Biometric Systems

Script from the Course

Christoph Busch

January 19, 2021

Contents

1 Biometric Performance

1.1 Overview

- Metrics (FTC,FTX,FTE,FMR, FNMR, FAR, FRR, ROC, DET)
- confidence of measured error rates

1.2 Basic Metrics

Biometric performance is based on measurements conducted on a set of samples, which are available in a corpus or gallery. This corpus may contain a set of image samples, voice samples or other representation of a biometric characteristic. The fundamental metrics in a biometric system is the comparison score $c(Q, R)$ that is generated, when the algorithm under test compares a biometric probe or query Q with a biometric reference R. According to the ISO/IEC Harmonized Biometric Vocabulary [2] a comparison score is defined as:

Definition 1 (comparison score) *numerical value (or set of values) resulting from a comparison [2].*

Note that the term *matching score* is deprecated by ISO/IEC 2382-37 [2]. The comparison score can be expressed by a similarity score $s(Q, R)$ or distance score $d(Q, R)$

Definition 2 (similarity score) *comparison score that increases with similarity [2].*

Definition 3 (distance score / similarity score) *comparison score that decreases with similarity [2].*

For the distance score d, we expect the following properties to hold.

$$d(Q, R) \geq 0 \tag{1.1}$$

and

$$d(R, R) = 0 \tag{1.2}$$

The later can be considered as a self-comparison, e.g. two templates potentially derived with two different algorithms from one and the same sample should have a distance of zero or at least close to zero. The distance score d can be converted to a similarity score s using a monotonically decreasing function f. Examples for common conversions are the following

$$s = -d \tag{1.3}$$

$$s = -log(d) \tag{1.4}$$

$$s = \frac{1}{d} \tag{1.5}$$

For the case that the feature vectors are represented in binary form then the dissimilarity between the two vectors is expressed by *Hamming Distance* that essentially counts the number of different bits.

1.2.1 Distance Score

For a n-dimensional metric space an example for a distance metric is the *P-norm*, which is also known as *Minkowski metric*. In Figure 1.1 a two dimensional feature space with a probe feature vector Q and a reference feature vector R are illustrated. For the given

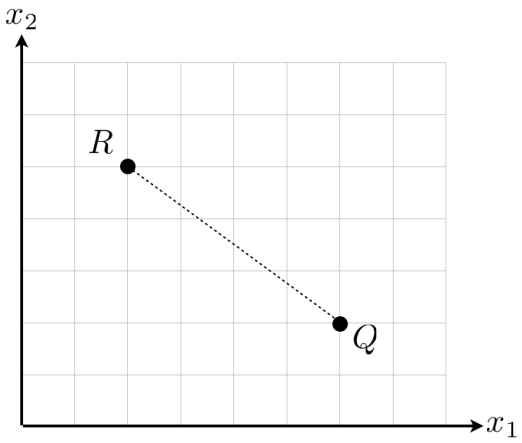

Figure 1.1: 2D feature space with probe Q and reference R

example the two vectors are represented by

$$Q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \quad \text{and } R = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

As distance measure between those two points we express the length of the difference vector X as P-norm

$$||X||_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \tag{1.6}$$

where n is the dimensionality of the feature space and p real number and $p \geq 1$.

City Block Norm For $p = 1$ the P-Norm simplifies to the the city block norm that is also called taxicab norm.

$$||X||_1 := \sum_{i=1}^{n} |x_i| \tag{1.7}$$

For the example from Figure 1.1 we derive

$$\begin{aligned}\sum_{i=1}^{n}|x_i| = \sum_{i=1}^{2}|q_i - r_i| &= |q_1 - r_1| + |q_2 - r_2| \\ &= |6-2| + |5-2| \\ &= 4 + 3 \\ &= 7\end{aligned}$$

Euclidian Nom For $p = 2$ the P-Norm simplifies to the Euclidian norm.

