Global Information Assurance Certification Paper

BIOMETRICS: Evaluation Criteria and Scenario Based Performance Testing

Warren Court

GSEC
Version 1.4b Option 1
June 30th, 2003

Abstract

This paper is meant as an introduction for corporations or organizations who want to conduct their own biometric studies. There are established methodologies and criteria from which to develop a test plan. This paper will present these methodologies and show how the author conducted his own testing.

No testing results are presented here and there are no references to actual biometric system vendors. Also the interpretation of the data extracted from any biometric test is beyond the scope of this paper.

Introduction

The biometric industry has exploded on to the computer security scene in the past 10 years. Sales are expected to grow to 1.9 Billion by 2005 [5]. Fuelled by increased security concerns after the terrorist attacks on September 11th, the biometric industry has been increasingly scrutinized to determine whether it brings to the table what it promises. There have been several well publicized roll outs of biometric security devices for trial periods [3]. Some of these have been perceived to have been performed badly and only add to the confusion of whether this technology is ready to be deployed in a real world situation.

How are the end users to determine which biometric system to go with and when they do decide which system how do they decide which particular vendor to choose? Is it feasible for a corporation to conduct its own tests of biometric systems?

There are several ways that an interested organization or corporation could go about testing a biometric system and garner the information they need to at least make better informed decision. The customer could contact one of the many organizations that are specializing in the testing of biometric devices. For a fee these test results can be made available or these organizations can conduct specialized tests at the requestor's behalf. There are also several online tests that could be utilized for free. It would be wise to include these tests as a source of more information to develop an opinion rather than relying on them alone.

An alternative would be that the prospective buyers of a biometric system could do tests themselves.

In House Tests

If the customer were to choose to conduct their own test there are established methodologies by which to follow but these methodologies alone might not establish whether a biometric technology was suitable for the intended purpose of the customer. There are other factors to consider such as perceived intrusiveness by the device among the potential user population.

Another problem a potential testing team might face is financial constraints. Most biometric systems perform reasonably well, therefore, to gather enough data by which to decide which system to pursue, could take extensive testing and taxing of resources. [2]

Basic Biometric Enrollment and Verification

Before discussing specific testing methodologies it is important to have an understanding of the generic process by which biometric systems function.

A Biometric sample, commonly referred to as a "corpus" is the extraction of a unique set of data from an individual. Almost all biometric systems perform in the same basic manner. They capture a biometric image and extract unique features from that input. Using an algorithm, the unique features are formed into a template and that template is encrypted and stored in a database. The user then comes back to the same type of system that he has enrolled on and tries to authenticate to that stored template.

- **Capture**

This refers to the capture of the whole biometric, for example, in the case of fingerprint scanning it refers to the copying of the whole fingerprint image.

- **Extract**

  Extracting a biometric is the means by which individual characteristics are obtained from the entire biometric input. ~~Again~~ In the case of fingerprint scanning, it is the extraction of minutia points. If not enough minutia points can be extracted from a submitted sample there may be a request for an additional sample. A limit is usually set to the number of attempts a system will make in trying to extract the necessary data. If this limit is reached with no satisfaction the user is deemed to have "failed to enroll".

- **Template**

  The extracted components are run through an algorithm and stored as a template. It is commonly accepted that a biological template cannot be reversed engineered to create a complete biometric input of a user such as creating a complete fingerprint from the biometric input. The template is usually encrypted and stored in a central database. After this template is created and stored the user usually tries to immediately verify against this template.

- **Verify**

  The user submits the same biological input and the stored template is extracted and compared to this new sample. If it matches the user is authenticated and if it does not match the user is rejected and usually locked out of a system after several failed attempts at verification.

One other concept that is necessary to understand is the difference between the two types of biometric input. The first is a physiological input. In general terms it is what you are. Examples are fingerprints, facial features, irises and retinas etc… Biological characteristics you cannot change, although they may change on their own over time.

The other type of biometric input is characteristic. Generally speaking, it is what you do. Examples would be speaking or writing. On a subconscious level we do these things on a consistent basis by speaking various words the same way or writing our signatures pretty much the same way all the time. From these consistencies enough data can be extracted to form a unique template of these actions.

