INTERNATIONAL STANDARD

ISO/IEC 19795-2

First edition
2007-02-01

Information technology — Biometric performance testing and reporting —

Part 2: Testing methodologies for technology and scenario evaluation

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

Partie 2: Méthodologies d'essai pour l'évaluation des technologies et du scénario

Reference number
ISO/IEC 19795-2:2007(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 19795-2 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 addresses two specific biometric performance testing methodologies: technology and scenario evaluation. The large majority of biometric tests are of one of these two generic evaluation types. Technology evaluations evaluate enrolment and comparison algorithms by means of previously collected corpuses, while scenario evaluations evaluate sensors and algorithms by processing of samples collected from Test Subjects in real time. The former is intended for generation of large volumes of comparison scores and candidate lists indicative of the fundamental discriminating power of an algorithm. The latter is intended for measurement of performance in modeled environments, inclusive of Test Subject-system interactions.

This part of ISO/IEC 19795 builds on requirements and best practices specified in ISO/IEC 19795-1, which addresses specific philosophies and principles that can be applied over a broad range of test conditions.

This part of ISO/IEC 19795 is meant to provide biometric system developers, deployers and end users with mechanisms for design, execution and reporting of biometric performance tests in a fashion that allows meaningful benchmarking of biometric performance within and across technologies, usage scenarios and environments.

Information technology — Biometric performance testing and reporting —

Part 2:
Testing methodologies for technology and scenario evaluation

1 Scope

This part of ISO/IEC 19795 provides requirements and recommendations on data collection, analysis and reporting specific to two primary types of evaluation: technology evaluation and scenario evaluation.

This part of ISO/IEC 19795 specifies requirements in the following areas:

— development and full description of protocols for technology and scenario evaluations;

— execution and reporting of biometric evaluations reflective of the parameters associated with biometric evaluation types.

2 Conformance

A test shall claim conformance to either the technology evaluation or scenario evaluation clauses of this part of ISO/IEC 19795.

The set of clauses to which a scenario test shall conform differs from the set of clauses to which a technology test shall conform. In addition, the set of clauses to which an identification-system test shall conform differs from the set of clauses to which a verification-system test shall conform. To conform to this part of ISO/IEC 19795, an evaluation shall conform to clauses of this part of ISO/IEC 19795 as shown in Table 1.

Table 1 — Conformance for evaluation methodologies and comparison types

Evaluation methodology	Comparison type	Required clauses
Technology or scenario	Identification or verification	Clauses 5 and 8
Technology	Identification	All of Clause 6, except 6.3.3
Technology	Verification	All of Clause 6, except 6.3.4
Scenario	Identification	All of Clause 7, except 7.3.4
Scenario	Verification	All of Clause 7, except 7.3.5

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 19795-1, *Information technology — Biometric performance testing and reporting — Part 1: Principles and framework*

4 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 19795-1:2006 and the following apply.

4.1 Biometric data

4.1.1
biometric reference
⟨template, model⟩ user's stored reference measure based on features extracted from enrolment samples

4.2 Components of a biometric system

4.2.1
feature extractor
apparatus that extracts features from a sample

4.2.2
biometric reference generator
apparatus that transforms a sample into a biometric reference

4.3 User interaction with a biometric system

4.3.1
acclimatization
reduction, over the course of an evaluation, in a temporal condition of a biometric characteristic that may impact the ability of a sensor to process a sample

4.3.2
effort level
number of presentations, attempts or transactions needed to successfully enrol or match in a biometric system

4.3.3
enrolment attempt
submission of one or more biometric samples for a Test Subject for the purpose of enrolment in a biometric system

NOTE 1 One or more enrolment attempts may be permitted or required to constitute an enrolment transaction. An enrolment attempt may comprise one or more enrolment presentations.

NOTE 2 See Annex B for illustration of the relationship between presentation, attempt and transaction.

4.3.4
enrolment attempt limit
maximum number of attempts, or the maximum duration, a Test Subject is permitted before an enrolment transaction is terminated

4.3.5
enrolment presentation
submission of an instance of a biometric characteristic for a Test Subject for the purpose of enrolment

NOTE One or more enrolment presentations may be permitted or required to constitute an enrolment attempt. An enrolment presentation may or may not result in an enrolment attempt.

