TECHNICAL REPORT

ISO/IEC TR 29189

First edition
2015-06-15

Information technology — Biometrics — Evaluation of examiner assisted biometric applications

Technologies de l'information — Biométrie — Évaluation des applications biométriques assistées par un examinateur

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

Reference number
ISO/IEC TR 29189:2015(E)

© ISO/IEC 2015

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2015, Published in Switzerland

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT), see the following URL: Foreword — Supplementary information.

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, Subcommittee SC 37, *Biometrics*.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved

Introduction

Biometric identification systems such as those used in forensic applications are typically examiner assisted and not automated to the extent that most biometric systems are. This is particularly the case for applications such as latent fingerprint searching where sample quality can be so poor that the system requires human input. Key processes such as sample capture and preparation, enrolment, template generation, matching result adjudication, and final decision that would otherwise require minimal manual intervention are instead heavily reliant on input from experts (fingerprint examiners in the case of AFIS). These experts can interact with the system at each of these stages to prepare, launch, and/or review the results of biometric searches. The execution and performance of the "end-to-end" search process is thus, a combination of the examiner's role (and capability) and the functionality of the automated biometric system.

This partially automated approach to biometrics using "*examiner assisted*" biometric systems provides value both in assisting the human examiner to perform their role more effectively, and in allowing the expertise of the human examiner to be exploited to assist the automated matching process. Therefore, such systems are most likely to be beneficial in non-real time scenarios where the search response is not necessarily required immediately but the throughput of the system is still high.

Understanding the role of the examiner is crucial, as it impacts on the design of the system, the manner in which it is used, how it is tested, and how the system performance and its individual subcomponents are defined and measured.

The main objectives of this Technical Report are to describe the characteristics of *examiner assisted* biometric applications and, where appropriate, to contrast such applications with mainstream biometric applications.

This Technical Report addresses the issues with assessing the system as a whole, or by testing the *examiner assisted* and automated elements separately.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

Information technology — Biometrics — Evaluation of examiner assisted biometric applications

1 Scope

The purpose of this Technical Report is to identify and characterize those aspects of performance testing that are unique to examiner assisted biometric applications.

An examiner assisted biometric system has the following characteristics:

- reliant on the interaction and skill of a human examiner for one or more stages of the complete biometric process, be it data capture, enrolment, template generation, or final decision;
- can incorporate identification functionality, verification functionality, or both;
- will use a combination of the examiner's input and the functionality of the biometric algorithm to execute the complete biometric process;
- will likely have inbuilt examination toolsets to assist the human examiner when enrolling biometric samples or when comparing the match results provided by the biometric algorithm.

Although there is a wide variation in the use of the term "examiner" in the context of an "examiner assisted biometric system", as defined in this Technical Report, an "examiner" typically has the following characteristics:

- field expert in the biometric modality being exploited;
- trained to use the system to an advanced degree of proficiency;
- authorized to override the biometric system's decisions in particular when accepting or rejecting a match decision based on their own examination of the biometric samples and the results returned.

Assessing an examiner's level of expertise is excluded from the scope of this Technical Report. However, the skill of the examiner does have a major bearing on system performance and vice versa. Measuring or assessing the ability of an examiner to employ their skills might be necessary to properly evaluate the performance of an examiner-assisted system.

Other individuals, such as administrative users, or subjects whose biometrics are used within the system are not considered in this Technical Report. It is outside the scope of this Technical Report to consider non-expert examiners.

2 Terms and definitions

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

2.1
examiner
person responsible for examining biometric data and biometric system outputs for the purpose of either preparing data suitable for a system or confirming, overriding, or modifying a decision output from the biometric system

Note 1 to entry: This decision output could be a match decision or simply the location of a biometric feature point (e.g. a fingerprint core and delta points, or the location of eye co-ordinates on a facial image).

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved

2.2
examiner assisted
feature or quality of a process, application, system, or any other element that refers to the fact that an examiner takes part by contributing his/her knowledge and expertise

2.3
suspected match
decision state indicating qualified support on the part of an examiner that a match exists, based on the outcome of the examination process and on the limitations of the relevant comparable data

2.4
suspected non match
decision state indicating qualified support on the part of an examiner that no match exists, based on the outcome of the examination process and on the limitations of the relevant comparable data

3 Symbols and abbreviated terms

AFIS	Automated Fingerprint Identification System

4 Example of an examiner assisted search process

Consider the diagram below in Figure 1 which illustrates at a very high level, some of the basic stages of a biometric search process. With the exception of the “search” which is fully automated, all other processes are potentially assisted by the interaction of a human examiner. Figure 2 shows each of the examiner assisted points in a diagram representing a generic biometric application.

