Information Security Technical Report, Vol. 3, No. 1 (1998) 70-76

Measuring and Comparing the Performance of Biometric Systems

By Tony Mansfield and Gavin Kelly, National Physical Laboratory

This paper describes some of the issues in measuring the performance of biometric systems. The aim is to explain what the often quoted measures such as the False Acceptance Rate and False Rejection Rate are actually measuring, and why performance in real applications can be significantly at odds with the manufacturer's quoted rates.

Introduction

- What advantages can biometric systems offer over conventional systems for automatic identification or verification of user identity?

- Which is the best biometric system for a given application?

Such questions cannot be answered without some way to measure and compare the performance of biometric systems. However, objective measurement of performance of biometric systems, allowing meaningful comparisons between devices, is made difficult by a number of factors.

1. The performance of a biometric system can depend heavily on the type of application. For example if the end-users are familiar with the system, willing to use it, and if their use of the system is being supervised then one would clearly expect performance to be better than with unsupervised, unwilling, untrained end-users.

2. Measures that are applicable to some biometric devices are meaningless with others. For example, in the case of behavioural biometric systems (such as signature, or voice), ease of forgery is important. However, there is not a direct analogy for physiological biometric systems.

3. There are trade-offs between the various performance measures. By relaxing the acceptance criteria, the false rejection rate can be improved at the cost of increasing the false acceptance rate. Allowing multiple attempts can also decrease the false rejection rate, but will worsen the throughput rate. Devices can be set to give the best possible performance for one particular application, but will then be less than optimal in different circumstances.

4. Manufacturer's quoted performance figures, obtained from in-house laboratory tests, are often not easy to relate to real life performance.

5. Moreover, for many of the possible measurements, there are different interpretations of how to make the measurement, how to present the results and what the results mean.

There are a number of initiatives working towards improving this situation and enabling authoritative performance measurement of biometric systems. Within Europe, the BIOTEST project has developed a methodology for measuring the key performance aspects of biometric systems, in a collaborative venture involving several European companies including manufacturers, users and evaluators. In America, methods and standards for testing

are being developed by National Biometric Test Centre; and the International Computer Security Association has launched a certification scheme for biometric products. Further details of these initiatives are given towards the end of this paper.

What to measure — What to compare

Biometric systems can be used for identifying an individual from all those enrolled in the system, or for verifying a claimed identity. In this paper we assume that the application is one of user verification. This is for convenience, to save repeated reference to both possible modes of operation. However, the problems and solutions mentioned are, by and large, equally applicable to both cases.

To select an appropriate biometric system, it is likely that we would wish to consider various aspects of performance:

- How accurate will the system be at verifying a claimed identity?
- Will end-users find the system easy to use, and will they be happy to use it?
- Will the system be fast enough in operation?
- Is the system secure enough to protect against attempted fraudulent use?
- In addition there are the usual considerations regarding cost, interfaces, capacity, etc.

The performance figures of most interest are those that tell how the whole system will operate in practice. If laboratory measurements of the systems' biometric components are to be used for this purpose, full account must be taken of the context of use. The type of end-users (employees or general public) must be considered along with how well they have been enrolled and trained in the system, (supervised by an expert in the system or self-enrolment), and all other external factors such as time-pressures, operating environment, etc.

To compare different systems, again it is best to use operational performance measures for the entire system, as different devices are weak or strong on different aspects. (A possible exception to this is where almost all system components are the same, the same underlying biometric technology is being used, and only the component capturing the biometric, or the biometric algorithm differs.)

The drawback to using operational performance metrics, obtained on the application under consideration, is that these can vary considerably between applications. Laboratory measurements, under controlled conditions, can provide reference performance metrics, that are not application dependant, are repeatable, and that can be independently verified.

In the next sections we shall discuss:

1. The main factors that cause differences between the operational and reference figures for the false rejection rate.

2. How the reference false acceptance rate has little bearing on the operational figure, and is only one factor in assessing the overall security of the system.

False Rejection Rate

The False Rejection Rate is defined as the probability that the biometric system will fail to verify the legitimately claimed identity of an enrolee.