4.3.6
enrolment presentation limit
maximum number of presentations, or the maximum duration, a Test Subject is permitted before an enrolment attempt is terminated

4.3.7
guidance
direction provided by an Administrator to a Test Subject in the course of enrolment or recognition

NOTE Guidance is separate from feedback provided by a biometric system or device in the course of enrolment or recognition, such as audible or visual presentation queues.

4.3.8
habituation
degree of familiarity a Test Subject has with a device

NOTE A Test Subject having substantial familiarity with a biometric device, such as that gained in the course of employment, is referred to as a habituated Test Subject.

4.3.9
comparison attempt
submission of one or more biometric samples for a Test Subject for the purpose of comparison in a biometric system

4.3.10
comparison attempt limit
maximum number of attempts, or the maximum duration, a Test Subject is permitted before a comparison transaction is terminated

4.3.11
comparison presentation
submission of an instance of a single biometric characteristic for a Test Subject for the purpose of comparison

NOTE One or more comparison presentations may be permitted or required to constitute a comparison attempt. A comparison presentation may or may not result in a comparison attempt.

4.3.12
comparison presentation limit
maximum number of presentations, or the maximum duration, a Test Subject is permitted before a comparison attempt is terminated

4.4 Performance measures

4.4.1
failure at source rate
proportion of samples discarded from the corpus either manually or by use of an automated biometric system prior to use in a technology evaluation

EXAMPLE A proportion of images collected in a face data collection effort may be discarded due to lack of a face in the image.

5 Overview of technology evaluations and scenario evaluations

This standard addresses two types of evaluation methodologies: technology evaluations and scenario evaluations. A test report shall state whether it presents results from a technology evaluation, a scenario evaluation, or an evaluation that combines aspects of both technology and scenario evaluations.

Technology evaluation is the offline evaluation of one or more algorithms for the same biometric modality using a pre-existing or specially-collected corpus of samples. The utility of technology testing stems from its separation of the human-sensor acquisition interaction and the recognition process, whose benefits include the following:

— Ability to conduct full cross-comparison tests. Technology evaluation affords the possibility to use the entire testing population as claimants to the identities of all other members (i.e. impostors) and this allows estimates of false match rates to be made to on the order of one in N^2, rather than one in N.

— Ability to conduct exploratory testing. Technology evaluation can be run with no real-time output demands, and is thus well-suited to research and development. For example, the effects of algorithmic improvements, changes in run time parameters such as effort levels and configurations, or different image databases, can be measured in, essentially, a closed-loop improvement cycle.

— Ability to conduct multi-instance and multi-algorithmic testing. By using common test procedures, interfaces, and metrics, technology evaluation affords the possibility to conduct repeatable evaluations of multi-instance systems (e.g. three views of a face) and multi-algorithmic (e.g. supplier A and supplier B) performance, or any combination thereof.

— Provided the corpus contains appropriate sample data, technology testing is potentially capable of testing all modules subsequent to the human-sensor interface, including: a quality control and feedback module(s), signal processing module(s), image fusion module(s) (for multi-modal or multi-instance biometrics), feature extraction and normalization module(s), feature-level fusion module(s), comparison score computation and fusion module(s), and score normalization module(s).

— The nondeterministic aspects of the human-sensor interaction preclude true repeatability and this complicates comparative product testing. Elimination of this interaction as a factor in performance measurement allows for repeatable testing. This offline process can be repeated *ad infinitum* with little marginal cost.

— If sample data is available, performance can be measured over very large target populations, utilizing samples acquired over a period of years.

NOTE 1 Collecting a database of samples for offline enrolment and calculation of comparison scores allows greater control over which samples and attempts are to be used in any transaction.

NOTE 2 Technology evaluation will always involve data storage for later, offline processing. However, with scenario evaluations, online transactions might be simpler for the tester — the system is operating in its usual manner and storage of samples, although recommended, is not absolutely necessary.

