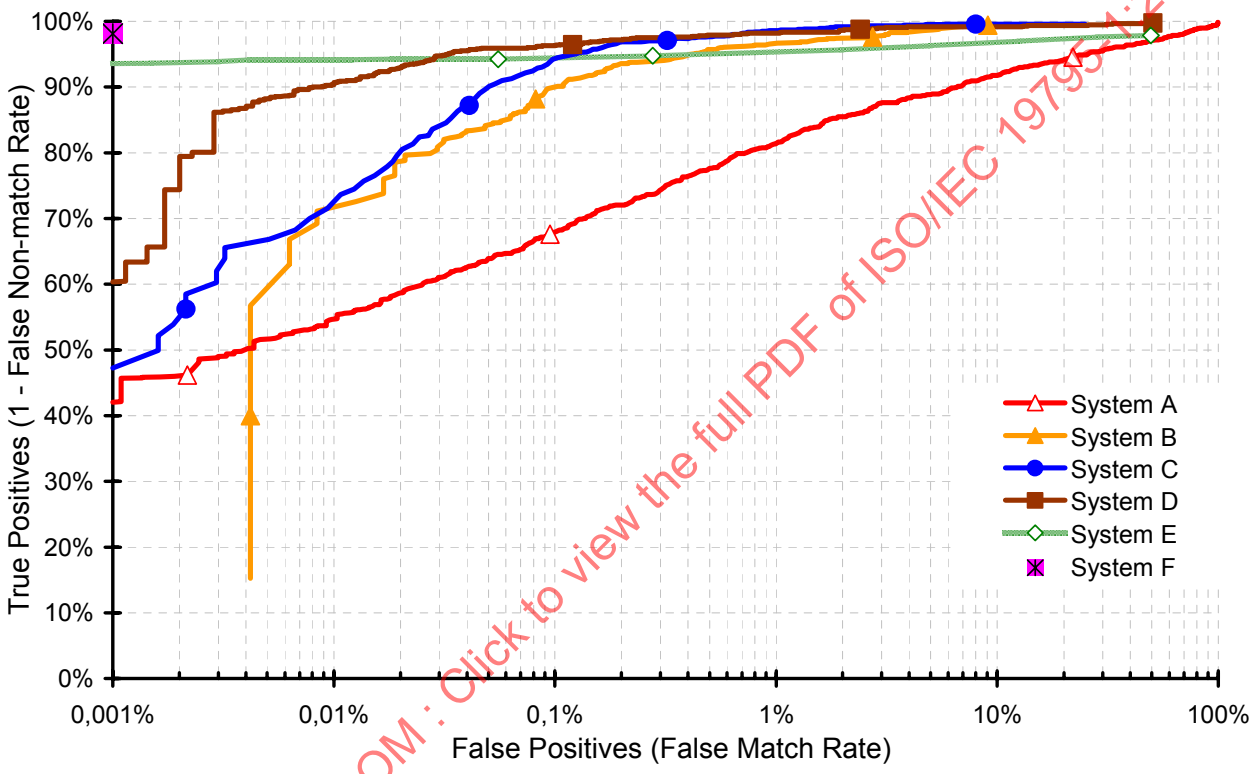
NOTE These DET curves plot the same data as the ROC curves in Figure 4.

Figure 3 — Example set of DET curves

10.6.3 ROC curves

10.6.3.1 ROC curves are a traditional method for summarising the performance of imperfect diagnostic, detection and pattern-matching systems. ROC curves are threshold-independent, allowing performance comparison of different systems under similar conditions, or of a single system under differing conditions. ROC curves may be used to plot matching algorithm performance (1-false non-match rate against false match rate), end-to-end verification system performance (1-false reject rate against false accept rate), as well as open-set identification system performance ((correct) identification rate against false-positive identification-error rate).

10.6.3.2 The x-axis may be plotted using a logarithmic scale, to help spread out the plot for greater clarity. In the case of logarithmic plots, an observed error rate of zero errors in N trials may be plotted as a value 0,5/N, or as the scale minimum if greater.