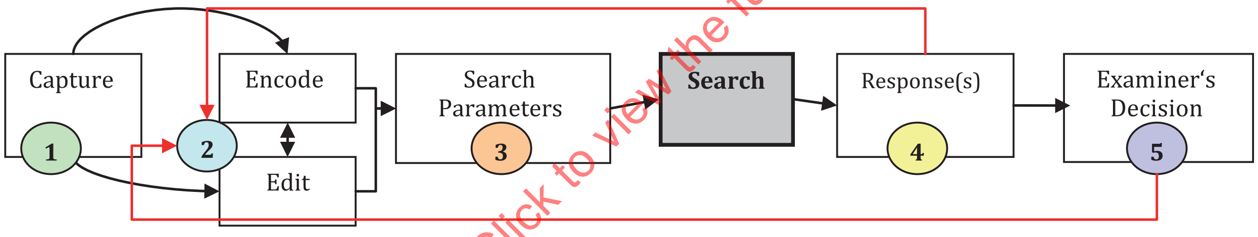

Figure 1 — Basic stages of a biometric search process

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved

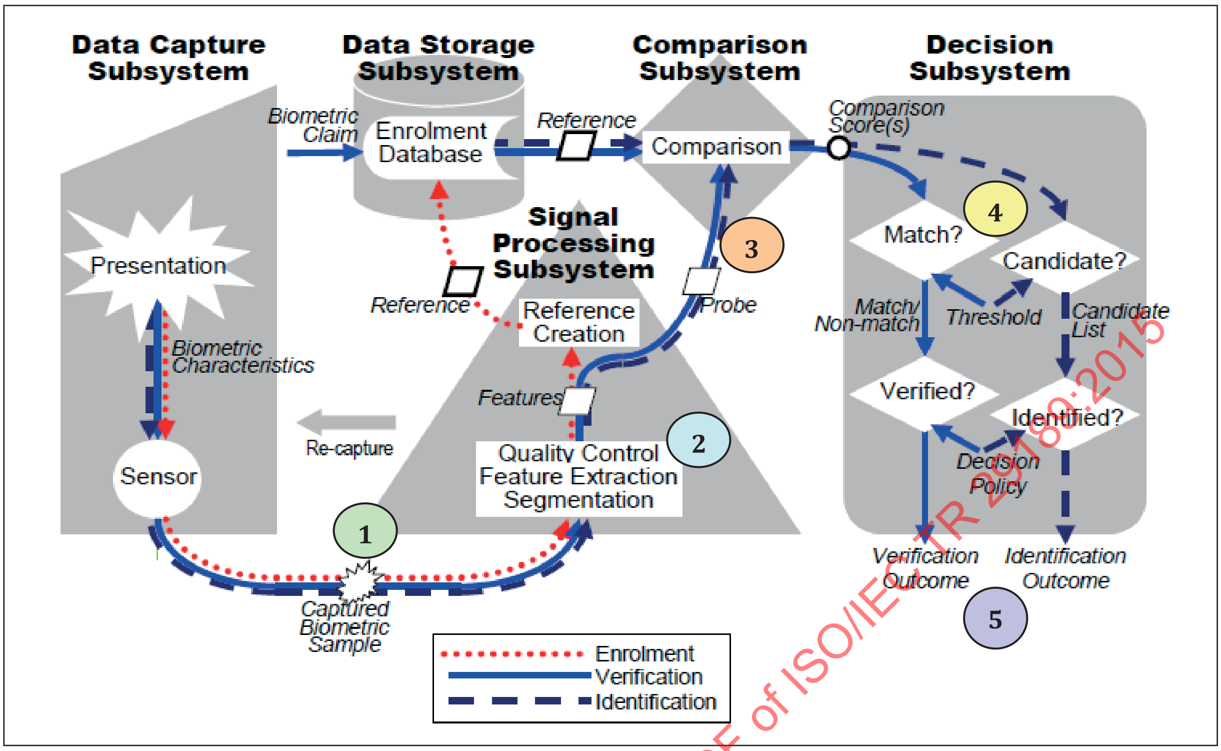

Figure 2 — A generic biometric application highlighted to indicate the examiner assisted points

To illustrate the importance of the examiner assisted stages, consider the role of forensic AFIS examiners. These are fingerprint experts, trained specifically to interact with the system, to fully exploit the functionality of the AFIS in order to prepare, launch, and review the results of the biometric searches. Their interaction at each of these examiner assisted stages shown in Figure 1 can be described as follows.

a) **Capture:** An image is scanned by an examiner or imported directly into the system. If multiple images are available the examiner may select the image(s) that they consider as (the most) suitable quality for searching.

b) **Edit and Encode:** The image is displayed for viewing on a monitor and may be enhanced or edited by the examiner to improve the visibility, and subsequent placement, of features by an examiner. User interface tools are provided to enable the examiner to manually encode features such as fingerprint minutiae, cores, deltas, etc. The examiner may also *override* system decisions about the placement of features such as minutiae, based on their skill and expertise. Some systems may iterate this process to gradually improve the quality of the data with each cycle of manual and automated processing.

c) **Search Parameters:** The examiner may specify search parameters to provide additional data to the matcher in order to maximise likelihood of the search resulting in a match if one exists in the database. Finger position, palm region, orientation, or pattern type may typically be input by the examiner following careful study of the biometric data being searched, based on their domain specific knowledge.

d) **Responses:** The matcher threshold may be manually configurable in order to adjust the number of responses returned. Alternatively, the desired number of responses may be configurable directly, within some system-defined bounds. Some searches may be assigned a certain level of priority (over other searches) depending on the importance of the search outcome. An example of this may be a search conducted on a police system relating to a serious crime.

e) **Decision:** When the output of the search is returned (typically as a ranked list of potential matches when used in a forensic context) the examiner compares the enquiry and respondent images in order to accept or reject possible matches. Even at this stage certain tools on the user interface may be utilised to assist the examiner in performing this comparison. In some cases the examiner may not be able to

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

make an acceptance or rejection decision and may either deem the result indeterminate or return to the edit and encoding step to initiate an augmented search. Such practice should be documented.

At stages 1 through 3, the aim is to provide additional information to the system that cannot be derived automatically. In the case of forensic AFIS the fingerprint data submitted for searching is generally of poor quality, highly varied and thus requires the input of the examiner in order to be able to accurately search the database. Therefore, the value of an examiner interacting with the system is the direct impact that their actions have on performance, especially where data quality is severely compromised.

Although forensic AFIS has been chosen as an example, a wide variety of biometric systems or applications could involve examiner interaction.