Scenario evaluation is the online evaluation of end-to-end system performance in a prototype or simulated application. The utility of scenario testing stems from the inclusion of human-sensor acquisition interaction in conjunction with the enrolment and recognition processes, whose benefits include the following:

— Ability to gauge impact of additional attempts and transactions on system's ability to enrol and recognize Test Subjects.

— Ability to collect throughput results for enrolment and recognition trials inclusive of presentation and sample capture duration.

NOTE 3 In online evaluations, the Experimenter may decide not to retain biometric samples, reducing storage requirements and in certain cases ensuring fidelity to real-world system operations. However, retention of samples in online tests is recommended for auditing and to enable subsequent offline analysis.

NOTE 4 Testing a biometric system will involve the collection of input images or signals, which are used for biometric reference generation at enrolment and for calculation of comparison scores at later 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.

Information on differences between technology and scenario evaluations is presented in Table 2.

Table 2 — Distinctions between technology and scenario evaluations

	Technology Evaluations	Scenario Evaluations
What is tested	Biometric component (comparison or extraction algorithm).	Biometric system.
Objective of test	Measure performance of algorithm(s) on a standardized corpus.	Measure performance of end-to-end system in simulated application.
Ground truth	Known associations between data samples and source of samples, subject to data collection errors and intersections in merged data sets.	Known associations between system decisions and independently recorded sources of presented samples, subject to data collection errors and tester failure to note unwanted Test Subject behaviour.
Test Subject behaviour controlled by Experimenter	Not applicable during testing. May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled (unless Test Subject behaviour is an independent variable).
Test Subject has real-time feedback of the result of attempt	No.	Yes.
Repeatability of results	Repeatable.	Quasi-repeatable (if test environment conditions and human factors variables are controlled).
Control of physical environment	May be known to be controlled when biometric data recorded, otherwise considered to be uncontrolled.	Controlled and/or recorded.
Test Subject interaction recorded	Not applicable during testing. May be recorded when biometric data recorded.	Recorded.
Typical results reported	Relative robustness of biometric components or versions of components (e.g., comparison or extraction algorithms). Determine critical performance factors.	Relative robustness of biometric systems. Determine critical performance factors. Measure simulated performance.
Typical metrics	Most error rates. Not end-to-end throughput. Good for large-scale identification system performance where difficult to assemble large test crew.	Predicted end-to-end throughput. False match rate, false non-match rate. Failure to acquire, failure to enrol. GFAR, GFRR.
Constraints	Appropriate test database, e.g., gathered with one or more sensors, the identity of which may or may not be known.	Operational, instrumented system.
Human test population	Recorded.	Real time participation.

NOTE 5 Although in some cases there may be exceptions to the entries in this table, these are the main distinctions.

6 Technology evaluation

6.1 Test design

6.1.1 Goals

An evaluation shall be designed to evaluate a system's enrolment, acquisition and matching functions on the target application.

6.1.2 Application realism

If the test intends to evaluate performance within an application or concept of operations, the test shall be designed and executed so that it mimics the functional (input to output) and procedural (e.g. enrolment or verification processes) aspects of such an application or concept of operations.

EXAMPLE If several images are typically gathered to constitute an enrolment transaction in a real-world enrolment attempt, technology test design should follow a similar process.

For testing purposes, the implementations under test should, if possible, return the comparison score of each comparison attempt.

6.1.3 Determination of appropriate performance measures

Experimenters shall determine which performance measures are applicable to their evaluation, in addition to those listed at clause 6.3.

Test design shall ensure that all required metrics can be generated.

Experimenters shall determine and report on the type(s) of comparison functionality to be incorporated within the technology test. One or more of the following types of comparison shall be specified:

a) verification

b) open-set identification

c) closed-set identification

The rationale for selection of one or more types of comparison functionality within a technology test shall be reported. The comparison functionality evaluated should be applicable to the algorithm in question, such that systems designed to conduct a specific type of comparison such as watchlist identification are tested in a fashion that generates the appropriate type of result.

NOTE Formulae for error rate calculation are provided in ISO/IEC 19795-1:2006, Clause 7.