Security and Performance Testing

The two types of testing are security and performance testing. Security testing tests an individual device against a security target not against other devices. The security target may be for testing purposes such as a database or server with fake data stored on it. A security test usually involves vulnerability testing and techniques that are used to test the effectiveness of other information security devices. An example is an advanced impostor trial where a silicon thumb with the impression of a real user's thumbprint embedded on it is used to try and obtain a false match against an existing template.

"*Every biometric can be defeated. If one allows sufficient amount of money and attempts. Employing biometrics increases thje security levels to such an extent that more often than not the cost of penetrating the system does not justify the rewards* [5]*".*

Security tests are a necessary step in determining the right biometric system to choose from, but also as equally as important is the performance test. The performance test simulates the real world conditions by which the biometric system will be used. A large number of biometric inputs are fed into the device, templates are extracted and verification attempts are made. Done repeatedly by using as many samples of corpus as readily available it is possible to compare the results of two similar biometric systems to see which one performs the best. [4]

Three Types of Performance Testing

Under the category of performance testing there are three types of tests: technological, scenario and operational.

Goal of technology evaluation is to compare competing algorithms using a single form of input device such as a finger print scanner. This test is best done using a very large database with thousands of either real or generated templates. Sometimes the input device is discarded all together and templates are run against the algorithm alone. This is called a simulation. The algorithm alone is isolated and bombarded with template after template in order to generate an Equal Error Rate (EER) which is the crossover between a False Acceptance Rate (FAR) and a False Rejection Rate (FRR). Testing one isolated part of a biometric system however will not give the potential buyer an overall feel for how that system will perform. It leaves out variables such as user reluctance or environmental conditions as well as how the other components such as the input device will operate.

The goal of scenario testing is to determine the overall performance of a complete system in as close to a real world situation as possible. The tests use

a variety of competing biometric devices from the same category such as two or three fingerprint scanning systems or multiple voice verification systems. Testing usually involves the use of a large pool of "live" inputs meaning real people sitting in front of the biometric device and being enrolled on it and then trying to re-verify against their stored template. As in a technology style test a database of templates can be used but extra configuration is needed as most out of the box biometric systems are not compatible with this type of testing.

The objective of operational testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population. The testing environment must be as close to the actual expected operating environment and may actually be used on a trial basis in the real world using real users. These types of tests are usually called trial runs and are typically publicized by the chosen biometric system [4].

EVALUATION CRITERIA

When trying to determine the biometric system to implement there are other factors besides which device performs the best. There are factors such as privacy and user reluctance as well as health concerns especially when scanning of the iris or retina is involved. Environmental conditions can also and should be factored into the testing. For example If a biometric system that captures a voice verification is to be used how is that verification protected from eavesdroppers who might be able to record their verification message.

Other concerns are user durability and these concerns can to a certain extend be examined with their own tests where how well the system performs from a verification standpoint might not be the issue but rather how well it stands up to repeated use by authorized users. This type of testing would be relevant for fingerprint or hand geometry scanners.

Biometric systems to roll out enterprise wide can be extremely expensive but beyond the ROI of a system there is the volatility in the biometric market, will the vendor chosen be around in a year or two after their system has been rolled out. This is of course a concern when choosing any type of information technology but it can become a deciding factor when choosing between two systems that performed equally as well in performance test.

There have been concerns raised about the use of the biometric input once it is obtained. Potential users might be ill-informed about the capabilities of the system for example can potential health issues such as AIDS or diabetes be detected through a retina scan. Whether or not this capability exists is

unimportant, if the user feels that this might be true they will be reluctant to use the system [6].

Performance Parameters

Most biometric systems have some sort of threshold function that will determine how easily a match is made against a stored template. If the threshold is set too high then very few legitimate users will be able to verify against their stored template and gain access. User annoyance will increase.

If the threshold is set too low then an unacceptable number of false match's will occur decreasing security. It is generally accepted to start testing by setting the biometric device threshold levels at their default setting. After sufficient testing at the default level the administrators of the test can adjust these levels and record the results.

