Simple Test Procedure for Image-Based Biometric Verification Systems

C. L. Wilson, R. M. McCabe
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

This report discusses a simple test method for image-based biometric verification systems. A fingerprint based computer login system is used as an example of the process used in this test method. Ideally the algorithmic part of these systems should be performed using standard reference data such as NIST special database 24 [1] but it is still possible to test blackbox versions of systems where it is not possible to enter previously stored image data into the system. The procedure presented here allows such a system to be tested using manual input of the data and manual recording of results when no software interface to the system is provided. This test procedure also allows the effect of image quality of the input sensor to be evaluated in the typical working environment where the system is to be used. For the system tested here, the quality of the input image was found to be both important and highly user dependent. The trade-off between false positives and rejection of valid users was approximately as expected and as specified by the system manufacturer.

1 Introduction

A wide variety of computer security products are being made available with biometric verification keys as an alternative or supplement to PIN and password keys. In these products, a biometric key, such as the user's face, fingerprint, voice, or signature, is recognized to allow the user to access or change some information. The utility of systems of this kind is based on how successfully they match the pattern of the biometric to a previously stored sample of the same biometric. In these systems, the accuracy of the pattern match, from the quality of image typically available, will determine both the level of security provided and the ability of the user to obtain authorized access to the secured system. The trade-off between easy access and a high level of system security is still present but is now controlled by the pattern matching efficiency of the biometric access system. The test discussed in this report should allow users, with little access to the internal operation of the software being used, to test the operation of a biometric verification system in the environment in which it will be used.

1.1 Need for a simple test

Most testing of pattern recognition systems is done at the API (Application Program Interface) level where entry points into the software are available to perform the recognition and return the result of the match to the user. When this level of software interface is used, data from a standard test collection can be loaded into the software and the result can be scored using standard procedures similar to those widely used in areas such as face recognition [2, 3]. In some cases the potential system user will not have an API package but will need to evaluate a COTS (Commercial Off The Shelf) biometric system. In this case a small scale test which provides information about the performance of the system for the users specific application area is needed. Tests of this kind are important because the results obtained by most biometric systems are correlated with image quality. Face recognition depends on lighting conditions and, as we will show here, fingerprint recognition is correlated with user skin condition.

1.2 The verification and identification tasks

Biometric systems are commonly used for two tasks, verification and identification [4]. The verification task requires that the biometric be used to decide if the input biometric identifies a specific user, are you who you say you are. The identification task requires that the biometric be used to find a matching biometric signature, if it exists, in a database of biometrics such as the FERET test data [2, 3] for face or NIST Special Database 24 [1] for fingerprints. This distinction is important in both the system design and testing. In the verification system design, a relatively slow matching method can be used since only one match is needed to make the decision. In the identification task, the simplest matching strategy requires that as many matches be made as there are items in the database. A match per second might be fine for a once an hour login verification test while a 1000 matches per second would be very slow for the 300 million fingerprints in the FBI's criminal database.

Both verification and identification matching results generated by automatic matching, as opposed to human verified matching, are statistical estimates. The systems say that the answer is correct to some prespecified security level or to some specified probability. In the criminal identification application this probable identification is usually checked by trained examiners to reduce the possibility of a false identification.

1.3 Definition of terms

When an image is presented to the verification software one of five possible results is possible 1) Image rejected (IR); 2) True positive (TP); 3) False positive (FP); 4) True negative (TN); and 5) False negative (FN). The IR condition exists when an biometric target is present, such as a finger on the fingerprint reader but no **acceptable** image is detected by the matching system. The other four results can only be calculated on those images that are accepted by the matching software. The TP condition exist when the biometric system correctly matches the stored biometric; a match should occur and it does. The FP condition exists when the biometric system incorrectly matches the stored biometric; a match occurs but it should not occur. The TN condition exist when the biometric correctly fails to match the stored biometric; no match should occur and there is no match. The FN condition exists when the biometric incorrectly fails to match the stored biometric; a match should occur but does not.