$$||X||_2 := \sqrt{\sum_{i=1}^{n}|x_i|^2} \tag{1.8}$$

For the example from Figure 1.1 we derive

$$\begin{aligned}\sqrt{\sum_{i=1}^{n}|x_i|^2} = \sqrt{\sum_{i=1}^{2}|q_i - r_i|} &= \sqrt{|q_1 - r_1|^2 + |q_2 - r_2|^2} \\ &= \sqrt{|6-2|^2 + |5-2|^2} \\ &= \sqrt{4^2 + 3^2} \\ &= \sqrt{25} \\ &= 5\end{aligned}$$

Maximum Norm For $p = \infty$ we get the infinity norm or maximum norm.

$$||X||_\infty := \sqrt{\sum_{i=1}^{n}|x_i|^\infty} = max\left(|x_1|, |x_2|, \cdots, |x_n|\right) \tag{1.9}$$

For the example from Figure 1.1 we derive

$$\begin{aligned}max\left(|x_1|, |x_2|, \cdots, |x_n|\right) = max\left(|x_1|, |x_2|\right) &= max\left(|q_1 - r_1|, |q_2 - r_2|\right) \\ &= max\left(|6-2|, |5-2|\right) \\ &= max\left(4, 3\right) \\ &= 4\end{aligned}$$

1.3 Biometric Failures

There are multiple failure associated with a acquisition of a biometric sample or with its processing. In Sections 1.3.1 to 1.3.3 we will discuss the failures that are associated with the deficiency of a biometric system to create a biometric reference for a data subject and subsequently in Sections 1.6 to 1.7 will consider errors that are attributed to biometric verification and identification systems.

1.3.1 Failure-to-Capture

A Failure-to-Capture Rate (FTC) is constituted, when the capture process could not generate a biometric sample of sufficient quality. This can be caused due to one of the following reasons:

1. The sample is not generated, as the characteristic is not placed properly on the capture device (e.g finger not covering the sensor area)
2. The captured signal is rejected by the automatic sample quality control algorithm.
3. The captured signal is stored as file, but rejected by the operator (staff expert) subsequent to visual inspection as it is not of sufficient quality

The ISO-definition [2] for the FTC is given by:

Failure-to-Capture Rate: *proportion of failures of the biometric capture process to produce a captured biometric sample that is acceptable for use.*

To estimate the FTC we use the following formula:

$$FTC = \frac{N_{tca} + N_{nsq}}{N_{tot}} \tag{1.10}$$

where N_{tca} is the number of terminated capture attempts, N_{nsq} is the number of images created with insufficient sample quality and N_{tot} is the total number of capture attempts. In consequence of a Failure-to-Capture are new capture attempt is initiated. This is illustrated in figure 1.2 .

1.3.2 Failure-to-eXtract

A Failure-to-eXtract is constituted, when the feature extraction process was not able to generate a biometric template. This can be caused due to one of the following reasons:

1. The algorithm itself declares that it cannot create a template from the input sample. This could be caused by a insufficient number of features that were identified e.g. only five minutia could be extracted from a fingerprint image.
2. Processing time of feature extraction algorithm exceeds the specified limit and thus the feature extraction is terminated
3. The feature extraction algorithm might suddenly crash during processing. In this case, some actions will be undertaken (e.g. start over application, repeat process, etc.) but if the crash happens all the time with the same sample then for this image a failure to extract feature will be constituted. There is currently no ISO-definition for the Failure-to-eXtract Rate.