Size of the testing pool

No accurate calculation as to how large a database of either real or simulated templates should be used for performance testing. Doddington's Rule states that testing should be done until 30 errors are achieved however this does not give an estimate of how large a database or sample group must be to achieve this. Best advice seems to be to use the largest possible sample group as possible. When using a CD-ROM full of user's templates it is necessary to invest in further configuration of a biometric input device and algorithm for this. This may lead to the decision to use only the algorithm and run the database against it. This therefore switches the test from a scenario one back to a technological one. [1]

Testing against a live person pool

If the test administrators are unable to configure a biometric system to receive a database of stored templates then they can use a "live" testing pool. This requires the recruitment of volunteers for the test. It is advisable to gather together as many volunteers as possible from as wide a demographic as possible. As with testing against a database the same rule applies for testing against a live pool, the larger the testing pool the better.

The potential volunteers for the test must be made aware of how much time the test will involve and more importantly how their biometric data will handled by the administrators over the course of the test and at its conclusion. This is necessary to alleviate user reluctance and fear of invasion of privacy. It may be of interest to the testing team to record the rate of user reluctance. An opinion

poll could be circulated before and after the test to gauge people's perceptions and acceptance of a biometric system.

DATA EXTRACTED

Enrollment

"Regardless of the accuracy of the matching Algorithm the performance of a biometric system is compromised if an individual cannot enroll or if they cannot present a satisfactory image at a later attempt" [1].

Enrollment failure can be divided into two categories. The first is called *Failure to Acquire* and means that the biometric system is incapable of extracting the required data to establish a template. *Failure to Acquire* is system related. No further testing is required with this particular system because in short it does not work [8].

The other type of failure is *Failure to Enroll* and it means that the subject who is attempting to enroll is unable to provide a sufficient quality of data to create a template. It is subject related. Further testing can be done by modifying the user input i.e. if a user is asked to provide a fingerprint sample from their right hand they are disabled in some way and cannot provide it they can be asked to provide a sample from their left hand. It is important however to record the number of failures to enroll because of the inability of the subject to provide the required biometric input as this will be indicative of the real world circumstances the administrator's of the system are likely to encounter.

Determining between the two types of enrollment failure can be difficult. Unless the organization conducting the test is preparing to publish the results it is easiest to consider *failure to acquire* and *failure to enroll* as one occurrence and call it failed enrollment.

False Match Rate (FMR) and False Non-Match Rate (FNMR)

False Non Match rates are individually recorded instances of a test subject failing to verify against their own template. False Match rates are a zero effort attempt by an impostor to verify against someone else's template.

The False Accept Rate (FAR) and False Reject Rate (FRR)

The FAR and FRR indicate overall system performance. They incorporate the entire FNM and FMR as well as failed enrollment data. From the FAR and FRR the Equal Error Rate is obtained and is usually the determining factor of a best performer in a scenario based performance test. Determining the equal error rate is outside the scope of this paper.

TESTING METHODOLOGY

When doing a scenario based performance test on a number of biometric devices it is best to group them by the method in which the biometric is captured. Within each type of biometric device category it is possible to further sub classify them. If a biometric device has a threshold setting that is adjustable by the administrator then it is best to group devices that have such a capability and ones that don't.

The next step would be to develop and engage in a run through of testing procedures and the recording of data before engaging in the test. This is to avoid user annoyance especially if the users are scheduled to come back for a return visit. Users who fail to show up for a re-verification after an initial time lapse from the enrollment date cancel themselves out from the test.

It is also wise to document and prepare for an instance where intervention in the test by the administrators if the subject fails in following the outlined procedure i.e. they present their index finger instead of thumb for finger print enrollment. This data should not be discarded however as it represents user ability and mimics real world frustrations the users may encounter when using the biometric device.