The following list provides some examples:

- AFIS — fingerprint matching system, typically using full ten-print enrolments and usually full 10-print probes, often very large scale. Human role is usually to perform a final match/non match decision from a candidate list
- Forensic AFIS — semi-automated fingerprint matching application, often using 10-print enrolments. The human role is to mark-up the latent print and make a final match/non match decision.
- Facial recognition — alias or duplicate enrolment detection. The human role is to perform a final match/non match decision from a candidate list (applications such as visa programs, drivers licences)
- Physical Access Control — a security guard making a human decision of facial match/non match — for example using printed face on ID card/passport — as part of a secondary check or a back-up process in the event of the biometric comparison resulting in a reject decision.
- Adjudication processes — Any decision output from a biometric system that is one of ‘Match’, ‘Non match’, ’Uncertain’ or ‘Suspected Match’, and where all “uncertain” and “suspected” instances are brought to the attention of a trained human agent to resolve, or will be left in the system in the suspected state in anticipation that new biometric or other data, or advances in technologies or changes in policy will allow a resolution of the match.
- Enrolment Quality Checks — a decision, made by a human (possibly aided by automated tools) following a check to determine if the quality of an enrolment sample(s) is of sufficient quality to accept, or if the subject needs to retry enrolment.

Forensic AFIS applications (used for latent searching) are reliant on the interaction of a fingerprint expert at each stage of the complete biometric process. This serves as a good example of an examiner assisted system as it is well defined or understood in comparison to other examiner assisted applications. Therefore, it will be used to illustrate many of the points made in this report.

NOTE To demonstrate the contrast of a forensic AFIS system from that of a standard AFIS consider the scenario of a (*civilian*) fingerprint system being used for checking identity documents at a point of exit or entry. The operator of the system might be involved at a number of stages in the overall system functionality — for example, to assist subjects during the enrolment, or to manually oversee subjects pass through an entry/exit point controlled by the system. However, the operator in this instance would not be an expert, or be required to examine the biometric data at the time; rather their actions would be prompted by the output of the system. It is outside the scope of this technical report to consider such (non-expert) operators or indeed all other users or administrators.

5 Factors to consider when evaluating examiner assisted biometric applications

5.1 General

Any sound evaluation should begin with a thorough examination of the context in which the biometric system is operating, as well as the business processes underlying its use. Such an assessment is generally qualitative in nature, and may consist of interviews or process mapping tools aimed at gaining a sound understanding of current processes and procedures.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

There are a number of factors to consider when evaluating examiner assisted biometric applications, and their relative importance varies with application. At each stage where an examiner is involved with the system, the test design must consider whether this interaction should be specifically addressed or accounted for in the test. It is beyond the scope of this technical report to make specific recommendations; however, this technical report will highlight some points to consider when evaluating such systems. Broadly speaking these considerations can be categorised as 'system-related' or 'examiner-related' factors.

5.2 System-related factors to consider when evaluating examiner assisted biometric applications

5.2.1 Dependencies in the flow process — Where does the examiner interact with the system?

Examiner interaction with the system, at any stage, may have an effect on overall processes and performance. Decisions made at one stage may also have implications for the level of interaction required by the examiner at subsequent stages. Ultimately, there is a trade-off between increasing reliance on an examiner, either in part or across the whole end to end process, against the benefits to performance overall.

The *Edit* and *Encode* process, described earlier in Clause 4, is an example of this. An examiner may be required to spend more time editing and encoding an image to pre-process the search for the matcher to perform better; thereby reducing the time required by the examiner to visually examine images at the decision stage. Alternatively, a system may be designed to minimise the amount of time allocated for pre-processing the search, with greater reliance placed on visual examination of search results.

Therefore, evaluations that measure the performance of examiner assisted biometric applications should take these interdependencies within the overall process into account in order to understand is there is merit in changing the level of reliance on an examiner at any stage of the process.

The test design should attempt to identify and quantify the level of impact that the examiner's actions have on the performance of the system. However, simply removing the examiner from the process may not be feasible or desirable, and where appropriate it may be better to impose controls around what the examiner can do at particular stages in order to isolate and understand the impact of their actions on subsequent stages.

For example in 5.2.2 below the concept of system and stage-level performance is introduced to decompose the overall process into stages that are automated (partially or fully) and those that are entirely reliant on the examiner.

5.2.2 System and stage-level performance measurement

5.2.2.1 Introduction

In order to clearly understand the contribution/impact of human input/interaction in the overall biometric process, it is necessary to decompose the overall process into individually measurable stages. Then when taken collectively, the overall system-level performance can also be computed. The differences between 1:1 verification and 1:N identification systems dictate that different performance measures be described for these two categories. Furthermore, there is a need to describe specific performance measures for the different automated and examiner assisted stages. The decision matrices described in the following sections provide examples of determining system and stage-level performance to assess the stage where the final match decision is made, or overridden, by a human examiner.

In the following tables, green cells indicate correct decisions, and red cells indicate incorrect decisions.

Identification application with human examiner input at the final decision stage only The following describes the means of defining and computing performance at each stage of an identification application where the final match/non match decision is made by a human reviewing a candidate list of potential matches.

There are 5 possible outcomes for the automated stage where a candidate list is generated and for which a true mate biometric reference sample is either previously enrolled (mated) or not enrolled (non-mated) on the system. See the Table 1 below for details.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

Table 1 — Decision matrix for identification application (Automated stages)

	MATED	NON MATED
List generated, Matching candidate ON LIST	True (Match) Positive Identification (known as "Reliability")	(n/a)
List generated, Matching candidate NOT on list	False Negative Identification	False Positive Identification
No list generated (no matches above threshold)	False Negative Identification	True Negative Identification

The two automated "List generated" outcomes in Table 1 devolve through the examiner assisted stage to the three possible outcomes shown in Table 2 below. For each of these outcomes the performance of the examiner can be determined.