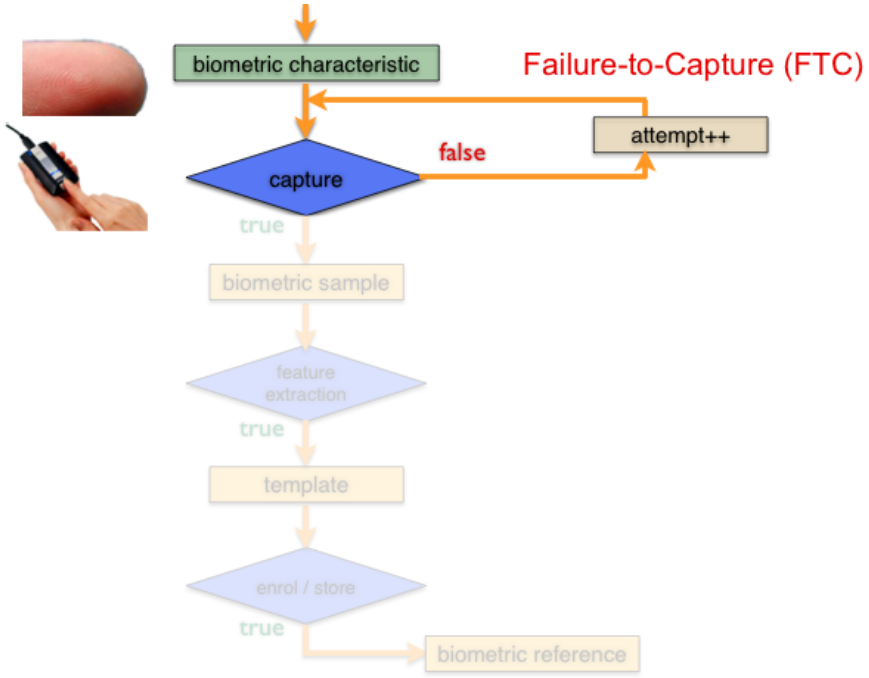

Figure 1.2: Failure-to-Capture (FTC)

To estimate the Failure-to-eXtract Rate (FTX) we use the following formula:

$$FTX = \frac{N_{ngt}}{N_{sub}} \tag{1.11}$$

where N_{ngt} is the number of cases, where no template was generated and and N_{sub} is the total number of biometric samples being submitted to the feature extraction component (i.e. the template generator). In an operational scenario the consequence of a Failure-to-eXtract is a new attempt including a new biometric sample creation and it subsequent processing. This is illustrated in figure 1.3 .

1.3.3 Failure-to-Enrol

A Failure-to-Enrol is constituted, when the biometric system is not capable to create for data subject a biometric reference. Thus the Failure-to-Enrol Rate (FTE) expresses the proportion of the population, for which the system fails to complete the enrolment process. This can be caused due to one of the following reasons:

1. The biometric characteristic of the subject (e.g. its fingerprint images) can not be captured at all.

2. For each evaluation setting, and if required instances of the same characteristic (e.g. left index finger instead right index finger) it is not possible to create for this

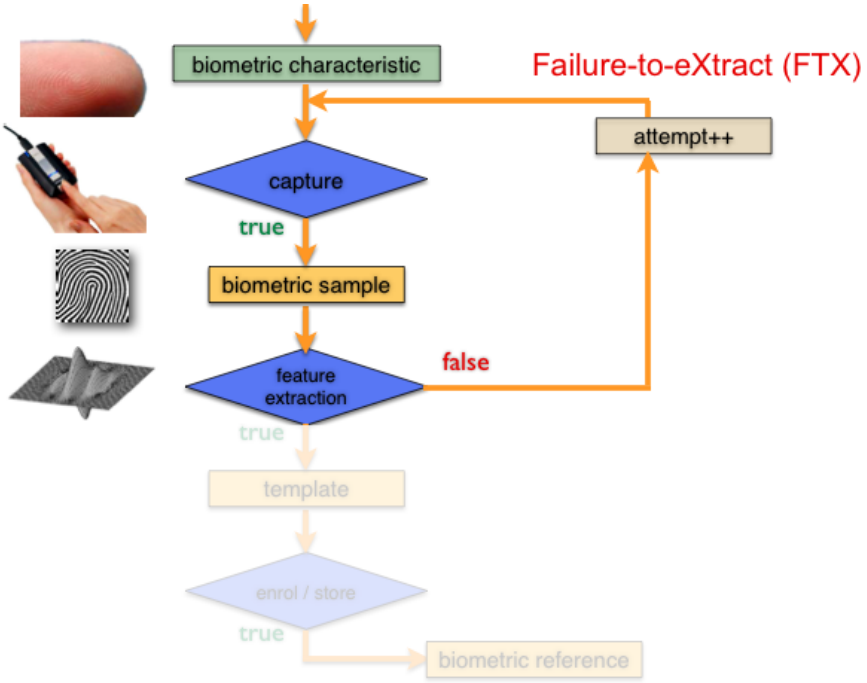

Figure 1.3: Failure-to-eXtract (FTX)

subject a template of sufficient quality (e.g. a feature set with minimum number of minutia)

