UNDERSTANDING BIOMETRIC PERFORMANCE EVALUATION

Introduction

When working with biometric systems or components, two very fundamental questions often arise:

- How could you measure the accuracy of a biometric system (or components thereof)?
- How to compare different systems with each other?

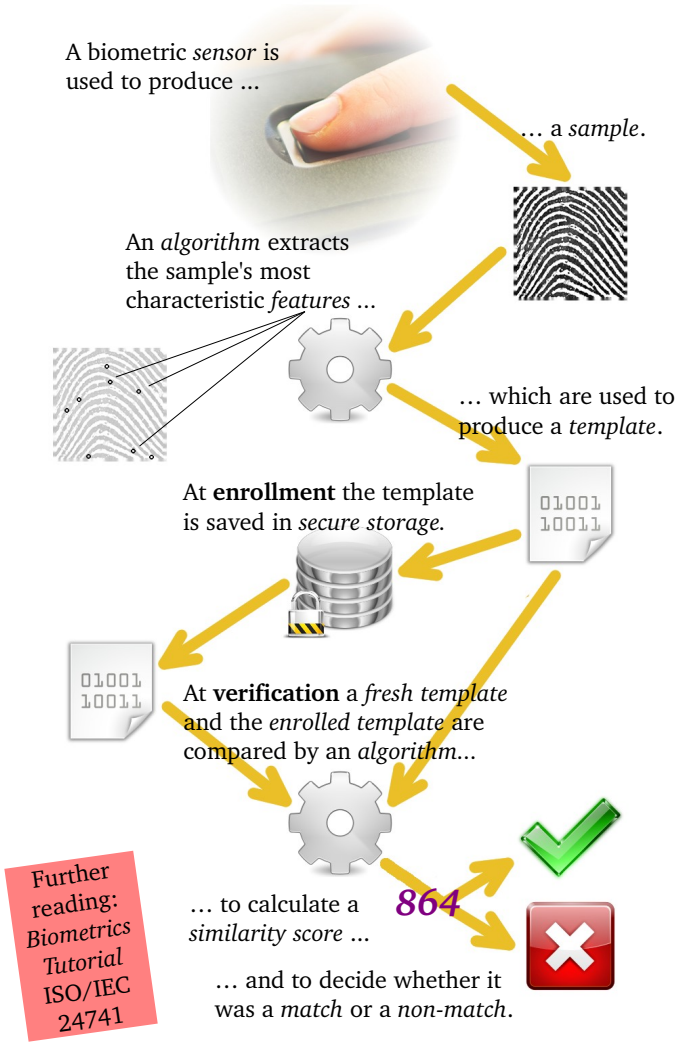
One way could be to check biometric performance figures reported by vendors. They often come in a form similar to this:

FRR 1% @ FAR 1 / 10,000

What do these figures actually say and how were they calculated?

The Fundamentals of Biometrics

Let's start with the fundamentals. Most biometric systems do verification in a similar manner (here shown for fingerprints, but the same principals apply to other modalities):



Further reading: Biometrics Tutorial ISO/IEC 24741

The purpose of the verification is to tell if the two templates being compared come from the same object, e.g. the same finger. The matching algorithm analyzes the templates to produce a similarity score and if the score reaches a certain threshold the algorithm decides that it is a match.

A perfect biometric system would always make correct decisions, but in reality this is not possible. Depending on

- the amount of useful information available in samples that could be used to characterize objects, and
- the capabilities of the complete biometric system (and the algorithms in particular),

the decision is more or less probable to be correct.

The amount of information available varies due to many factors. In the case of fingerprint biometrics it depends on e.g. the number of fingers used, the fingerprint sensor size and resolution, image quality, and overlap between samples.

The capability of the algorithm does not solely depend on the matcher calculating the score – precise extraction of features from the sample plays an equally important role. Thorough analysis of the biometric features is also often limited by computational constraints and time limits.

Genuine attempt

A single attempt by a user to match his/her own stored template.