Table 2 — Decision matrix for identification application (Automated and examiner stages)

	Automated stage	Examiner Stage	
List generated, Matching candidate ON LIST (mated)	True Positive Identification (mated)	True Positive Identification Confirmed	True Positive Identification Rejected
List generated, Matching candidate NOT on list (mated)	False Negative Identification	False Negative Identification Confirmed	False Positive Identification by Examiner
List generated, Matching candidate NOT on list (non mated)	False Positive Identification	False Positive Identification Rejected	False Positive Identification by Examiner

For each cell in the tables above, the appropriate performance metric can be determined.

In the first row the examiner stage increases the system-level False Negative Identification rate when "True Positive Identification Rejected" is the outcome. In the second row, the examiner stage increases the system-level False Positive Identification rate when "False Positive Identification by Examiner" is the outcome.

Note that the overall system-level error rates are only modified when the examiner decision differs from that of the automated stage. Decisions made by examiners can increase or decrease system error rates. If an examiner makes a correct decision when the automated decision was incorrect, the system error rate will decrease. If an examiner makes an incorrect decision when the automated decision was correct, the system error rate will increase.

5.2.2.2 Verification application with examiner input at decision stage

There are several examiner stages possible for a verification application. This section will not address any quality review stages that may be performed during enrolment and assumes all tests are performed with enrolled test subjects. The predominant examiner stage is invoked when a genuine user is falsely rejected, and submitted to an examiner assisted "secondary" inspection. A third examiner stage could occur, but only if an attended access point application utilized an additional layer of security, for example by checking the face printed on an ID against the ID holder even after successfully matching the primary biometric (this may be a rare situation).

The decision table for the automated stage has 4 outcomes for genuine and impostors, either declared matching or non matching".

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved

Table 3 — Decision matrix for verification application

	Genuine	Impostor
Automated MATCH decision	True Match	False Match (A)
Automated NON -MATCH decision	False Non-match (B)	True Non-match

For the case of a false non match (B), the examiner "secondary" stage has two possible outcomes:

Table 4 — Examiner stage decision matrix for False Non Match outcome (B)

	Examiner stage	
Automated NON -MATCH decision (genuine)	True Match	False non -match confirmed (C)

To compute the system and stage-level performance rates, see the formulas below, given that the number of genuine = Ng, number of impostors = Ni

Table 5 — Decision matrix with system and human stages (Verification)

	Automated Stage	Examiner Stage	System-Level
False Match Rate acceptance	(A)/ Ni	(n/a)	(A) / Ni
False Non Match Rate rejection	(B)/ Ng	(C) / (B)	(C) /Ng

These examples will increase in complexity as more intricate or dynamic decision policies are employed.

5.2.2.3 Suspected match and Suspect non match decision states

"Suspected Match" and "Suspected Non Match" decision states compensate for low-quality biometric probes and/or references that cannot be adjudicated at the confidence level of the system and training policies, but are in the examiner's opinion likely or unlikely to be from the same biometric data subject.

A system that supports these decision options provides examiners with flexibility not present in biometric systems where the comparison decision is limited to an "either-or" outcome, associated to a "match" or "non match" decision.

A Suspected Match option can avoid a Non match result in such cases, and instead promote further biometric and non-biometric searching to confirm an outcome, resulting in a more accurate or confident final decision outcome of "match" or "non match".

There are some potential disadvantages of a system supporting the use of "Suspected Match/Non match" decision states. It may lengthen the process of reaching a final decision outcome; it may influence the decision of examiners subsequently reviewing the search responses; and it may generate additional work for other examiners who may also be required to follow up or review "Suspected Match/Non match" outcomes.

5.2.3 Measuring 'true' operational performance

Even if a system correctly returns the true match at top position, if the examiner mistakenly fails to confirm this match it will result in a 'miss' (false reject) as operationally their decision has resulted

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

in a missed identification. Therefore, the examiner's ability to correctly identify true matches has implications for the operational end-to-end accuracy achieved.

In this scenario a true measure of operational performance, with respect to accuracy, needs to include the examiner's decision, not just the decision output by the biometric application.

Therefore, tests of examiner assisted systems will typically include this final stage. This has a number of implications, specifically on the analysis of search results. If examiners are required to review search results, the amount of time necessary to perform the test increases. This may limit the number of transactions that can be processed during the trial period.

Conversely excluding the examiner decision from the scope of the tests has implications on the design, execution, time and resources required for the evaluation. However, it also limits what can be measured or inferred from the observed results, in terms of their relevance to the system's overall performance in live operation, where the examiner assisted parts are included.

In addition, testing accuracy is dependent on having a known ground truth. However, in some cases the ground truth may itself be created as a result of an examiner assisted process, which may be the very process being evaluated. Hence the creation of such ground truth datasets should also be considered in test design.

5.2.4 The impact of prior probabilities on human performance

In some applications human examiners will rarely encounter a match outcome because the occurrence of match encounters is naturally low. For example, in a counter terrorism application in which a biometric system matches samples against a watch-list the human examiner will rarely encounter a match from even an accurate biometric system simply because the underlying event is rare. This quantity, the prior probability of a match, can vary over many orders of magnitude between applications. For example, in routine law enforcement, the prior probability might approximate the recidivism rate i.e. the fraction of offenders who reoffend.[35] In the case when prior match probabilities are low, and the system produces many non match candidates, the human examiner might become complacent. Indeed system owners sometimes deliberately supplement the legitimate workload with planted matches so as to maintain and test examiner alertness. Evaluations of examiner-assisted systems should therefore:

a) Estimate prior match probabilities

b) Consider whether to inject match occurrences into the workflow.

c) Consider surveying the examiners to assess perceived prior probabilities

d) Consider periodic or secular trends in prior probabilities e.g. diurnally.

5.2.5 Confidence Levels

A system that supports an examiner-specified confidence level along with the Match, Non match or other outcome, will allow examiners to qualify their decisions as part of the overall system outcome. Decision states described in 5.2.2.3 can work in conjunction with examiner-provided confidence levels to provide additional flexibility in post-biometric processes

A confidence level provides further information to others involved in the biometric and post biometric processes which allows for more complex decision outcomes or policies to be applied that are better for integration into the overall system or business. It also allows for qualification where one or both biometric probes and/or references are too low in quality to allow for a completely confident response to be made.