There are currently two ISO-definitions for the FTE. The original definition in the performance testing standard [1] and the more recent one from the harmonized biometric vocabulary [2]:

Failure-to-Enol Rate (ISO/IEC 19795-1): *proportion of the population for whom the system fails to complete the enrolment process.*

Failure-to-Enrol Rate (ISO/IEC 2382-37): *proportion of biometric enrolment (that did not fail for non-biometric reasons), that resulted in a failure to create and store an enrolment data record for an eligible biometric capture subject, in accordance with an enrolment policy.*

To estimate the FTE we use the following formula:

$$FTE = \frac{N_{nec}}{N} \tag{1.12}$$

where N_{nec} is the number of cases, where we meet one of the two Failure-to-Enrol criteria and N is the total number of subjects, intended to be enroled in the biometric application. The consequence of a Failure-to-Enrol In an operational scenario is that for the capture subject a fallback procedure must be activated that should treat the

individual in a non-discriminatory manner. The Failure-to-Enrol is illustrated in figure 1.4 .

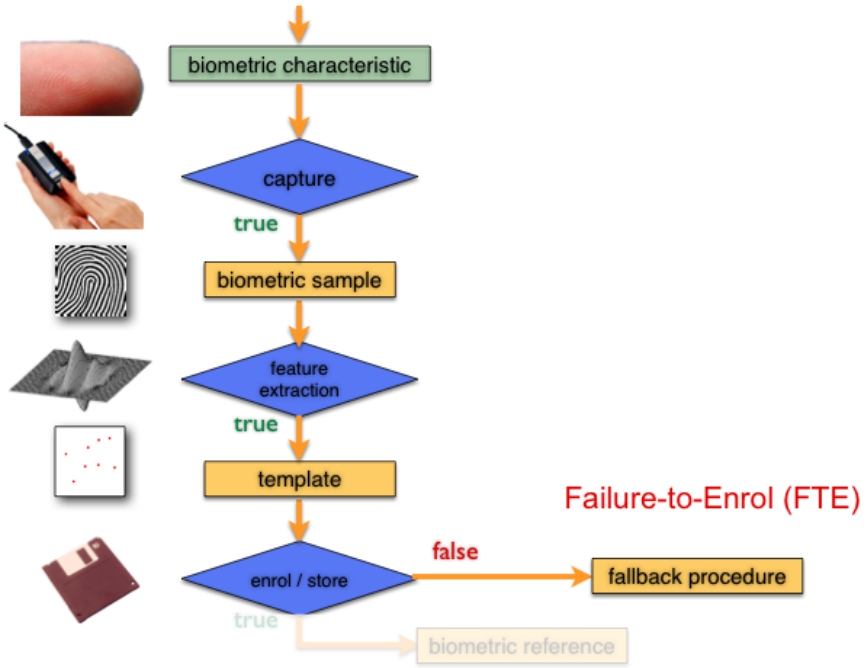

Figure 1.4: Failure-to-Enrol (FTE)

1.3.4 Failure-to-Acquire

The Failure-to-Acquire Rate (FTA) is essential for the verification process and estimates the likelihood that biometric comparison can not be completed due to potential deficiencies in the live sample that is submitted as a probe. If there is no feature vector that can be compared to a biometric reference this can be caused due to one of the following reasons:

1. The is no biometric sample generated, which is expressed by the FTC.
2. The feature extraction componen failed to extract features as the number and/or quality of extracted features is not sufficient. This is expressed by the FTX.