1.4 Specific Example

In this report we consider a computer login application. The login uses a fingerprint biometric read from a fingerprint scanner located in a modified keyboard. The computer system operates using Microsoft's NT 4.0 operating system [1]. The software provided by the biometric system vendor allows a fingerprint to be used in place of a password for login and screen locking. When an account is authorized on the system, a fingerprint is also registered in an account database. This fingerprint can then be used as an alternate or a substitute for task that would usually require a user password.

In section 2, we describe the test procedure used to obtain data on the rate of occurrence of the five match results defined in section 1.3 for a biometric login system. In section 3, we discuss the level of security that is provided by the system used as a test example and the implications of the required security level on the measured system error rates. In section 4 we discuss the test results for our specific example system and in the concluding section we discuss some of the implications of the test on the applicability and usability of the specific biometric system and biometric systems in general.

2 Test procedure

The test performed here uses a round robin technique in which a specified number of biometrics, in this case fingerprints, are registered in the system. After registration, each biometric is tested against all of the others to obtain the error rates discussed above. If n biometrics are used then a total of n^2 tests are performed. Since this is a strictly external black box test, each test requires that the input sensor acquire a new image of the fingerprint. The frequency of failure of the capturing of the fingerprint image is then used to generate the IR error rate.

2.1 Registration of fingerprints

For the system used as the test example, each biometric was registered as a login key for an account on a PC running NT 4.0[1]. The account creation procedure was a modified version of the usual account creation process in which the administrator creates accounts. In addition to the usual account parameters the system can accept a fingerprint biometric which can be used as an alternative to a password. Since the account user's fingerprint is used, the account user must be present during account creation. The fingerprint is presented twice to the system once for registration and once for verification. The system generates an image quality score for each print and minimum scores are required before each print is accepted. During registration a match score is provided for each fingerprint imaged during the process. In the example system low scores during registration usually were predictive of above average image capture problems.

[1] Certain commercial software may be identified in order to adequately specify or describe the subject matter of this work. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the software identified is necessarily the best available for the purpose.

2.2 System testing

The test is carried out by instructing each registered user to attempt to gain access to each account. The login time, when an account is opened, slows the test process significantly but only should occur once for a single account. If IR error was detected by the system, each user was instructed to attempt access only five times; this would result in five IR errors being recorded. If access to the correct account was obtained, the test was scored as TP. If access was not obtained to the account and should have been the test was scored as FN. Accounts that should have been accessed but were not, were repeatedly tested using up to five attempts. Accounts that should not have been accessed were tested either five times or until the expected TN result was obtained. Accounts that should not have been accessed but were accessed were scored as FP and were tested up to five times to determine repeatability. Accounts that should not have been accessed and were not were scored TN and were not retested.

2.3 Sample size

In this example 40 fingers, using all ten fingers from four individuals, were used as a test sample. This required each person to make 400 tests and was large enough to test the minimum, vendor specified, false positive rate of 0.1%. The amount of effort required by each individual was strongly dependent on the IR rate that each person encountered. The minimum time required for the 400 test was approximately one hour. The maximum time used including the repetitions required by frequent IR errors was approximately four hours. When IR errors occurred, the delay in the system response increased sharply because multiple scans were made by the software before the IR error was generated. In cases where FP errors were encountered, the test was repeated to measure the sensitivity of the system to image quality in the FP situation.

2.4 Repetition

The dependence of livescan fingerprint, fingerprint taken from a real-time electronic image, matching on image quality factors that relate to the skin condition of the finger has been observed by many users of livescan equipment. To provide some estimate of this effect all of the individuals testing the system attempted to gain access to their ten accounts, one account for each finger, two weeks after the initial test and one individual ran the entire 400 test over. One of the individuals tested access to a single account, similar to the expected commercial use, several times a day over a period of weeks. In all the cases tested the system response was similar to the results obtained from a one to two hour intensive testing experiment. The individual who did the 400 tests again had high IR errors during the first test run. The IR errors were apparently better in more humid weather but these errors still occurred frequently.