Impostor attempt

The opposite – a user's template is matched against someone else's template.

(For impostors it's distinguished between *zero-effort* and *active* impostor attempts depending on the user intentions – an accidental match or active means to fake characteristics. Zero-effort impostor attempts are usually implicitly understood when evaluating biometric systems.)

Incorrect decisions come from that poor genuine attempts in some cases score lower than the highest scoring impostor attempts. Selection of the score threshold used by the matcher will determine the proportions of each attempt type that is falsely categorized. The score threshold is thus used to tune biometric systems: A lower threshold targets the system at user convenience (fewer genuine attempts are rejected) while a higher threshold targets it at security (fewer impostor attempts are accepted).

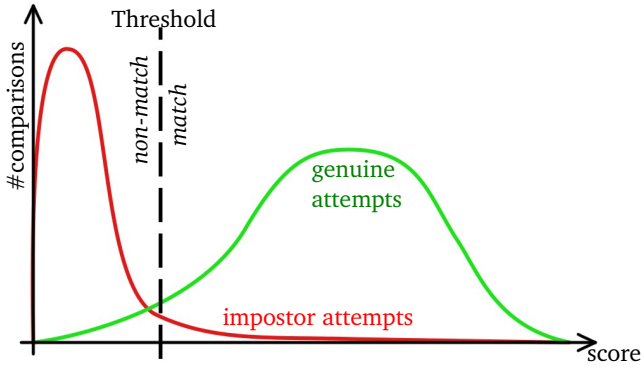


Illustration 1: Distribution of scores by attempt type.

As the score threshold increases, fewer impostor attempts will falsely be considered matches, but at the same time more genuine attempts will falsely be classified as non-matches.

PRECISE™ BIOMETRICS

Performance Evaluation

To evaluate how accurate a biometric system is, i.e. to measure its *biometric performance*, many genuine and impostor attempts are made with the system and all similarity scores are saved. By applying a varying score threshold to the similarity scores, pairs of FRR and FAR (or FNMR and FMR) can be calculated.

Results are presented either as such pairs, i.e. FRR at a certain level of FAR, or in plots (see below). Rates can be expressed in many ways, e.g. in percent (1%), as fractions (1/100), in decimal format (0,01) or by using powers of ten (10^{-2}). When comparing two systems, the more accurate one would show lower FRR at the same level of FAR.

Some systems don't report a similarity score, only the match/non-match decision. In that case it is only possible to gain a single FRR/FAR pair (and not a continuous series) as result of a performance evaluation. If the mode of operation (the security level) is adjustable (i.e. we have a means of controlling the internally used score threshold), the performance evaluation can be run again and again in different modes to obtain further FRR/FAR pairs.

There are two common ways of plotting performance evaluation results:

- DET graph (Detection Error Trade-off)
  plots FRR (Y-axis) vs. FAR (X-axis), i.e. false negative vs. false positive rate, often using logarithmic scale (at least for the FAR axis). As the Y-axis shows the number of match errors, the curve that is closest to the bottom of the plot corresponds to the best biometric performance.
- ROC graph (Receiver Operating Characteristic)
  plots true positive (1 - FRR) vs. false positive rate (FAR). Best biometric performance near the top of the plot.

DET curves are generally far better at highlighting areas of interest, the critical operating level, and are thus the most commonly used.

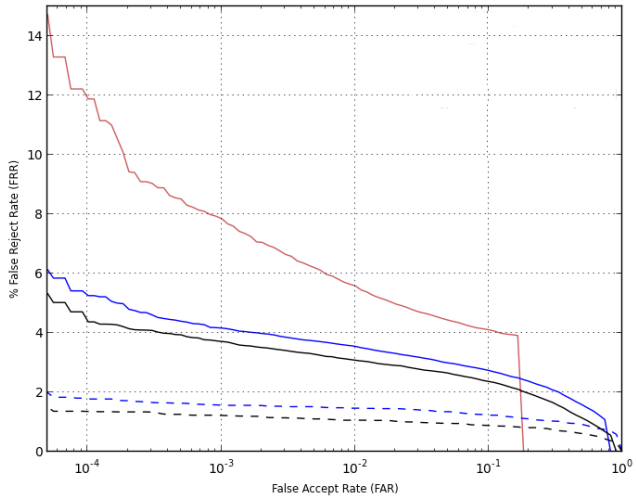


Illustration 2: Example DET graph

The top-most red curve corresponds to the worst biometric performance in this example. It shows an FRR of almost 8% at FAR 1/1,000.