6.1.4 Implementation primacy

The test plan shall not dictate the method(s) by which the biometric recognition system implements its functions. It is the responsibility of the biometric recognition implementation to perform its functions in its own way.

NOTE The separation of what a tested biometric system does from how it does it is the fundamental construct for allowing offline testing to be done. It is primarily useful in establishing the responsibilities of tester versus supplier. The system under test should be regarded wherever possible as a black box: Its essential function is to render decisions on input samples. The internal details of how this occurs may be proprietary, but in any case, are of no concern to the tester. This construction facilitates the testing of arbitrary biometric samples.

EXAMPLE 1 If a fingerprint is sampled at 1000 dpi, and a test device is known to process only half that, then the tester should a) not execute the down-sampling on the basis that the method for doing so is non-trivial, and b) apprise the supplier of the need to handle the down-sampling internally.

EXAMPLE 2 A set of simultaneously acquired non-frontal face images could be processed by a biometric system and device in at least three ways: selection of the best image; fusion of all images; or stereoscopic synthesis of a 3D model. In any case the biometric system or device decides.

EXAMPLE 3 Most automated fingerprint identification system (AFIS) machines (i.e. those identifying multi-fingerprint records) implement some binning mechanism to partition the database according to some criterion (most simply, the Henry class) and to search only that part of the database of the same category as a user or impostor sample, thereby obtaining throughput benefits, but possibly incurring accuracy losses. Such a tradeoff is achieved by the supplier setting internal binning parameters, and measured by conducting full-scale repeats of the test for each configuration.

EXAMPLE 4 In a study seeking to demonstrate the utility of multiple fingers' prints in a recognition system, the tester should not pass separate samples through the device and perform subsequent score-level fusion, but should instead compose all the imagery as a sample (e.g. in an American National Standards Institute [ANSI]-National Institute of Standards and Technology [NIST] record, or a Common Biometric Exchange Formats Framework [CBEFF] wrapped instance of 19794-2) so as to let the biometric device perform the fusion internally. See ISO/IEC 19794-2, *Information technology — Biometric data interchange formats — Part 2: Finger minutiae data* for further information on CBEFF wrapped instances. See ANSI/NIST-ITL 1-2000 NIST Special Publication 500-245 for information on ANSI-NIST records.

6.1.5 Policies on disclosure of information to suppliers

The tester shall formulate policies before testing begins that govern what information will be disclosed to the suppliers a) before test equipment is configured, shipped or installed, and b) at execution time.

6.1.6 Non-interchangeability of identification and verification attempts

Comparison scores that result from a one-to-many identification search shall not be presented as the results of verification attempts without justification.

NOTE 1 The principle of operational realism indicates that performance shall be estimated from outcomes of attempts (i.e. rejects and acceptances). A verification system shall be evaluated on the results of a sequence of user claims of identity. Likewise an identification system shall be tested over one-to-many searches. Even in the case that a one-to-many search produces a full candidate list, the candidate list is atomic, meaning that it should not be regarded as the result of N verification attempts (to be used in the computation of verification performance).

NOTE 2 The non-equivalency of a single identification attempt and N one-to-one verifications arises because verification can be improved by comparing the user sample with additional hidden samples in a process known as cohort-normalization. The method adjusts the raw single comparison score in order to drive down false accept rates by effectively setting user-specific thresholds. The method trades off performance for throughput because the additional comparisons mean 1:1 verification incurs the expense of 1:M, where M is the size of the hidden biometric reference set.

NOTE 3 The use of cohort normalization is properly conducted internally to the device, making private use of an internally selected enrolled population.

6.1.7 Acknowledgement of models

If a model, approximation or prediction of identification performance is reported in place of, or in addition to, an empirical trial, the model shall be verified to the extent possible with the available data and fully documented.

6.1.8 Sequential use

The test plan shall define the order of use of the test data. This order shall be appropriate to the application. The implementation should process the test data in this sequence.

NOTE 1 Transactions are ordinarily executed separately. Therefore the implementation would need to complete one transaction before commencing the next transaction.