2.5 Application to other situations

The simple round robin test method used here is applicable to tests of a biometric system that can not be tested with externally stored and generated data. Tests with other biometrics, such as face, would require a larger testing population since each user has only one face but would not be very much more difficult to do. The difference in this test from tests based

on stored images is that the ability of the input sensor to capture the input image is also tested. For many proposed biometrics this can be important. For fingerprints, the user's skin condition is important. For face recognition the angle and intensity of illumination is important. Since no broadly accepted standard for biometric image quality exists, the only way to test the input sensor under typical applications is to perform tests of the type presented here.

3 Level of security provided

Biometrics have the potential to discriminate between very large numbers of individuals. Expert testimony in legal cases suggests that this limit for fingerprints is 1 in 10^{97} [5]. For face recognition the limit should approach the identical twin limit of 1 in 10^4. These values can be obtained by careful examination of high quality images by humans. This does not imply that a automatic system using an image of uncertain quality can approach these levels of discrimination.

3.1 ROC curves

The ability of a biometric recognition system to discriminate between individuals is a trade-off between correct results and false alarms. An effective way to visualize the trade-off between correct results and errors treated as false alarms is to plot the ROC (Receiver Operating Curve) shown in figure 1 for an optical fingerprint verification system [6]. This set of three image resolution curves shows the effect that image resolution has on accuracy as a function of false alarm rate. The curves are generated by changing the recognition threshold from a low value where no false alarms are generated to higher values and calculating the percentage of correct results at each threshold. In terms of the errors listed above, the TP percentage is plotted against the FP percentage.

The only adjustable parameter for the system used in this example test allows the desired FP rate to be set in three steps of 0.1%, 0.01%, and 0.001%. This only allows a small part of the extreme left part of the ROC curve to be generated. Generating the ROC curve using manual input would in any case be very time consuming, since the entire test sequence would need to be repeated for each point. The curves are usually generated using sequences of prestored test image. This ensures that the same sequence of test images is used for each point on the curve. The procedure presented here usually will not provide as much information as is provides using a full ROC curve and for this reason may not be suitable of comparison of different biometrics.

3.2 Strength of required match

The strength of the required match as represented by the FP error rate, the rate at which unauthorized users are allowed to access the system, is a measure of the security provided by the biometric system. The FN rate is a reasonable measure of the rate at which valid users are denied access to the system. In tests with the relatively small sample size, used here, it may only be possible to bound the FP error. With 1000 samples it would be possible to detect FP errors that were greater than 0.001% but errors lower than this might not be detected. An application in which this error rate occurred would be only as good as a three digit PIN (Personal Identification Number) but the biometric can not be lost or stolen.