Environmental Conditions

Generally speaking the environmental conditions should closely match the real world scenario in which the biometric device is going to be used. However to get accurate scientific results if possible the environmental conditions during the enrollment phase should remain constant for at least the initial test. During the secondary visit after a time interval, environmental conditions might be allowed to vary after a stable set of environmental conditions is first used. Especially if the changing environmental conditions can definitely affect the test for example if voice verification is being used in a loud environment after initial testing it might be interesting to simulate this loud noise in the test environment.

The Testing Plan

Once a testing plan is created and tested a large subject pool or database acquired and any extraneous environmental considerations factored in it is time to begin testing.

First Visit

The users are assigned with usernames and are given instructions on how to enroll on the system. In most cases the biometric is so user friendly that it gives the instruction directly to the user.

Careful notation of the number of attempts to enroll should be recorded with a maximum set at a reasonable number. Usually a system will stop trying to enroll a person automatically. Sometimes a quality number or score is assigned to each enrollment or verification attempt. It is doubtful that this data will be of any help in determining a better biometric system as it comes down to either the user can enroll or they can't.

Once the subject has successfully enrolled, they can immediately try to verify against their stored template. Again the number attempts to re-verify should be set at a specific number and most systems will have an automatic time out rate set if a user repeatedly fails to verify.

Follow up Visits

It is advisable at least for characteristically based systems i.e. voice verification or signature recognition, to have a time lapse between the first and second visit. With physiologically biometric systems it is not as important as the inputs required for them are more stable than characteristic inputs. Fingerprints might scar but the chances of having a substantial number of test subjects having degraded fingerprint scans over say a six week period is so low it's not worth calculating.

It should be made clear to the testing subjects that they have a commitment to return for a second round or even a third round of testing. During the return visit the test subjects should only attempt to verify against their previously created template. There is no point in re-enrolling the subjects.

Ideally if a sufficiently large database of "live" subjects can be acquired the subject should attempt to enroll once, if successful would attempt to verify once and after a time lapse would come back to try and verify one more time only. This is usually impractical as it would require hundreds of users to obtain sufficient data from which to make any kind of judgment.

Impostor Trials

Once the subject pool has been put through the database it is usual practice to try and attempt an impostor trial. Impostors are people who try and verify as someone else using a zero effort attack meaning they are relying on their unaltered biometric matching someone else's template. The impostor may also try and better his chances by using an artifact (rubber finger with the person's captured fingerprint) or mimicry (trying to alter their appearance or their specific characteristic such as voice) to better their chances at obtaining a successful match. With mimicry the impostor can even better their chances by studying the existing user's of the biometric system and isolating a user that has similar characteristics to their own such as facial appearance or tone of voice.

Example of an Actual Test

The author of this paper conducted a real test of several commercially available biometric systems. The following describes the methodology used and problems encountered.

Selecting the Competitors

Research was done on several types of biometric systems. Several vendors were contacted and met with and eventually four biometric systems were obtained for testing. There were two fingerprint scanning systems and two voice verification systems. The two fingerprint scanning systems both had an adjustable threshold level so the testing team did not have an issue with grouping them together and conducting tests wherein the threshold level was manipulated. Of the two voice verification systems one system had adjustable threshold capability but it was decided to test the two systems against each other and to keep the adjustable threshold on the one system at the default level for the entire test. After the test was completed the team went back to the system with the adjustable threshold and conducted an independent test using this feature.

Testing Plan:

The team devised for the most part the entire testing plan up front with the exception of the impostor trials. We determined that after the initial test using "live" subjects was conducted we would have a better feel for the system and would be able to come up with inventive means of conducting an impostor trial.

Selection of Testing Pool

We did obtain a large database of voice templates from one organization in the United States but the time involved to develop a third party interface to allow it to be put through the voice verification systems would have delayed the test. It was kept as an option for further testing of the voice verification system that

performed the best in our initial test. We decided then to conduct the initial test using a pool of "live" subjects. We were able to gather 40 volunteers who were of diverse backgrounds and ages and best represented the potential user pool of a biometric system deployed across a large company's enterprise.

For the voice verification unit we grouped the volunteers by age, gender and by ethnic background. There were a number of people of an Eastern European background in our pool and we felt it best if they were grouped together at least for part of the test. Overall statistics of the entire group were calculated for both the fingerprint and voice verification and these ultimately led us to our conclusions.