The option or requirement to provide a confidence level for the decision provides further value from the outcome of the biometric process by improved reliability and efficiency of the biometric application within a business's overall management of information.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

Table 6 — Example of examiner-provided confidence levels

Examiner Confidence Level Scale	Description of examiner decisions and confidence levels
Match	Based on the outcome of the examination process and the presence of sufficient relevant comparable data, the examiner's opinion is that of very strong support for a match decision.
Suspected Match	Based on the outcome of the examination process and on the limitations of the relevant comparable data, the examiner's opinion is that of qualified support that a match exists.
Suspected Non match	Based on the outcome of the examination process and on the limitations of the relevant comparable data, the examiner's opinion is that of qualified support that no match exists.
Non match	Based on the outcome of the examination process and the presence of sufficient relevant comparable data, the examiner's opinion is that of very strong support for a non match decision.
No Opinion	Based on the outcome of the examination process and the limitation of the relevant comparable data, the examiners is unable to form an opinion

This example adapted from[34] which could be applied to a variety of biometric modalities where a forensic approach is taken by examiners to the outcomes of biometric system matches.

5.2.6 The impact of automated systems on human performance

Automated systems can affect a human examiner in several ways.[3] Firstly, if the examiner is not presented with a matching image they are unable to make a positive identification. For example, as is the case with some face recognition systems, only images that achieve a score above a certain threshold setting will be presented to the examiner. Therefore, it is possible that if the matching image does not score highly enough that this image will not be presented to the examiner as a possible match.

Secondly, the biometric system needs to be appropriately calibrated to allow for the human examiner to operate at an optimal level of performance. This calibration should recognise that the capabilities of different human examiners will vary. Although the human can contend with a certain number of false alarms, overloading the examiner can lead to poor decision making. The system should work with, rather than against, the examiner to provide the highest level of performance.

Finally, if the examiner is shown an image of inadequate quality this may compromise the ability of the examiner to make a decision. For example, research into human face recognition suggests that humans are vulnerable to the effects of viewpoint pose and illumination, so much so that if the two facial images under examination are in different poses, or if the two images are taken under vastly different lighting conditions, the human examiner may have significant problems identifying them as matching subjects.[4]

5.3 Examiner-related factors to consider when evaluating examiner assisted biometric applications

5.3.1 An Examiner's perception of the system's accuracy

The *Examiners* of examiner assisted systems can be considered as 'educated' users in the sense that they will normally have sufficient understanding of how the technology works in order to be able to interact with it effectively. Indeed, training courses in many forensic disciplines often include modules on biometrics technologies, pattern classifiers and algorithms, designed to develop a basic level of competence in these areas in the context of an examiner's work. As a result, in addition to their own domain subject matter expertise, such users often have a detailed understanding of the functionality of the biometric technology being used.

However, there is another notable effect to take into consideration in employing examiners. By understanding the technology, examiners are able to indirectly gauge the accuracy of the system, through their general interaction with it. They may over time observe a trend in the outputs of the system,

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

compare this to their own findings, and subsequently form an opinion of the system's performance. For example, an examiner may come to have little confidence in a system that routinely returns correct matches at poor ranks in a candidate list and may even start to override decisions in the belief (rightly or wrongly) that they can improve on the performance. Therefore, the examiner's *perception* of accuracy can have a direct effect on the performance of the system, for better or for worse, as this perception is likely to impact on their use of, and confidence in, the system.

Biometric systems that provide quality metrics in relation to the biometric probes and biometric references, for example during enrolment and matching stages, can help to manage an examiner's perception of performance, especially where data quality is poor, as well as inform how the examiner chooses to interact with the system based on the quality of the data. Overall this may help improve performance.

EXAMPLE In a facial recognition system the examiner can be provided with key data to understand how useful or reliable each image is for accurate matching and comparison, which can include:

- a system derived quality value for each image;
- data on the number of pixels between the eyes for the images;
- information on the image's size, dimensions, format and compression;
- the angle and orientation off centre of the subject's heads;
- EXIF data on the image capture;
- ICAO requirements compliance etc.

If the examiner is aware of any quality shortcomings relating to the biometric sample(s) being processed this may help to account for any resulting poor performance observed. Moreover, the examiner may be able to take pre-process the data, where acceptable to do so, to compensate for any deficit in quality of the sample.

5.3.2 Usability and examiner acceptance

To be effective, biometric systems need to be usable and accepted by examiners.[5] A system is considered usable if:

- the intended examiners can meet a desired level of performance operating it;
- the amount of learning or practice required to achieve this level of performance is appropriate;
- the system does not place any undue physical or mental strain on the examiner;
- the examiners are satisfied with the experience of interacting with the system.[6]

User acceptance is imperative for any system to be efficient and effective. It directly contributes to the establishment of trust in a system, which in turn is one of the main determinants of performance.[7] Therefore evaluation design should assess the usability and acceptance by the examiner of the system under test. Such measurements may be in the form of surveys, functional reviews, observational appraisal by human factors experts, or other assessment mechanisms.

Access to appropriate tools that examiners need to efficiently and confidently examine system generated matches will be essential for examiners to use their training and expertise. For example, facial image comparison as part of a facial recognition system operation may require a large visual display, with tools to pan and zoom both images to view them in a synchronised manner and normalise them for size and orientation. Evaluation of examiner-assisted applications should utilize such tools in a fashion consistent with that of real-world operations.