Manufacturers will normally quote a false rejection rate obtained from in-house laboratory tests. In this section we consider how this reference false rejection rate differs from the operational false rejection rate that is encountered in the field on a selected application. This gives an indication of the other factors that need to be taken into account to predict operational performance from the reference figures.

Reference False Rejection Rate

Laboratory tests for the False Rejection Rate usually estimate the value as

$$\text{False Rejection Rate} = \frac{\text{Number of False Rejections}}{\text{Number of Enrolee Verification Attempts}}$$

The verification attempts are made in a controlled environment, in which problems and variability in capturing the biometric sample are minimized.

Since the value for the False Rejection Rate is an estimate, and its accuracy depends on the number and quality of verification attempts, confidence intervals should be provided.

The value of the False Acceptance Rate at the same accept/reject decision threshold should be given, as this threshold can be adjusted to lower the false rejection rate by allowing an increase in false acceptances. It should also be clear, in the case of devices that allow more than one attempt at verification, whether the figure gives a one-attempt or several-attempt rejection rate.

To interpret what the result is really telling us, it is useful to know the conditions under which enrolment and verification take place. Details are needed of the types of user, the operating environment, the level of supervision during enrolment and at subsequent verification attempts, etc. Quoted performance figures typically omit this information, in which case the most likely scenario is that the users are trained in use of the device, are supervised in enrolment and verification, and the tests are conducted in a laboratory environment; i.e. ideal conditions for a minimal number of false rejections.

It should also be noted that there is a difference between a few individuals making many verification attempts, as opposed to many individuals making few verification attempts.

A further issue is what this rate really means. A False Rejection Rate of 1% can mean that any user has a 1% chance of being rejected, or could mean that 1% of users are always rejected and the other 99% are never rejected. Experience seems to show that with most (if not all) biometrics, the chance of a false rejection is not spread evenly among users. For example, some signatures exhibit more variability than others, some fingerprints are finer than others or have shallower valleys making recognition harder, a wearer of glasses might be more susceptible to false rejections on an iris recognition device, etc.

Operational False Rejection Rate

To obtain an objective measure of the operational false rejection rate we must consider all the stages of using the device, and measure how likely a failure causing rejection is at each stage. Many of these potential failures are ignored in the laboratory calculation of the false rejection rate.

Use of a biometric system requires two processes, as shown in *Figure 1*: first enrolment and subsequently verification. In both processes, the first stage is for the end-user to present their biometric feature (e.g. fingerprint) to the system, and an a biometric sample (e.g. fingerprint image) is captured. The second stage converts this biometric