Fingerprint Testing

Where possible we tried to conduct all fingerprint testing in one day meaning we enrolled a subject on both fingerprint systems and then had them repeatedly try and verify against their templates. On average the number of attempted verifications for each subject was 10. We did not feel there was a need to put in a time delay in between the initial enrollment and verification and follow up verifications for this particular method of biometric authentication.

Voice Verification Test

For the voice verification the sample group was enrolled on the system and a series of verification tests were done. The group was then rescheduled for an additional round of verifications.

Impostor Trials

After the initial tests were completed with the voice verification and fingerprint systems at our disposal we conducted several impostor trials. At first these impostor trials involved a zero effort attempt to verify against somebody else's template.

We then expanded the impostor tests to include the use of mimicry and artifacts. For the voice verification systems the impostors attempted to mimic the voice and inflection of the entire testing pool at first. This attempt was then zeroed down to a sub group within the testing pool that closely matched the age and background of the impostors.

After these tests we rigged a recording device and had several members of the original testing group come back and record their voice authentication message onto tape to be played back against the two systems in an attempt to verify against their original template.

For the fingerprint systems we attempted a zero effort attack at first. We were unable to go any further in-depth with this impostor attack due to lack of experience of specialized knowledge in fingerprint retrieval etc.

Problems Encountered

One of the problems we encountered was the availability of the user group for repeated tests. We were unable to keep the time interval between the first and second visit at a constant and were forced to schedule the return visits whenever the members of the user group were available. This was a minor problem and it was determined that it did not substantially affect our test but it would be a consideration if further testing were done.

We also encountered some user reluctance. We did not make a conscious effort to record this data but instead gathered a general impression that contributed to the conclusions drawn from the test. Another factor that we observed and made note of in a similar fashion was overall user friendliness.

Conclusion

Biometric devices are here to stay. The number of vendors providing this type of technology is growing every day. Existing technology is being improved upon. Such is the case with fingerprint scanning. The quality of fingerprint scanning device has vastly improved. Coupled with another biometric device or a smart card increases this type of technology substantially.

There are also new ways of capturing a unique biometric being developed. However relevant these systems will be remains to be seen; as of right now the most widely used biometric systems seem to be fingerprint, iris, face and hand geometry scanning and voice verification.

This paper has given an overview of the testing methodologies used and hopefully will encourage any corporation no matter what the size to conduct their own tests if for no other reason than to gain a better understanding of the technology available.

References

1. Biometrics Working Group, Best Practices and Reporting Performance of Biometric Devices (version1.0) http://www.cesg.gov.uk/site/iacs/itsec/media/protection-profiles/BBP.pdf

2. Government of Canada, Communications Security Establishment, Biometric Technology Security Evaluation Under The Common Criteria. http://www.csecst.gc.ca/en/documents/services/ccs/ccs_biometrics121.pdf

3. P. Jonathan Phillips, Alvin Martin, C.L. Wilson, Mark Przybocki, An Introduction to Evaluating Biometric Systems http://www.frvt.org/DLs/FERET7.pdf

4. Common Criteria Biometric Evaluation Methodology Working Group , Common Criteria, Common Methodology for Information Technology Security Evaluation, Biometric Evaluation Methodology Supplement

5. Stanley Glancy, Biometrics: The Body of Evidence http://www.expresscomputeronline.com

6. David Hochman, Biometrics: Security and Privacy Issues Come to a Head http://www.contingencyplanning.com/PastIssues/Mar2002/4.cfm

7. Tony Mansfield, Gavin Kelly, David Chandler and Jane Kane, Biometric Product Testing Final Report http://www.cesg.gov.uk/site/ast/biometrics/media/Biometric%20Test%20Report%20pt1.pdf

8. International Biometric Group, Comparative Biometric Testing: Official Test Plan http://www.biometricgroup.com/

9. Harold F. Tipton and Micki Krause, Information Security Management Handbook 4th edition, Auerbach Publications