5.3.3 Training and expertise

Training and expertise are considered key factors that impact on examiner performance, however, the link between training, expertise, and performance is poorly understood. The type of training examiners complete tends to vary markedly from intensive training to ad hoc on-the-job training. Expertise levels

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

also vary markedly. Both are, however, central to the performance of human examiners and need to be considered in any evaluation of an examiner assisted biometric system.[8]

5.3.4 Workload

Workload is another factor that can have an impact on examiner performance.[3] Examiners' work demands will vary greatly according to different environments. Examiners may work to achieve a daily quota, to cope with a high volume of alarms, to multitask, and/or as a member of team who together make the final decision. All of these factors impact on the performance of the examiner. In some environments, many alarms will be false but examiners will still need to remain vigilant in order to detect the rare true alarm. A person's ability to maintain vigilance and attention reduces over time. Inattention can be induced by fatigue, which has been widely studied as the cause of errors. In[11] it was found that humans often perform worse when placed under a time pressure as opposed to normal conditions. Such pressure can cause stress and promote risky or biased decision making.

In some circumstances identification may not be the only responsibility of the examiner. Most research in the area of multi-tasking suggests that the introduction of subsequent tasks distracts from the original task, impacting negatively on performance.[12] The requirement to consult a decision aid or check multiple screens for information can direct attention away from the core task.

There may also be performance differences between individuals and teams. These differences may have important implications for resourcing and the accuracy of the system. It is also necessary to note that different examiners working in the same environment may have different reactions to workload stressors, and/or be less susceptible to fatigue and distraction.[13]

Therefore, evaluation design should characterize examiner and/or team workload factors in the examiner assisted system.

5.3.5 Bias in decision making

Human examiners may be subjected to a number of different biases that can directly or indirectly affect their decision making and performance. It may be important in an evaluation to determine if biases are present in examiners' actions or decisions.

— Cognitive Constraints — Biases may relate to cognitive constraints, such as perception, judgment and decision making. These issues have been reviewed in.[14]

— Conformation bias — The role of confirmation bias,[12] that is, a tendency to confirm any initial theory or preconception whilst avoiding disconfirming information in the forensic identification setting has been studied extensively.[13],[14],[15]

— Threshold bias — Research has found that decision makers have a threshold that must be reached before a certain decision, such as to identify a person, can be made. Many factors can influence this, including time pressures, accountability, expectations and emotions surrounding the task.[18] It has also been found that human examiners often base their decision making on past experiences more so than on logic or rationality.[12]

— Reliance on technology — The over reliance of examiners on the technology to automatically flag a threat (or not) may lead to a failure of examiners to adequately interpret output themselves, potentially leading to biases in decision making. In addition, the consequences of making an inaccurate decision may need to be balanced against the need to process individuals as quickly as possible to meet logistical (or administratively imposed) requirements. This can lead to a speed/accuracy trade-off (speed bias versus accuracy bias).

— Speed trade-off bias — Responses made under speed bias can be faster than those observed under accuracy bias; the reduction in reaction time may be accompanied by an increase in error rates, with speed being traded for accuracy.[17]

In some cases, an evaluation may be designed to determine why biases exist and how they can be removed.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

5.3.6 Individual differences between examiners

There is some evidence to suggest that individual differences (certain abilities or personality characteristics for instance) may be associated with better examiner performance.[18] Knowledge of innate abilities or other characteristics important for examiner performance are useful as they may guide personnel selection and/or development of training. Little research, however, has gone into:

a) identifying what these specific abilities are,

b) predicting performance in the real-world application on the basis of these abilities, and

c) the extent to which any such abilities may be learned or innate.

A consideration of individual differences during evaluations may, therefore, be useful.

As it is likely that particular individual abilities may be associated with better performance on examiner tasks, tests of cognitive, perceptive and other abilities that may be of relevance (e.g., personality) could be administered to examiners. An example of this approach can be found in[1] while details of appropriate tests can be found in,[22],[23],[24],[25],[26],[27] and.[28] The results obtained from these tests could then be related to the examiner's detection performance, discrimination and bias measures, speed and confidence ratings. This may provide a useful explanation for the results obtained for these measures. Additionally, these results may potentially be used as a basis for identifying the types of individuals that may be best suited to examiner roles.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

6 Performance evaluation of examiner assisted systems

6.1 Types of Evaluation

Evaluations of examiner assisted biometric systems generally fall into three categories: component, scenario and operational.[19]

Component tests evaluate the performance of a discrete, examiner assisted subcomponent of a biometric system (e.g. encoding software or a capture device). A component evaluation may be able to isolate performance factors more efficiently than a scenario test, or operational evaluation of an entire system.

For example, a component evaluation may be appropriate when comparing biometric feature location and sample mark-up software whose workflows and underlying technologies differ. One could assess (as a first-order measure of the component's effectiveness) examiner throughput or the proportion of samples whose automated decisions conflicted with the examiner decision.

Scenario evaluations assess the examiner-assisted biometric systems in a simulated real-world environment in order to determine overall system performance in an operational context. In general, the role of the human examiner should be an important consideration in scenario tests, with regards to both the performance of the technology and the impact of the human examiner on overall performance. It is important to recognise that results of scenario tests may not necessarily be representative of those achievable in the operational environment due to differences in the environment, subject behaviour, and the various impacts on the human examiner (i.e. fatigue, workload, consequence of actions and bias).

Operational evaluations employ an examiner assisted biometric system in a real world environment. One of the many purposes of an operational evaluation may be to determine if a system is sufficiently mature to meet operational performance requirements.

In addition to examining the technology and the human examiner, such evaluations may encompass the ability of the business processes to support decision making and the capacity of the system to function for extended periods without system failures. However, whilst operational tests can provide the most realistic assessment of the suitability of a system for deployment, there are typically many challenges that hinder conducting them. For example, it may be impermissible to conduct tests on a live operational system, especially one using biometric data, due to privacy restrictions other other risk factors that prevent testing a system in a live environment). As such, additional controls (in the operational environment) can be incorporated into the operational trial to mitigate some of the limitations of

© ISO/IEC 2015 – All rights reserved

operational trials. This may include launching test transactions in which ground truth is known with live transactions or on the system in an offline state. Following the operational tests, it may be possible to reuse some of the data collected to conduct additional analyses, offline.