There are currently two ISO-definitions for the FTA. The original definition in the performance testing standard [1] and the more recent one from the harmonized biometric vocabulary [2]:

Failure-to-Acquire Rate (ISO/IEC 19795-1): *proportion of verification or identification attempts for which the system fails to capture or locate an image or signal of sufficient quality.*

Failure-to-Acquire Rate (ISO/IEC 2382-37): *proportion of a specified set of probe acquisitions that failed to create a biometric probe.*

Note that in ISO/IEC 2382-37 a *probe* is defined as *biometric data input to an algorithm for comparison to a biometric reference(s).*To estimate the Failure-to-Acquire Rate we use the following formula:

$$FTA = FTC + FTX * (1 - FTC) \tag{1.13}$$

1.4 Performance measures

Given a corpus of biometric samples the following measures provide insight in the recognition accuracy of a feature extractor and biometric comparison subsystem. When computing the measures it should be considered to achieve independent trials. This is relevant, when for example multiple instances of a biometric characteristic are captured (e.g. 10 fingerprints or 2 eyes per data subject). According to Clause 7.6.1.3 of ISO/IEC 19795-1 the evaluator should respect that within-individual comparisons are not equivalent to between-individual comparisons, and shall thus not be included in the set of impostor transactions [1].

1.4.1 False-Match

For impostor comparisons a False-Match constitutes the undesired case that an impostor probe is matching a biometric reference, which has not been created for himself. There are currently two ISO-definitions for the corresponding False-Match-Rate (FMR). The original definition in the performance testing standard [1] and the more recent one from the harmonized biometric vocabulary [2]:

False-Match-Rate (ISO/IEC 19795-1): *proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template.*

False-Match-Rate (ISO/IEC 2382-37): *proportion of the completed biometric non-mated comparison trials that result in a false match.*

$$FMR(t) = \int_t^1 \Phi_i(s)ds \tag{1.14}$$

Together with the False-Non-Match-Rate (FNMR) the FMR is the key metric to be used in biometric technology testing and is understood to characterize a security property of a biometric system. Note that some literature is using the term False-Accept-Rate in the meaning of FMR.

1.4.2 False-Non-Match

For genuine comparisons a False-Non-Match constitutes the undesired case that an genuine probe is not matching to biometric reference, which has been created for the same subject from the same source (e.g. same index finger). There are currently two ISO-definitions for the corresponding False-Non-Match-Rate (FNMR). The original definition in the performance testing standard [1] and the more recent one from the harmonized biometric vocabulary [2]:

False-Non-Match-Rate (ISO/IEC 19795-1): *proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same data subject supplying the sample.*

False-Non-Match-Rate ISO/IEC 2382-37): *proportion of the completed biometric mated comparison trials that result in a false non-match.*

$$FNMR(t) = \int_0^t \Phi_g(s)ds \tag{1.15}$$

Note that in computing the FNMR we will count the Genuine-Match-Rate (GMR) first. The relationship between GMR and FNMR is as follows:

$$FNMR(t) = 1 - GMR(t) \tag{1.16}$$

Together with the False-Match-Rate (FMR) the FNMR is the key metric to be used in biometric technology testing and is understood to characterize a convenience property of a biometric system. Note that some literature is using the term False-Reject-Rate in the meaning of FNMR.

1.4.3 Similarity Matrix

In order to compute the above performance measures we need first to analyze the similarity scores achieved by our comparison subsystem. Figure 1.5 indicates the similarity scores for a face recognition subsystem for 3 subjects and 3 instances. The gallery was capture in 2 session, such that we have one enrolment sample and one probe sample per subject.

Similarity scores in green represent genuine scores (i.e. stemming from mated comparison trials) and similarity scores in red represent impostor scores (i.e. stemming from non-mated comparison trials).

If, while considering the recommendations regarding independent trials, multiple instances per subject are observed and evaluated, then the similarity matrix becomes more complex. In Figure 1.6 we can observe the similarity matrix for an evaluation, where we have multiple instances per subject and in addition multiple samples per instance captured. The number of similarity scores and impostor scores is significantly increasing.