Biometric performance evaluation is actually standardized. It's done jointly by **ISO/IEC** in the **19795**-series of standards. Even when doing less formal evaluations, the standard documents are a great source of information and help to avoid common pitfalls.

FMR – False Match Rate

Proportion of impostor attempts that are falsely declared to match a template of another object.

FNMR – False Non-Match Rate

Proportion of genuine attempts that are falsely declared not to match a template of the same object.

FTA – Failure-to-Acquire Rate

Proportion of the attempts for which the system fails to produce a sample of sufficient quality.

FAR – False Accept Rate and
FRR – False Reject Rate

Roughly the same as FMR and FNMR respectively, but the definition distinguishes between *attempts* and *transactions*. A transaction may consist of a sequence of attempts and depending on the system's configuration the outcome of individual attempts affects the transaction different.

FAR and FRR also takes the Failure-to-Acquire Rate into consideration. In case a transaction consists of exactly one attempt, FAR and FRR are calculated like this:

FAR = FMR * (1 - FTA) FRR = FTA + FNMR * (1 - FTA)

EER – Equal Error Rate

The point where the proportion of False Matches is the same as False Non-Matches (FNMR = FMR).

Understanding the figures

Going back to the example in the introduction, we can now explain what the figures mean:

FRR 1% @ FAR 1 / 10,000

means that...

When the system operates in a mode where

one out of ten thousand impostor attempts is (falsely) considered a match,

one per cent of the genuine attempts would fail (be falsely considered non-matches).

That seems rather straight forward, right?

However having answered this question, several new ones do pop up. For example:

1. Where did all the samples used for the comparisons come from?
2. Can the performance figures be directly compared with figures produced for other systems?
3. Are these figures really relevant for all conditions where the system can be used?
4. How reliable are these figures?

OK, some more details about biometric performance evaluation is certainly needed.

How evaluations are made

There are basically three types of performance evaluations: technology, scenario and operational evaluation.

When doing evaluation of biometric algorithms, technology evaluations are by far the most common and often most feasible. Since this type of evaluation is done using saved samples, the results are reproducible and doing an evaluation is not that time-consuming or complicated.

The great disadvantage with technology evaluations is that they do not necessarily reflect the conditions where the system will eventually be used. Because of this, it could be beneficial to collect a dedicated set of samples trying to mimic the conditions of the target system when preparing for an evaluation.

Technology evaluation

Evaluation using saved data, e.g. previously acquired fingerprint images.

Scenario evaluation

Evaluation of an end-to-end system using a prototype or simulated environment.

Operational evaluation

Evaluation in which the performance of a complete biometric system is determined in a specific application environment with a specific population.

Databases

The saved samples used in technology evaluations are collected in *databases*.

Data collection is done using a group of volunteers of which at least some provide multiple acquisitions of the same biometric modality (e.g. the same finger) to allow for genuine attempts. To make collection efficient, samples of multiple objects may be collected from each volunteer, e.g. all ten fingers.

The characteristics of the database have huge impact on the outcome of an evaluation. As stated before, except for the capabilities of the biometric algorithm, the amount of information available that could be used to characterize the objects being compared is what determines the biometric performance.

In databases the amount of usable information will vary depending on e.g.