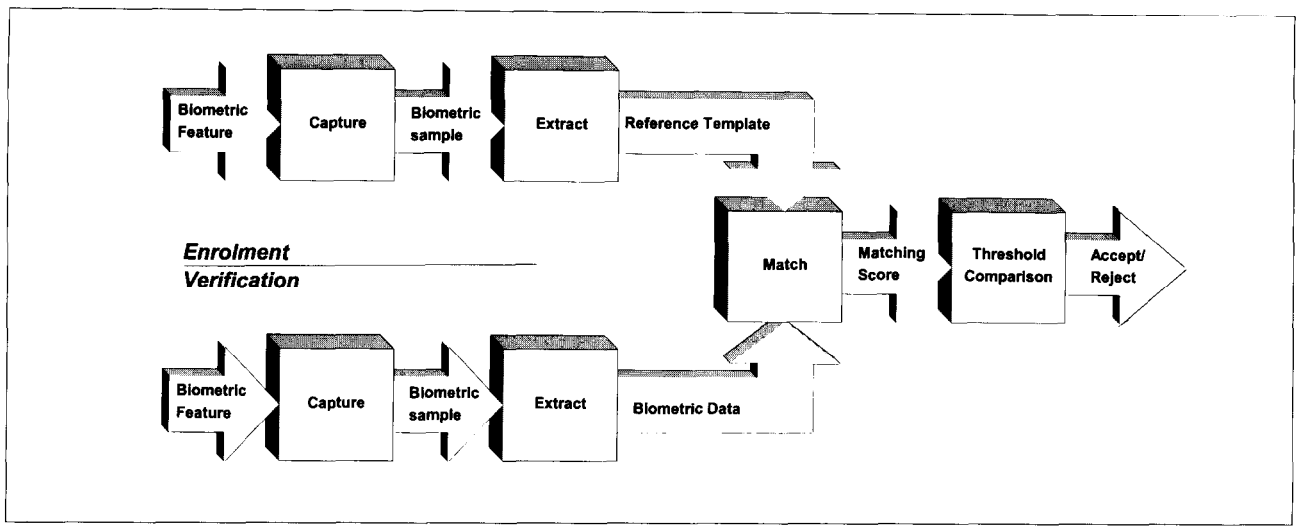

Figure 1: Stages in use of a biometric device.

sample to biometric data (e.g. minutiae coordinates) for matching. The final stage of the enrolment process is to form the reference template for the individual; in the case of verification the final stage is to compare the biometric data with the reference template. We consider the possible faults that can occur at each stage that will cause a false rejection.

Presentation of biometric

The end-user may have difficulty in presenting the system with the required biometric feature. This could be due to the positioning of the device, for example the device may be too high for wheelchair users; or it could be that the user does not have the appropriate biometric feature, for example due to amputation.

For systems that use a card or PIN together with the biometric, there is also the possibility that the end-user has forgotten their card, forgotten their PIN, or that the card is damaged in some way, thus making the transaction impossible.

Extracting biometric data from the biometric sample

Even if the end-user can present a biometric sample to the device, it is quite possible that the system fails to convert the biometric sample to biometric data which can then be matched. This situation is normally called a failure to acquire.

An example is a fingerprint system failing to recognize the presence of a finger because the fingertip is not making sufficiently good contact on the device, or because the finger is sufficiently misplaced that the system cannot locate the reference points needed for extracting the biometric data for matching. Clearly some systems are going to be more tolerant of such variations than others.

Further attempts at presenting the biometric may sometimes be successful in such circumstances. If the system is supervised, the supervisor may be able to give advice in how to present a sufficiently good sample.

Matching biometric data with reference template

Case I. Badly presented biometric sample

It is also possible that a biometric sample is presented to the system badly, but the system does manage to extract biometric data. Examples here are the end-user presenting the wrong finger, or presenting a fingerprint at an angle that is out-of-tolerance for the matching algorithm

In a supervised system the supervisor will try to ensure that this does not happen. In these circumstances it is expected that the system will fail to match the biometric data and reference template. The situation sometimes occurs during enrolment, producing a reference template that will never match the end-user's biometric data at subsequent verification attempts.

Since quoted False Rejection Rates are normally obtained from results obtained with experienced users, or in a supervised environment, such figures will not include such mis-presentations

Case II. Well presented biometric sample

The estimate for the False Rejection Rate obtained from laboratory tests normally only covers cases where biometric data from a well-presented biometric sample fails to match with the appropriate reference template from another well-presented biometric sample. Indeed, as badly presented biometric samples are user-induced false rejections, there are good technical arguments for excluding such cases from the false rejection calculations!

The operational false rejection rate clearly depends on much more than the false rejection rate calculated under ideal conditions. It will also depend on usability issues, such as how likely it is that the user will have difficulty in presenting a good biometric sample. In many instances the failure to acquire rate and user-induced false rejections can dominate over the reference false rejection rate.

False Acceptance Rate

The False Acceptance Rate is defined as the probability that the biometric system will fail to reject an impostor.

In this section we consider how the reference false acceptance rate is calculated and whether this relates to false acceptances occurring in the real application.

Reference False Acceptance Rate

Laboratory tests for the False Acceptance Rate usually estimate the value as

$$\text{False Acceptance Rate} = \frac{\text{Number of False Acceptances}}{\text{Number of Imposter Verification Attempts}}$$

The impostor verification attempts are usually simulated by attempting to match biometric data from one enrolee against the biometric template for another. While the resulting measure does have relevance in determining how easily one enrolee can be confused with another, it is hard to infer from this figure the likely operational false acceptance rate.

Firstly, a genuine impostor is likely to make some effort to increase the likelihood of a false acceptance. This is not the case with the simulated verification attempt. For example, with a signature recognition system, an impostor is unlikely to use his own signature, but will produce a signature based on the identity claimed. *Figure 2* shows the difference between the simulated and actual impostor verification attempt.