In this case we have N instances (e.g. 10 fingers) and U samples captured per instance, which should stem from session, which are separated by at least one week. Figure 1.6

Figure 1.5: Similarity matrix for 3 subjects and 3 instances

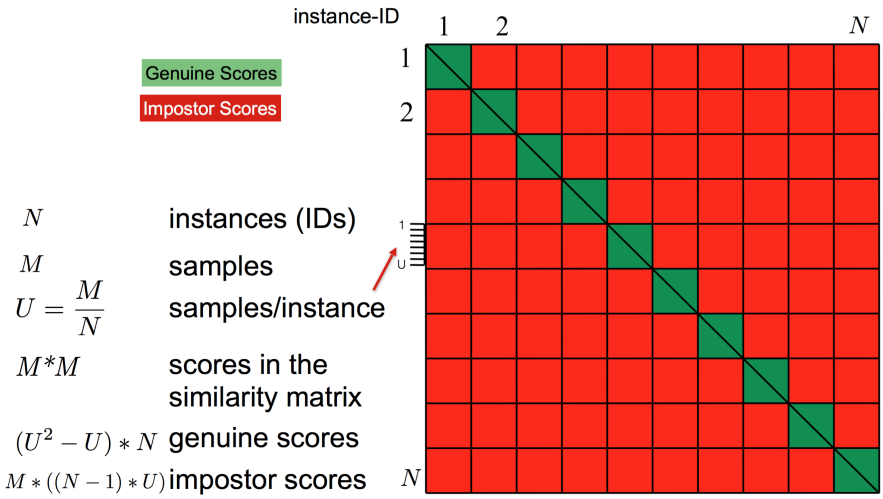

Figure 1.6: Similarity matrix for 3 subjects and 3 instances

indicates that in total we have $N*U=M$ samples, which can be used as either reference or probe sample. The size of the similarity matrix is $M*M$. For such a similarity matrix it becomes obvious that the number of impostor scores is significantly larger than the number of genuine scores. Counting the number of genuine scores, we have U samples (per instance) but from the possible U^2 we have to subtract U self-comparisons. Counting the number of impostor scores, we have $(M-1)*M$ comparison scores to consider.

A concrete example for the size of the similarity matrix: For a fingerprint system evaluation we invite 150 data subjects and record all 10 instances in 12 sessions. Our similarity matrix will then be of size

- Number of data subjects S: 150
- Number of biometric instances per subject: 10

- Total number of instances N: $1.500 = S * 10$
- Number of samples per instance U: 12
- Total number of samples M: $18.000 = 12 * 1.500 = U * N$
- Number of genuine scores per instance: $132 = 12^2 - 12 = U^2 - U$
- Total number of genuine scores: $198.000 = (12^2 - 12) * 1.500 = (U^2 - U) * N$
- Total number of impostor scores: $323.982.000 = 18.000 * 149 * 12 = M * (N-1) * U))$

As we realize the total number of impostor scores is by two magnitudes larger than the number of genuine scores. Thus in order to avoid bias towards impostor scores it is common practice to reduce the impostor scores to corresponding instances from different subjects. For example one would include in the non-mated comparison trials only those samples, that are captured from the equivalent finger types (e.g. from the left index finger).

On the way to progress from comparison scores to performance measures we define for a given corpus of biometric samples as follows:

- Ω_g: set of all genuine scores
- Ω_i: set of all impostor scores
- $\Omega_g(t)$: set of all genuine scores $> t$
- $\Omega_i(t)$: set of all impostor scores $> t$
- $|| \Omega ||$: number of elements in Ω

Then we can compute

$$FMR(t) = \frac{\Omega_i(t)}{\Omega_i} \tag{1.17}$$

$$GMR(t) = \frac{\Omega_g(t)}{\Omega_g} \tag{1.18}$$

and then

$$FNMR(t) = 1 - GMR(t) \tag{1.19}$$

1.4.4 Example

Now we apply the measures from the previous section on the example, which was introduced in Figure1.5.

First in Figure 1.8 we set the threshold t to 0,66.