- Equipment used to collect the data (e.g. type, area and quality of the fingerprint sensor)
- Environmental conditions (air temperature and humidity, ergonomics)
- Type of volunteers that provided samples (level of experience, age, occupation, sex, ethnicity)
- Data collection constraints (number of attempts allowed, time limits)
- Time between acquisitions (one user's acquisitions all done at the same time or weeks apart)

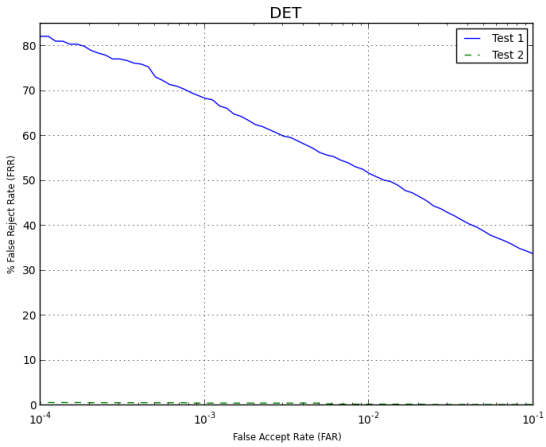


Illustration 3: Example: The same algorithm used with two different databases.

Test 1 resulted in more than 100 times as high False Reject Rate as Test 2 (dashed curve close to the X-axis) although run with the same algorithm. Different sensors and type of volunteers were used when collecting the two databases.

- If collection was supervised and what assistance volunteers were offered
- Whether it was a positive or negative use-case (users want to be recognized or not)

Due to all these factors **it is not possible to compare evaluations done using *different* databases**, as can be seen in Illustration 3. As there is no industry standard database to use, figures about claimed biometric performance can generally not be compared.

One way to overcome this would be to use a publicly available database that all vendors could utilize. Due to data protection legislation there are rather few available though. For fingerprint biometrics, the most commonly known are the ones produced for the Fingerprint Verification Competitions (FVC) organized by the University of Bologna. One concern about publicly available databases is that algorithm vendors may optimize against them and thus gain biased (better) results.

Confidence

To be able to make a statement about e.g. the FRR @ FAR = 1 / 1,000,000 *(that is the False Reject Rate at the mode of operation where only one out of one million impostor attempts are accepted)*, at least one million impostor attempts needs to be done. However if making a statement based on exactly one million impostor attempts, the FRR calculated would depend on one single accepted impostor attempt only. *(What is the similarity score of the 1,000,000 / 1,000,000 = 1 highest scoring impostor attempt and what proportion of the genuine attempts score lower.)* It is not hard to understand that the uncertainty in such a claim would be rather high – the outcome relies heavily on how the two most similar samples (of different origin) in the database happen to score.

When comparing figures or viewing a DET graph it is important to keep this in mind. In a DET graph the uncertainty is higher near the edges – what matters is if the significance is sufficient at the operational mode of interest.

The number of comparisons made is only one important factor that affects the confidence though. The key to gain better statistical significance is to make as many *uncorrelated* attempts as possible.

- It would be possible to do 1,000,000 impostor attempts using only two objects with 1,000 acquisitions of each. Since the 1,000 acquisitions of the same object would be very similar, corresponding similarity scores would have high correlation. Calculated FRR/FAR would vary a lot depending on how similar the two chosen objects happen to be and the figures would not tell much about what to expect from the system in wider use.
- On the other hand, to altogether avoid any kind of correlated data would require a huge group of volunteers. If only using each object for one impostor attempt, 2,000,000 objects would have to be sampled to gain 1,000,000 comparisons.

Often a compromise between these extremes is chosen:

- One sample of an object is compared to one sample each of several (often all) other objects.
- Different objects from the same person are not used for impostor attempts. E.g. two fingers from the same person are in general more alike (thus correlated) than two fingers from different persons.