EXAMPLE Biometric face images could be reprocessed by the algorithm against different watch lists, or latent prints could be replayed to a larger pool of human examiners for evaluation.

6.2 Performance measures for examiner assisted biometric systems

6.2.1 Introduction

Performance measures are a critical component of any evaluation. There are a range of performance measures that can be used to assess examiner assisted biometric systems, including those focused specifically on the technology and those specific to the human examiner.

6.2.2 Measures of accuracy

One expression of the performance of an examiner assisted biometric system is the proportion of transactions in which the examiner decision is confirmed, altered or differs from the automated system decision. A system in which a high percentage of examiner decisions differ from automated system decisions will typically be inferior to one in which a low percentage of examiner decisions differ from automated system decisions. This proportion can be expressed as a function of many variables, including but not limited to:

- Transactions in which the examiner marked up the submitted sample(s)
- Sample quality
- Priority or rank of reviewed results

In an evaluation in which the examiner's decision is also subject to an external assessment (e.g. to tabulate examiner errors), accuracy can be expressed as the proportion of transactions in which incorrect decisions are made at the automated stage and examiner stage.

The tables in 6.2.3 are an example of one such approach of breaking down the automated and examiner elements of performance with respects to the impact on match decision. By determining the metrics for each table one can quantify the performance of the overall system and compare this against the performance of either including or eliminating the examiner assisted element.

Measures for accuracy for examiner assisted systems may, especially for identification (one-to-many) systems, have a lower system threshold comparison score to increase the likelihood of identification of persons on watch lists. This may increase the automated stage false positive identification rate, requiring examiners to review more matches to finalise the outcome of possible positive matches.

This is done to compensate for poor biometric references and poor biometric samples, and where the business process is prepared to have examiners review these possible positive matches to confirm or override the system results. It may also be done so that staff in a capture zone are eliminated automatically from being referred to an examiner, in effect making sure they do not trigger any other watch lists and waste time and resources. System accuracy then is better measured in terms of the ground truth (if known) outcome of system and examiners, and may in operational environments be measured in terms of detection improvements over other non-biometric detection methods. For facial recognition systems this may require examiners to review an image gallery in rank match order based upon a threshold or a gallery size to determine or not if a match exists and possible provide a confidence rating and other data as to why they have made this decision.

A system may also have multiple watch lists with different matching thresholds. Some watch lists may have high threshold that have a high reliability of being accurate to keep a low false alarm rate for examiners. Whilst other watch lists may have lower or variable thresholds based on the importance of detection, and these variable thresholds may change based upon examiner resources, system throughput and/or threat/risk levels.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

6.2.3 Examiner-assisted performance considerations in watch list scenarios

Examiner-assisted performance in watch list scenarios may be impacted by the system's ability to perform the following functions:

— to aggregate all the images for a single individual to avoid requiring examiners to adjudicate multiple examinations in a short interval, to avoid clearing false matches against known members of staff (a watch list to override some lower priority lists but may still be overridden by other higher priority lists),

— to prioritise the more important matches for examiners in queues or even to interrupt ongoing examinations (based on matching threshold and watch lists)to send the comparison to a secondary examiner (for review or finalization) or to response resources to act upon,to detect persons who appear to be deliberately evading biometric collection in systems that are passively/covertly collecting,,

— to provide quality information on probes and references, as well as quality assessment tools, to allow the examiner to quickly and accurately resolve a potential match ,

— to provide relevant, useful data about the matched subject when a match is confirmed such that the subject can be quickly identified, and,

— to support a confirmed match outcome in a timeframe that supports the business processes workflow to respond with other actions and resources.

For example, in a surveillance application, performance may be measured in terms of face recognition matches confirmed by the examiner rapidly and with sufficient identifying information. This enables a timely, efficient response on the part of authorities.

6.2.4 Discrimination and bias

Methodologies for software development testing provide two additional measures (discrimination and bias) that can be useful in further analysing the accuracy of human judgments. Discrimination relates to a person's ability to distinguish target and non-targets, whereas bias relates to a person's tendency to raise an alarm. Two people may have the same discrimination ability, but they may differ in their bias. These measures are important for two reasons. Firstly, discrimination and bias are measures that describe human performance in a way that is not easily obtained by examination of only the detection performance in terms of hits and alarms. Secondly, these measures can assist in locating the source of poor performance (i.e., they may provide potential insight into why performance is poor and how it might be improved). Further information about the calculation of these statistics is contained in.[20]

6.2.5 Examiner Decision Confidence

In conjunction with examiner decisions, the human examiner's confidence in their decision can be measured. This data may be collected using the categories in Table 5 (in 6.2.4 of this document). Alternatively it may be collected on either a Likert scale, such as a five-point scale (e.g. 1 = Very Confident to 5 = Not Confident at All); or a continuum scale (e.g., as a percentage). Confidence ratings should be collected immediately after the examiner has made a decision. Confidence measures may be very useful in examining the decision making processes of the examiner. For example, inaccurate responses that are associated with high confidence suggest a lack of awareness of the quality of their responses and may reflect unwarranted faith in the abilities of the technology and/or the examiner's abilities. Similarly a significant correlation between time and confidence (i.e., that longer response times are associated with lower confidence ratings) could indicate tentativeness of response.[3]

6.2.6 Processing speed

In terms of assessing the speed of biometric systems it is often important to consider the time taken by both the technology and the human examiner to make a decision.[23]

The mean time and standard deviation should be considered for both the human and technical parts of the system individually, as well as the system overall. The collection of processing times involves

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

defining appropriate start and stop times (and associated cues for determining these) and noting the relevant time on the basis of a time date stamp (video recording evaluations may assist with this).