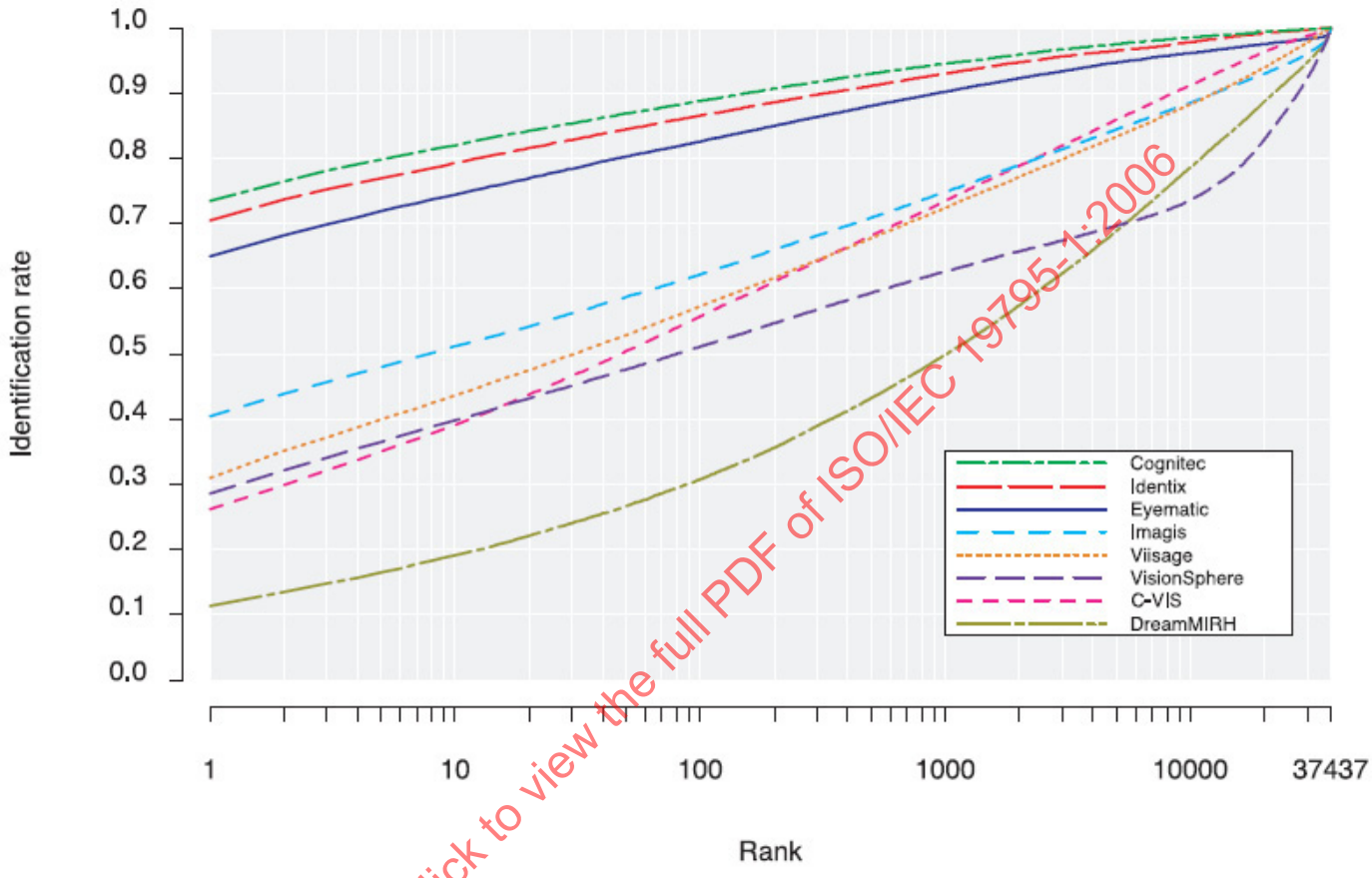


NOTE These ROC curves plot the same data as the DET curves in Figure 3.