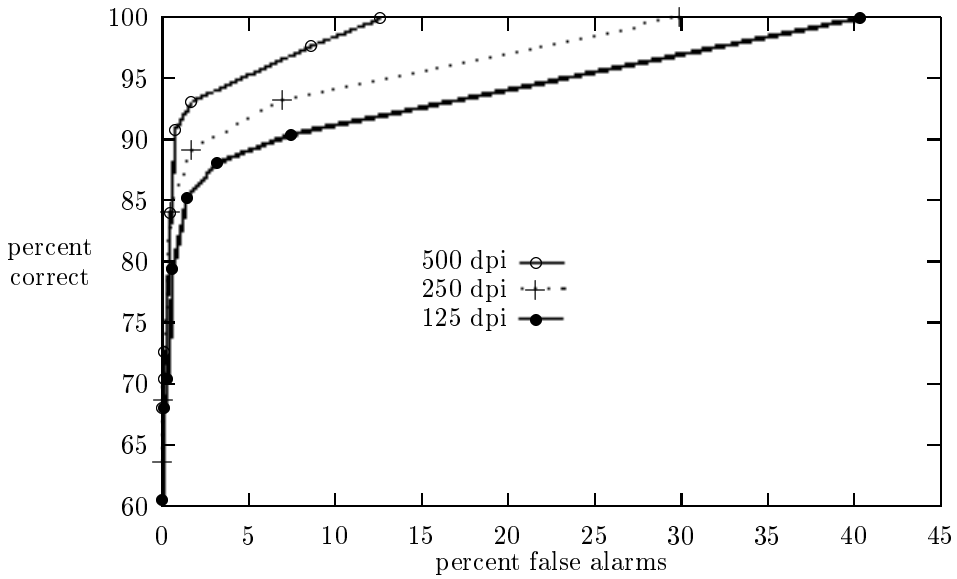

Figure 1: Example of a ROC curve for an optical fingerprint verification system test on 3500 fingerprint pairs.

Many biometric verification systems allow the user to specify an acceptable level of match strength, security, or probability of FP errors. As these criteria are made more stringent the sample size needed to test the system increases. If the expected FP error rate is 0.01% then the test sample to measure it must exceed 10,000. This makes the testing of higher security biometric verification system more time consuming and more expensive.

3.3 Nature of automatic matching

In any automatic pattern matching application, there is always some probable error associated with the result. In some types of application, such as character recognition on hand printed forms, a human correction procedure can be put in place to attempt to correct incorrect character recognition results. The biometeric matching systems of the type tested here are not subject to correction but require the user to try the match again. In character recognition, a recognition threshold can be used to limit the error rate of the recognition process. The same kind of threshold in biometric systems can be used to trade off security for user inconvenience. When a stringent matching requirement is used the security provided will be high but the possibility of denying access to a valid user is also high. When a low recognition threshold is used the possibility of denying access to a valid user is low but the security provided will also be lower. The trade-off between recognition threshold and false positives is often presented as a ROC curve.

4 Results of example tests

An estimate of the FP error, which was set at the highest allowed system level of 1 in 1000, should require at least a 1000 trial test. These tests were performed by creating 10 accounts, one for each finger, for each of four system users. Every account was then tested against

every finger. This resulted in approximately 1600 tests. One subject, user B, had such poor results that this set of 400 tests was repeated. Additional tests were required to overcome IR errors so a total of 2260 tests were performed. The results of these tests are shown in table 1.

Subject	$Tests$	IR-rate	TP-rate	FP-rate	FN-rate
A	414	1%	97%	0.2%	3%
B-1	630	36.6%	56%	0.0%	44%
C	413	5%	70%	0.2%	30%
D	422	0.5%	100%	0.0%	0%
B-2	481	16%	77%	0.2%	23%

Table 1: Table of error rates for four test subjects. These percentage error rates are definfined as: IR-rate $= 100IR/Tests$, TP-rate $= 100TP/(Tests - IR)$, FP-rate $= 100FP/(Tests - IR)$, NP-rate $= 100NP/(Tests - IR)$.

The FP error rate in the system tested could be set to three values, 0.1%, 0.01%, and 0.001%. Since several false positive finger combinations were found in our test with the 0.1% FP error rate we tested these combinations and the combinations that resulted in TP results to estimate the effect of these settings on FP errors and the expected increase in FN errors at more secure systems settings. These results are shown in table 2.

Subject	Level 1/1K	Level 1/10K	Level 1/100K
A-FP	0.2%	0%	0%
A-FN	3%	12%	20%
B-2-FP	0.2%	0%	0%
B-2-FN	23%	61%	79%
C-FP	0.4%	0.4%	0%
C-FN	30%	30%	47%

Table 2: Table of error rates, as defined in section 1.3, for three test subjects as a function of the system security level as controlled by the projected FP rate.

4.1 Failure to acquire image - very user sensitive

In our limited test, we found that the most variable results were associated with the inability of the system to detect images of adequate quality; this resulted in high IR error rates. Examination of table 1 shows that this was both user and time dependent. The least subject to IR errors was D with 0.5% while user B-1 had a 36.6% IR error rate. When user B repeated the test this IR error rate was reduced to 16%. Spot checks of results by the users with low IR error rates, users A and D, found no increase in IR error rate. In all the test performed here, a correlation between the IR error rate and the FN error rate was observed.