We can observe that 2 impostor scores exceed the threshold. Thus we compute according to equation 1.17 the $FMR(0,66) = 2/6$ and according to equation 1.18 the $FNMR(0,66) = 1/3 = 1 - 2/3$

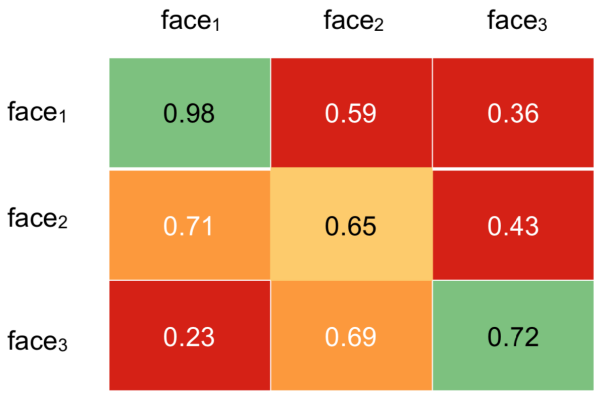

Figure 1.7: Performance metrics for threshold t=0,66

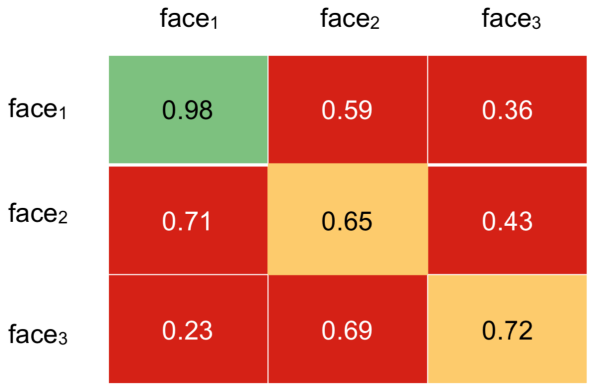

Figure 1.8: Performance metrics for threshold t=0,73

Next we modify the threshold and evaluate the system for $t = 0,73$.

Now we can observe that no impostor scores is exceeding the threshold. But as a consequence of the modified threshold now tow genuine comparison trials are not successful any longer. We compute the $FMR(0,66) = 0$ and the $FNMR(0,66) = 2/3 = 1 - 1/3$. We can clearly identify the dependency between FMR and FNMR.

1.5 Reporting

In order to benchmark various algorithms we use a graphical representation plotting the GMR/FNRM in relationship to the FMR for different thresholds t. In Figure 1.9 we observe the Receiver Operating Characteristic (ROC) and most recommended graphical reporting in Figure 1.10 the Detection Error Trade-Off (DET) Curve.

The benchmark of various algorithms (biometric comparison subsystems) is illustrated in Figure 1.11. We can say that technically speaking that system performs best that

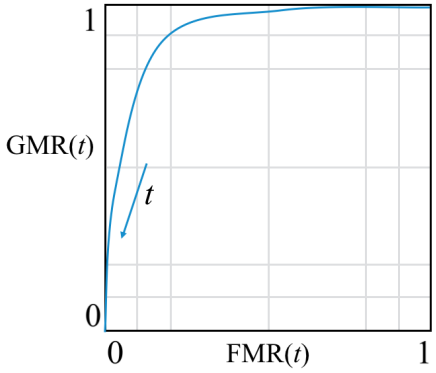

Figure 1.9: Receiver Operating Characteristic (ROC)

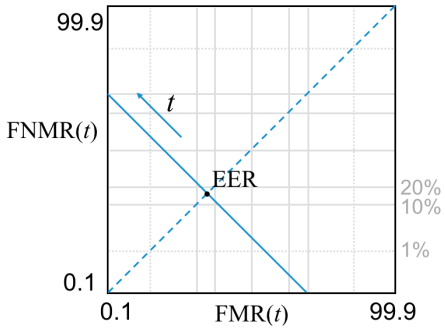

Figure 1.10: Detection Error Trade-Off (DET) Curve

shows the lowest area under curve. Note that this is not System D.