A second problem with the estimate of the false acceptance rate is that false acceptances

	Comparison to be made Sample v Template
Simulated Impostor Verification Attempt	*J M Postor* =? *A N Other*
Actual Impostor Verification Attempt	*A N Other* =? *A N Other*

Figure 2: Real v Simulated Impostor Verification Attempts.

are unlikely to be evenly distributed across all enrolees. In practice only a few individuals may attempt to verify themselves against another enrolee's identity. However, once a false acceptance has been found, it is likely that the impostor will try and succeed with that claimed identity again.

Operational False Acceptance Rate and System Security

To determine how vulnerable the system might be to fraudulent use it is necessary to consider security aspects other than the reference (zero-effort) false acceptance rate.

Where the biometric has a behavioural aspect (for example signature or speaker recognition) ability to detect forgeries can be tested. However, some people are better than others at forging a signature or mimicking a voice, and it is not appropriate to calculate a forgery acceptance rate averaged over all users. Those attempting fraud by forgery are likely to be those most skilled!

The main difficulty in obtaining an objective measure of the operational false acceptance rate and the security of the system is that of considering all possible attacks on the system. This will normally require some knowledge of how the system works, and so is a particular problem with relatively new devices where the potential weaknesses have yet to be found. Proper security evaluation might be carried out in a similar manner to ITSEC evaluations.

Biometric Testing / Measurement Initiatives

Biometric Testing Services (BIOTEST)

BIOTEST is a European collaborative project aimed at developing standard metrics for comparing performance of biometric devices. The consortium is made up of companies with a diverse range of interests in the biometric field including vendors, users and evaluators. The main partners are: STI (in Spain); Sagem, CR2A-DI and CNET (in France); SIAB (in Italy) and NPL (in the UK).

The project has developed methodologies for measuring the key aspects of performance of biometric systems in terms of accuracy of identification, usability, and security. To aid in the testing of identification accuracy, databases of biometric samples are being built. Other aspects are measured with trials involving real users, and by expert assessment.

Testing services are to be offered by NPL, CR2A-DI and STI. Further details are available on the project's Web page:

http://www.npl.co.uk/npl/cise/this/biotest/

National Biometric Test Centre

The National Biometric Test Centre at San Jose State University is developing methods and standards for biometric systems at multiple levels of security. The centre is involved in evaluation of biometric systems, and in the

building of databases of biometric samples. The centre also has an educational role both for its students and for the biometrics industry. For further details see:

http://www-engr.sjsu.edu/~graduate/biometrics/

International Computer Security Association

The ICSA has recently established a certification programme for biometric products. The scheme looks at false acceptance and false rejection error rates measured in a laboratory environment. Details are available from the Commercial Biometric Developer's Consortium Web pages at:

http://www.ncsa.com/services/consortia/cbdc/

Conclusions

This paper has discussed some of the reasons why performance at the application level can be at variance with quoted reference figures based on laboratory tests. Reference error rates can be useful in helping predict reliability of verification, but only when usability factors are also measured to determine the extent that these affect performance. In ideal circumstances, for example with well-trained end-users, or in a supervised system the operational error rates may be close to the reference rates. In other conditions the operational rates may be significantly worse.

Though for meaningful comparison between systems it is the operational performance that should be considered, it is easier to infer this from application-independent reference figures than from figures obtained on a completely different application.

It is important that reference figures give enough information about the context in which they were obtained, to allow this extrapolation to be done, or to warn of the conditions required to attain the given rates. Current practice is deficient in this respect, though it is hoped that the recent initiatives mentioned might begin to improve matters.

Finally it should be noted that, on their own, the false rejection rate and false acceptance rate are not very useful in predicting performance. For a more complete view of performance, measurements of usability and security are also needed. The operational false rejection rate is dependent on several usability factors, while the operational false acceptance rate is affected by security issues.

Acknowledgements

The authors are involved in the BIOTEST project, and the paper reflects some if the issues raised in the course of this project. The work has been supported with DTI and EC funding.