Figure 4 — Example set of ROC curves

10.6.4 CMC Curves

For closed-set identification applications, performance results are often illustrated using a cumulative match characteristic curve. The curve plots, as a function of *k* the proportion of transactions where a test subject's identifier is included among the top *k* identifiers returned.



NOTE This example is taken from FRVT2002 [11, Figure10]. The graph shows the identification rate for a database of 37437 face templates.)

Figure 5 — Example CMC curves

Annex A
(informative)

Differences between evaluation types

Table A.1

	Technology	Scenario	Operational
What is tested	Biometric component (matching or extraction algorithm)	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 behaviour	Dependent on available controls and instrumentation to establish whether data is from genuine users or impostors
User behaviour controlled by test administrator	Not applicable during testing; may be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled (unless user behaviour is an independent variable)	Uncontrolled
User has real-time feedback of the result of attempt	No	Yes	Yes
Repeatability of results	Repeatable (corpus fixed)	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; may be recorded when biometric data recorded	Recorded	Recorded during enrolment; may be recorded during verification/identification
Typical results reported	Comparison of biometric components, 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 test crew	Predicted end-to-end throughput, FMR, FNMR, FTA, FTE, FAR, FRR	End-to-end throughput; reliable testing of operational FAR and FRR requires some knowledge of ground truth
Constraints	Appropriate test corpus, 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

NOTE Although in some cases there may be exceptions to the entries in this table, these are the mainstream, fundamental characteristics and distinctions.

Annex B
(informative)

Test size and uncertainty

B.1 Confidence intervals and test size assuming independent identically distributed comparisons

B.1.1 Rule of 3

The Rule of 3 [22, 28-30] addresses the question "What is the lowest error rate that can be statistically established with a given number N of independent identically distributed comparisons?" This value is the error rate p for which the probability of zero errors in N trials, purely by chance, is (for example) 5%. This gives:

$$p \approx 3/N$$

for a 95% confidence level.

EXAMPLE A test of 300 independent samples returning no errors can be said with 95% confidence to have an error rate of 1% or less.

NOTE 1 $p \approx 2/N$ for a 90% confidence level.

NOTE 2 The assumption of independent identically distributed (i.i.d.) attempts may be achieved if each genuine attempt uses a different user, and if no two impostor attempts involve the same user. With n test subjects, there would be *n* genuine attempts and *n/2* impostor attempts. However, cross-comparisons between all submitted sample features and enrolled templates generates many more impostor attempts and, according to [24], achieves smaller uncertainty despite dependencies between the attempts. Thus, except perhaps in the case of operational testing, there is little merit in restricting data to a single attempt per user to achieve the i.i.d. assumption.