4.2 False positives - as expected

The FP errors in tables 1 and 2 are near those expected from the product specifications. No correlation between FP errors and IR and FN errors was measured. The 1600 tests in the top four rows of table 1 average to 0.1% exactly as predicted by product specifications. Increasing the strength of match, as illustrated by table 2, decreases the false positive rate. This rate is zero when a match strength of 1/100K is used. For users A and B, increasing the match strength to 1/10K eliminates all false positives. The false positives that are not eliminated by the 1/10K matching level for user C are due to small sample size.

4.3 False negatives - related to image quality and match strength

The FN appear from table 1 to be associated with IR which is visually observed to be related to image contrast and quality. The users who get above average IR errors also get most of the FN errors. The criteria that are used to reject images appear to be unrelated in the system tested to the strength of match. The number of IR errors generated using higher match strength was similar to the number found at low match strength. This was not true for the FN errors as was discussed above. FN errors increased with increasing match strength and security level.

4.4 Daily Systems Use

The most basic test of any biometric is to duplicate the conditions of use expected in the deployed system. In table 3, we present the results of one month of daily use by a user with high IR and FN error rates. The user typically accessed the system once a day and had a 44% chance of being given access to a valid account. Daily usage statistics are time consuming to collect and require that the test subjects have system access over the period of the test. Running a large enough set of daily test to accumulate good statistics on infrequent FP events may not be practical so a more exhaustive test is needed to accumulate these statistics.

Days	26
Attempted logins	28
TP	44%
FN	8%
IR	48%

Table 3: Table of error rates, as defined in section 1.3, for daily use by subject B over a period of one month. Subject B has the highest IR and FN error rates observed.

4.5 Sensitivity to users not easily predicted

The most surprising result obtained in this set of example test is that the IR rate is strongly dependent on user. User B over a period of one month has a consistently above average rate of both IR and FN errors. The manufacturer of the system has suggested that this is the result of dry skin. No correlation between factors like relative humidity that might effect

dry skin were measured. Users A and D retested their accessible accounts on the system one month later and found that the previous results were repeated. This suggests that any test of biometric access systems needs a large enough sample of users to detect user related differences in performance.

5 Conclusions

We concluded from this experiment that it is relatively easy to test systems for FP errors if these errors are in the 1 in 1000 range. A small number of users can perform enough tests in a relatively short time period, average test time was one to two hours, to check system performance for all five types of errors. If higher security levels are needed and FP errors are low the number of users and the test time grow linearly with the level of FP errors to be detected. In addition to the FP error rate in real applications the IR and FN rates may be very important in determining ease of use and the technology which achieves user acceptance. These rates appear to be user dependent and should be tested with worst case users under realistic conditions.

References

[1] C. I. Watson. Live-Scan Digital Video Fingerprint Database. Technical Report Special Database 24, National Institute of Standards and Technology, July 1998.

[2] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer-Verlag, Berlin, 1998.

[3] S. Rizvi, P. J. Phillips, and H. Moon. The FERET verification testing protocol for face recognition algorithms. *Image and Vision Computing Journal*, (to appear) 1999.

[4] P. J. Phillips, R. M. McCabe, and R. Chellappa. Biometric Image Processing and Recognition. In *Proceedings of the IX European Signal Processing Conference (EUSIPCO-98)*, 1998.

[5] M. R. Stiles, R. H. Levine, and P. A. Sarmousakis. Government's Combined Report to the Court and Motions In Limine Concerning Fingerprint Evidence. volume 704MITC2.PAS.vda, Philadelphia, March 1999. United States District Court for the Eastern District of Pennsylvania.

[6] C.I. Watson, P.J. Grother, E.G. Paek, and C.L. Wilson. Composite Filter for Vanderlugt Correlator. In *Optical Pattern Recognition IX, SPIE Aerosense Proceedings*, volume 3715, pages 53–59, Orlando, April 1999. SPIE.