6.3 Usability assessment

6.3.1 Introduction

Usability assessments of examiner assisted biometric systems may provide insights into factors, not previously considered, that affect overall system performance. The identification of usability issues can lead to improvements in the technology and/or business practices. These factors can be assessed using a variety of methods as outlined below and in.[29]

6.3.2 Qualitative observations

During an evaluation, qualitative observations of examiners interacting with the system can be undertaken. This may highlight common errors that are experienced in using the system, from which frequency observations could be collected to generate statistics on the prevalence of these issues. An understanding of the common issues experienced by examiners may suggest ways in which the system could be modified to improve performance. These results could be linked to processing speed to provide an indication of how errors affect processing time and functionality. The procedure for the collection of qualitative observations is discussed in further detail in.[30]

6.3.3 Questionnaires

Usability questionnaires may also be used to complement findings obtained from qualitative observations. A usability questionnaire (for examples see,[31],[32] and[33]) is designed to solicit the human examiner's view on each of the elements of the system under examination, covering both the performance and usability of the system. This type of questionnaire often consists of multiple parts, including optional demographic questions (e.g. details of the examiner's current employment, training and experience) and a series of performance and usability questions. The performance and usability section of the questionnaire consists of system related questions, typically in terms of the examiner's perception of the effectiveness and efficiency of the system, and their overall satisfaction with using the system. Examiners are often asked to respond on a Likert scale about various aspects of the system, including rating themselves as users on the specific system being examined. Finally, the examiners may be asked to list, in a free-text format, the things that they liked and disliked about the system under examination. If deemed necessary, follow-up questionnaires can be given to the examiners which target specific issues uncovered in earlier questionnaires.

6.3.4 Interviews and focus groups

In order to obtain more in-depth information about usability and performance issues in the specific operational context, focus groups and/or interviews may be conducted with examiners. It is necessary to conduct interviews with a number of different users to obtain a broad range of views from that operational setting. Interviews should be flexible and consist of open-ended questions designed to elicit a descriptive narrative of the procedures and practices used. Additionally, opinions of the examiners on the functionality of the system, and ways in which it may be modified and improved, could be discussed.

6.4 Reporting results

Any sound evaluation will ensure that the results are not only presented in the right context, but that they also address the fundamental questions concerning the examiner.[3] Conventional methods of reporting performance by means of FAR, FRR, etc. are often inappropriate for examiner assisted systems. These metrics are defined with respect to first rank only, or for verification-based systems, as opposed to a ranked candidate list which is returned for examiner assisted (or forensic AFIS) systems. Measures of accuracy should be reported in terms that are meaningful to the examiner; that is, they should report in business terms how the system is performing, and may also seek to quantify the impact that the examiner's actions are having on performance.

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

In 2004, the UK Police Information Technology Organisation performed operational benchmarks of forensic AFISs. The method used to report accuracy was based on measures of "Reliability" and "Selectivity". This is described in more detail in.[2] Reliability gives an indication of how many matches are returned correctly and Selectivity helps determine the level of examiner/examiner effort required in making the comparisons. Taken together these measures report accuracy in a manner that allows one to maximise the efficiency of the system, based on balancing matcher performance with examiner intervention.

These measures represent an alternative approach to reporting matcher performance in a way that is readily understood by both the business and the examiners.

— 'Reliability', can be defined as: the likelihood of the match (if it exists) being returned on a respondent list.

— 'Selectivity' can be defined as: the average number of potential but non-true matches an examiner would have to review manually until the first true match (if present) is detected.

NOTE In testing, reliability can be defined more generally as the fraction of searches for which a match exists that yields the matching candidate in one of the top R ranked candidates with comparison score above a system threshold. During analysis, either the threshold value and/or the rank condition can be tightened or relaxed.

Likewise, selectivity, defined over searches for which no match exists, is the expected number of candidates returned above a system threshold.

6.5 Applying controls in evaluations

6.5.1 Introduction

Some factors can be controlled without jeopardising the operational quality of the test. These vary from system to system and on the purpose of the evaluation, thus there is no generic rule of what should or should not be controlled. The papers referenced in[1]and[2]report on two real life operational benchmarks conducted in the mid 90's and 2004, where the approach described here was employed to evaluate forensic AFIS systems in the UK. The sections below provide examples of factors that were controlled in those particular tests whilst preserving the operational relevance of the evaluations.

6.5.2 Controls for examiner expertise

Examiners will have varying levels of both professional expertise and familiarity with the system. A large number (e.g. 20+) of examiners selected from a cross section of the examiner community (selected randomly or stratified) to reflect the range and varying levels of experience and skill will minimise any resulting bias. This is particularly important when using this testing approach to differentiate between two or more systems for the purpose of evaluation.

6.5.3 Controls for examiner decision bias

Examiner decision bias has an effect on all examiner assisted processes throughout the system. The causes can be numerous — be it training, familiarity, usability, etc.

Biometric system toolsets for examiners that have not been thoroughly tested for assisting matching accuracy and reliability, or for which appropriate training has not been provided, may in themselves produce decision bias. Experimenters may attempt to discern detection bias resulting from use of unfamiliar or untested biometric system toolsets, perhaps by executing parallel tests with alternative biometric system toolsets, or without such unfamiliar and untested toolsets.

For example facial recognition identification systems require the making of face comparison decisions but there are at present no internationally accepted standards on the minimum image qualities for human comparison, the correct process to compare, or what tools are necessary and how they should be applied. There are also studies that show that some tools such as face image on top of face image

STANDARDSISO.COM : Click to view the full PDF of ISO/IEC TR 29189:2015

© ISO/IEC 2015 – All rights reserved