B.1.2 Rule of 30

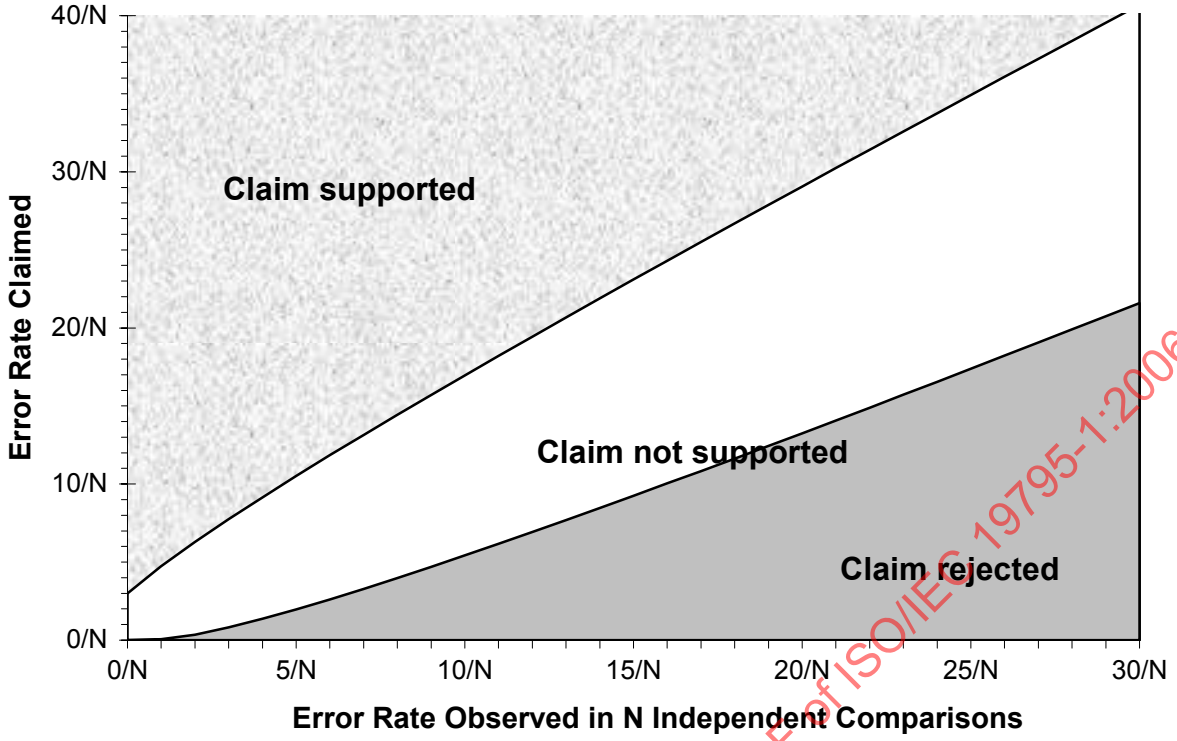
The Rule of 30 states that to be 90% confident that the true error rate is within ± 30% of the observed error rate, there should be at least 30 errors [13]. So, for example, if there are 30 false non-match errors in 3000 independent genuine trials, we can say with 90 % confidence that the true error rate is between 0,7 % and 1,3%. The rule comes directly from the binomial distribution assuming independent trials, and may be applied by considering the performance expectations for the evaluation.

EXAMPLE Suppose the performance goals are a 1 % false non-match rate, and a 0,1 % false match rate. This rule implies 3000 genuine attempt trials and 30 000 impostor attempt trials. Note however, the key assumption that these trials are independent. This would require 3000 enrolees and 30 000 impostors. The alternative is to compromise on independence by re-using a smaller set of test subjects, and to be prepared for a loss of statistical significance.

NOTE The rule generalizes to different proportional error bands. For example, to be 90 % confident that the true error rate is within ±10% of the observed value, at least 260 errors are needed. To be 90 % confident that the true error rate is within ±50% of the observed value, at least 11 errors are needed.

B.1.3 Number of comparisons to support a claimed error rate

B.1.3.1 The number of statistically independent comparisons required to support a claimed error rate is illustrated in Figure B.1. For example, no false matches in N independent impostor comparisons would support a claimed false match rate of 3/N, with 95 % confidence, while 30 errors would support a claim of 41/N.



NOTE This chart provides a reasonable approximation when the claimed error rate is 1% or below.

Figure B.1 – 95% confidence decision regions for accepting (or rejecting) an error rate claim with N independent comparisons

B.1.3.2 To ensure statistical independence, the impostor and impersonated templates in all comparisons would need to be different, and selected randomly and uniformly from the target population. This approach is unlikely to be efficient for low false match rates, as N independent comparisons will require 2N volunteers.

B.1.3.3 An alternative cross-comparison approach will often be adopted, though this does not ensure statistical independence. With P people, cross-comparison of attempts/templates for each (unordered) pair, may exhibit a low degree of correlation. The correlations within these P(P – 1)/2 false match attempts will reduce the confidence level for supporting an FMR claim compared with the same number of completely independent comparisons.