The set of rules that decides what comparisons to actually make is called the *comparison scheme*. For the Fingerprint Verification Competition, whose databases contain 100 fingers with eight acquisitions of each, the comparison scheme is accordingly:

- All possible genuine comparisons are made (i.e. all eight attempts of each finger vs. all remaining seven of the same finger, but the same pair of samples are only compared once: A-B but not B-A, thus "/2" below). This results in 100 * 8 * 7 / 2 = 2,800 genuine attempts.
- The first acquisition of each finger is compared against the first acquisition of all other fingers (again the same pair of samples are only compared once). This leads to 100 * 99 / 2 = 4,950 impostor attempts.

To help estimating the uncertainty in observed FRR/FAR values, there are methods to calculate confidence intervals based on the number of errors observed. Such methods are e.g. found in the ISO/IEC 19795-1 standard.

Manipulating evaluations

The errors that are observed in performance evaluation (False Rejects and False Accepts) are typically not evenly distributed between volunteers. Some people have biometric features that are either hard to capture or in themselves show few unique characteristics. For fingerprints, examples of the former are

- worn friction ridges (e.g. due to manual labor),
- dry skin (e.g. due to cold weather), and
- skin disease.

Examples of the latter are

- few minutiae in the ridge pattern, and
- inability to present proper part of the finger, e.g. only showing the top of fingers that contain fewer characteristic structures.

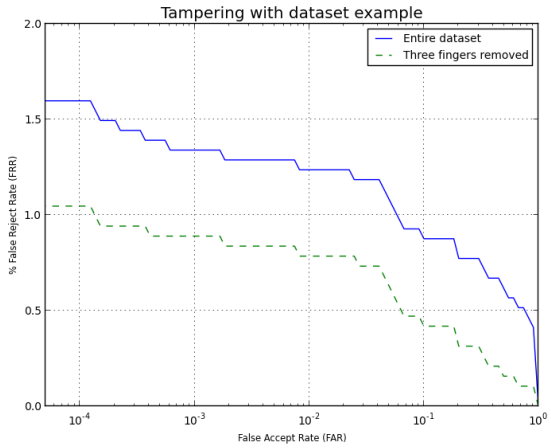


Illustration 4: Example: Tampering with data

Here only 3 out of a total of about 550 fingers were removed from a database after analyzing what genuine comparisons gained lowest similarity scores. At FAR of 1/1,000 the original (untampered) database would show about 50% higher FRR!

By selectively removing the worst samples of a database, the calculated biometric performance can easily be hugely improved. Thus **results that claim performance**

- **obtained from a *subset* of a database,** or alternatively
- **not presenting the Failure-to-Acquire Rate** or means used to sort out low-quality samples

are more or less useless. See the example in Illustration 4!

To avoid the risk of manipulation, it's better to self evaluate a system than to depend on fabricated numbers. As an alternative, evaluations done by independent entities could be trusted, but then the test environment and conditions are determined by someone else. Examples of such evaluations:

- "MINEX II" and "Ongoing MINEX" run by NIST (http://www.nist.gov/itl/iad/ig/minexii.cfm)
- "FVC-onGoing" (https://biolab.csr.unibo.it/fvcongoing/)

Conclusions

Biometric performance evaluation is done by performing many genuine and impostor comparisons and analyzing produced similarity scores or match decisions.

Error rates are calculated as the proportion of

- impostor attempts that are falsely accepted (FAR), and
- genuine attempts that are falsely rejected (FRR).

FRR at fixed levels of FAR are used to compare systems with each other or to determine if a system has sufficient accuracy for a specific use case.

Samples used in technology evaluations are stored in databases. The characteristics of a database have huge impact on the achieved biometric performance. Because of this

- evaluation results obtained on different databases cannot be compared, and
- the better the collected samples mimic the target environment, e.g. in terms of sensor used, types of volunteers and physical conditions, the more reliable predictions can be made from the evaluation.

Investigated error rates are often very low (e.g. FAR = 1/1,000,000) so it's important to consider the statistical confidence in the figures calculated.

Very few people have difficulties using biometric systems, so these few people have huge impact on estimated biometric performance. It's easy to tamper results by omitting them.

PRECISE™ BIOMETRICS