1.6 Verification System Performance

The first order estimation of the performance for a verification system that is based on transactions allowing multiple attempts can be derived from the detection error trade-off curve. However if this is applied the potential correlations between the attempts are neglected. Such correlations could be due to habituation of the capture subject with the human- computer interface of the biometric system. The relevant measures for a verification system are the False-Accept-Rate (FAR) and the False-Reject-Rate (FRR). The ISO-definition [1] for both metrics are the following:

False-Accept-Rate (ISO/IEC 19795-1): *proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed.*

False-Reject-Rate (ISO/IEC 19795-1): *proportion of verification transactions with truthful claims of identity that are incorrectly denied.*

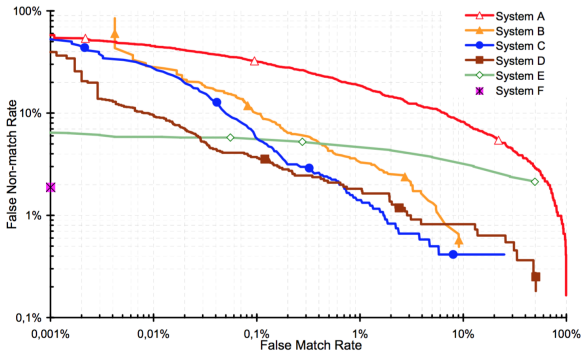

Figure 1.11: Detection Error Trade-Off (DET) Curve

For the simplified case that the verification system does allow only a single attempt per transaction then the FAR and FRR can be estimated as follows.

$$FAR = FMR * (1 - FTA) \tag{1.20}$$

and

$$FRR = FTA + FNMR * (1 - FTA) \tag{1.21}$$

If the biometric application is likely to be confronted with a large number of failure to enrol cases (e.g. as it is a fingerprint system for mine workers) and the biometric performance shall be predicted based on a gallery that was collected for a technology testing then the equations 1.20 and 1.21 do not sufficiently express the performance to be expected. The reason for this is that in a technology evaluation biometric references are generated from the gallery that do not cause a failure-to-enrol and probes that do not cause a failure-to-acquire. For such a case the generalized versions of the above equations are more appropriate, which are given by:

$$GFAR = FMR * (1 - FTA) * (1 - FTE) \tag{1.22}$$

and

$$GFRR = FTE + (1 - FTE) * FTA + (1 - FTE) * (1 - FTA) * FNMR \tag{1.23}$$

1.7 Identification System Performance

The first order estimation of the false positive and false negative identification rates for open-set systems, can be derived from FMR and FNMR and the DET curve. However, such estimates cannot take account of correlations in the comparisons involving the same data subject, and consequently can be quite inaccurate [1].

$$FPIR = (1 - FTA) * (1 - (1 - FMR)^N) \tag{1.24}$$

where $FPIR$ is the False-Positive-Identification-Rate. For a small FMR we can substitute in equation 1.24

$$(1 - FMR)^N \approx 1 - N * FMR \tag{1.25}$$

and thus under the assumption of $FTA = 0$ we derive

$$FPIR = (1 - 0) * (1 - (1 - N * FMR) \tag{1.26}$$

$$FPIR = N * FMR \tag{1.27}$$

This is comparable to the scenario that has been conducted at the Mainz as well as Berlin Suedkreuz trainstation

1.8 Testing Standards

Test procedures as such are well known since the biometric performance testing standards ISO/IEC 19795-1 was established in 2006 [1]. That framework for Biometric Performance Testing and Reporting was developed on the basis of established concepts such as the *Best Practices in Testing and Reporting Performance of Biometric Devices*[4] and it defines in which way algorithm errors such as false-match-rate (FMR) and false-non-match-rate (FNMR) as well as system errors such as false-accept-rate (FAR) and false-reject-rates (FRR) must be reported.

1.9 Reading

Complementary reading for performance testing is the ISO standard on Biometric Performance Testing [1]. Details on the rule of 3 are given in the paper of Jovanovic and Levy [3].

Bibliography

[1] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, March 2006.

[2] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 2382-37:2012 Information Technology - Vocabulary - Part 37: Biometrics*. International Organization for Standardization, 2012.

[3] Jovanovic, B., and Levy, P. A look at the rule of three. *The American Statistican* (1997), 137–139.

[4] Mansfield, T., and Wayman, J. Best practices in testing and reporting performance of biometric devices. CMSC 14/02 Version 2.01, NPL, August 2002.