B.2 Variance of performance measures as a function of test size

As the test size increases, the variance of estimates will decrease, but the scaling factor depends on the source of variability.

a) If test subjects each make multiple genuine attempts, then the variance of the observed false non-match rate has components due to:

   — variability of test subjects, scaling as 1 / (number of test subjects); and

   — residual variability of genuine attempts, scaling as 1 / (number of genuine attempts).

b) If test subjects make multiple attempts, and impostor attempts are generated offline by cross-comparison of these genuine attempts against enrolment templates from a different set of users, then the variance of the observed false match rate has components due to:

— variability of test subjects, scaling as 1 / (number of impostor test subjects);

— variability of impersonated templates, scaling as 1 / (number of impersonated templates);

— variability of genuine samples (other than that accounted for by variability of test subjects), scaling as 1 / (number of genuine attempts); and

— residual variability of the generated impostor attempts, scaling as 1 / (number of impostor attempts).

NOTE Doddington et al [16] show that biometric systems can have “goats”, “lambs” and “wolves”. Goats have a personal false non-match rate significantly higher than that for the overall population, lambs are those whose templates incur a disproportionate share of false matches, while wolves are those whose samples are particularly successful at giving false matches. This would imply that, for the false non-match rate, the component of variance for test subjects is non-zero; and for the false match rate, the components for test subjects and for templates are non-zero.

B.3 Estimates for variance of performance measures

B.3.1 General

This Clause presents formulae and methods for estimating the variance of performance measures. The variance is a statistical measure of uncertainty and can be used in estimating confidence intervals, etc. The applicability of these formulae depends on the following assumptions about the distribution of matching errors:

— the test crew is representative of the target population. This will be the case if, for example, the test subjects are drawn at random from the target population;

— attempts by different subjects are independent. This will not always be true. Users’ behaviour will be influenced by what they see others do. However, the correlations between test subjects are likely to be minor in comparison to the correlations within a set of attempts by one test subject;

— attempts are independent of threshold. Otherwise, the estimates for the error rates may be biased except at the threshold used for data collection;

— error rates vary across the population. Different subjects may have different individual false non-match rates, and different subject pairs may have different individual false match rates;

— the number of observed errors is not too small. In cases with no observed errors, the formulae would give a zero variance, but the Rule of 3 would apply.

B.3.2 Variance of observed false non-match rate

B.3.2.1 False non-match rate - Single attempt per test subject

In the case where each test subject makes a single attempt:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{B.1}$$

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} \tag{B.2}$$

where

n is the number of enrolled test subjects;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{V}(\hat{p})$ is the estimated variance of observed false non-match rate.

NOTE 1 A derivation of this estimate may be found in many statistical textbooks (e.g.,[31]).

NOTE 2 These formulae have sometimes been misapplied to cases where subjects make several attempts. The replacement of the number of test subjects n by the number of attempts is generally not valid.

NOTE 3 These formulae will also be appropriate for estimating variances of failure-to-acquire and failure-to-enrol rates when there is one attempt per test subject.

B.3.2.2 False non-match rate - Multiple attempts per test subject

In the case where each test subject makes the same number of attempts, the appropriate estimates are given by the following formulae [31]:

$$\hat{p} = \frac{1}{mn}\sum_{i=1}^{n} a_i \tag{B.3}$$

$$\hat{V}(\hat{p}) = \frac{1}{(n-1)}\left(\frac{1}{m^2 n}\sum_{i=1}^{n} a_i^2 - \hat{p}^2\right) \tag{B.4}$$

where

n is the number of enrolled test subjects;

m is the number attempts made by each test subject;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{V}(\hat{p})$ is the estimated variance of observed false non-match rate.

NOTE 1 When $m = 1$, the estimates are the same as those in formulae (B.1) and (B.2).

NOTE 2 These formulae will also be appropriate for estimating variances of failure-to-acquire rates when there are multiple attempts per test subject.

B.3.2.3 False non-match rate - Unequal numbers of attempts per test subject

There will be occasions when the number of attempts per subject varies. Some subjects might not complete the desired number of attempts. Acquisition failures may also cause attempts to be missing from the false non-match rate calculations. Provided there is no correlation between the number of attempts made and the differing success rates of individuals, the appropriate formulae are as follows.

$$\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \tag{B.5}$$

$$\hat{V}(\hat{p}) = \frac{\sum_{i=1}^{n} a_i^{\,2} - 2\hat{p}\sum_{i=1}^{n} a_i m_i + \hat{p}^2 \sum_{i=1}^{n} m_i^{\,2}}{\frac{n-1}{n}\left(\sum_{i=1}^{n} m_i\right)^2} \tag{B.6}$$

where

n is the number of enrolled test subjects;

m_i is the number attempts made by the i^{th} test subject;

a_i is the number of false non-matches for the i^{th} test subject;

$\hat{p}$ is the observed false non-match rate;

$\hat{V}(\hat{p})$ is the estimated variance of observed false non-match rate.

NOTE 1 This formula for the variance (from [31]) is an approximation to give an expression in a usable form.

NOTE 2 When all m_i are equal, we obtain the same estimates as in formulae (B.3) and (B.4).

NOTE 3 Sometimes the different frequency of use by test subjects will be correlated with the differing success rates: for example, test subjects that are rejected may make additional attempts to be recognized; or those using the system more frequently may get better performance through the effects of habituation. In such cases the formulae (B.5) and (B.6) cannot be directly applied, as results may then be dominated by a small group of excessively frequent but unrepresentative users.

B.3.3 Variance of observed false match rate

In the case where a full set of cross-comparisons is made, the observed false match rate, and an estimate of the variance are given by:

$$\hat{q} = \frac{1}{mn(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \tag{B.7}$$

$$\begin{aligned} \hat{V}(\hat{q}) &= \frac{1}{m^2 n(n-1)(n-2)(n-3)} \left\{ \sum_{i=1}^{n} (c_i + d_i)^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} (b_{ij}^{\,2} + b_{ij} b_{ji}) \right\} - \frac{(4n-6)}{(n-2)(n-3)} \hat{q}^2 \\ &\approx \frac{1}{m^2 n^2 (n-1)^2} \sum_{i=1}^{n} (c_i + d_i)^2 - \frac{4}{n} \hat{q}^2 \end{aligned} \tag{B.8}$$

where

n is the number of test subjects (and of enrolment templates);

m is the number of samples per test subject;

b_{ij} is the number of samples from the i^{th} test subject falsely matching the template of the j^{th} test subject (and $b_{ii} = 0$);

c_i is the number of false matches in total against the template of the i^{th} test subject ($c_i = \sum_{j=1}^{n} b_{ji}$).

d_i is the number of false matches in total by the i^{th} test subject ($d_i = \sum_{j=1}^{n} b_{ij}$);

$\hat{q}$ is the observed false match rate;

$\hat{V}(\hat{q})$ is the estimated variance of the observed false match rate.

NOTE The second line of this estimate (in the case m = 1) is the formula given by Bickel [22], which has been experimentally verified [24].

B.4 Estimating confidence intervals

B.4.1 General

B.4.1.1 With a sufficiently large number of attempts, the central limit theorem [31] implies that the observed error rates should follow an approximately normal distribution. However, because we are dealing with proportions near to 0%, and the variance in the measures is not uniform over the population, some skewness is likely to remain until the number of test subjects is quite large.

B.4.1.2 Under the assumption of normality, $100(1-\alpha)\%$ confidence bounds on the observed error rates are given by:

$$\hat{p} \pm \mathrm{z}(1-\alpha/2)\sqrt{\hat{V}(\hat{p})} \qquad \text{(B.9)}$$

where

$\mathrm{z}(\,)$ is the inverse of the standard normal cumulative distribution – i.e. the area under the standard normal curve with mean 0, variance 1 from $-\infty$ to $\mathrm{z}(x)$ is x. For 95% confidence limits, the value of $\mathrm{z}(0{,}975)$ is 1,96;

α is the probability that the confidence interval does not contain the true value of the error rate;

$\hat{p}$ is the observed error rate;

$\hat{V}(\hat{p})$ is the estimated variance of the error rate.

B.4.1.3 Often when the above formula is applied, the confidence interval reaches into negative values for the observed error rate — but negative error rates are impossible. This is due to non-normality of the distribution of observed error rates. Non-parametric methods, such as the bootstrap can be used to obtain confidence intervals in such cases [32-34].

B.4.2 Bootstrap estimates of the variance and confidence intervals

B.4.2.1 Bootstrap estimation reduces the need to make assumptions about the underlying distribution of the observed error rates and the dependencies between attempts. The distributions and dependencies are inferred from the data itself. By sampling with replacement from the original data, a bootstrap sample can be created, from which an alternative estimate of the error rate would be produced. With a large number of such bootstrap samples, an empirical distribution for the estimators can be obtained. This can be used to construct confidence intervals, estimate uncertainties, etc.

B.4.2.2 To illustrate the process, suppose we are estimating the false match rate using a full set of cross comparison with n test subjects, each providing m attempts to be compared against all $(n-1)$ non-self templates. If $x(v, a, t)$ denotes the result of the matching of the a^{th} attempt by test subject v against template t. The dataset X for estimating the false match rate consists of the results of all $mn(n-1)$ cross-comparisons

$X = \{ x(v, a, t) \mid t \neq v \in \{1, \ldots, n\}, a \in \{1, \ldots, m\} \}$. Each bootstrap sample shall be constructed from X in a way that replicates the structure and dependencies in the original data. The procedure is as follows:

a) sample n test subjects with replacement: $v(1), \ldots, v(n)$. (Sampling with replacement means the list is likely to contain more than one occurrence of the same item);

b) for each $v(i)$ sample with replacement $(n - 1)$ non-self templates: $t(i, 1), \ldots, t(i, n–1)$;

c) for each $v(i)$ sample with replacement m attempts made by that test subject: $a(i, 1), \ldots, a(i, m)$;

d) the bootstrap sample produced is:

$$Y = \{ \ (v(i), t(i,j), a(i,k)) \mid i \in \{1, \ldots, n\}, j \in \{1, \ldots, n–1\} \ a \in \{1, \ldots, m\} \}.$$

Many bootstrap samples are generated, and a false match rate obtained for each. The distribution of the bootstrap values for the false match rate is used to approximate that of the observed false match rate.

B.4.2.3 The bootstrap values allow a direct approach for constructing $100(1 - \alpha)\%$ confidence limits: choosing L (lower limit) and U (upper limit) such that only a fraction $\alpha/2$ of bootstrap values are lower than L, and $\alpha/2$ bootstrap values are higher than U. At least 1000 bootstrap samples should be used for 95% limits, and at least 5000 bootstrap samples for 99% limits.

B.4.3 Subset sampling

B.4.3.1 A further approach to inferring the error margin on the observed error rates is to divide the collected test data into disjoint subsets of users, and then generating a DET curve for each subset. The FRVT2002 evaluation [11], for example, used this approach to generate error ellipses.

B.4.3.2 The basic approach to deriving error ellipses is as follows:

a) gather performance results using T test subjects;

b) divide test population into M (e.g. $M=10$) disjoint sets of size $N=T/M$;

c) compute DET curve for each subset;

d) assume a threshold t;

1. find $\mathbf{x}_i = (FMR_i, FNMR_i)$ at the threshold for all sets $i=1,\ldots,M$;
2. compute sample mean $\mathbf{m}=\text{sum}(\mathbf{x}_i)/M$ and sample covariance $\mathbf{\Sigma}=\text{sum}((\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T)/(M-1)$
3. $\mathbf{m}$ and $\mathbf{\Sigma}/\text{sqrt}(M)$ provide an estimate of the distribution of FMR and FNMR observations (calculated for the whole test population) at threshold t which, under the assumption of normality, can be used to determine a 95%(say) confidence ellipse around $\mathbf{m}$.

e) repeat for further thresholds t.