Best Practices in Testing and Reporting Performance of Biometric Devices

Version 1.0
12 January 2000

Contents

1. Introduction

1. A review of the technical literature on biometric device testing reveals a wide variety of conflicting and contradictory testing protocols. Even single organizations produce multiple tests, each using a different test method. Protocols vary because test goals and available data vary from one test to the next. However, another reason for the various protocols is that no guidelines for their creation exist. The purpose of this draft document is to propose, for more general review by the biometrics community, "best practices" for conducting technical testing for the purpose of field performance estimation.

2. Biometric testing can be of three types: technology, scenario, or operational evaluation. Each type of test requires a different protocol and produces different results. Further, even for tests of a single type, the wide variety of biometric devices, sensors, vendor instructions, data acquisition methods, target applications and populations makes it impossible to present precise uniform testing

protocols. On the other hand, there are some specific philosophies and principles that can be applied over a broad range of test conditions.

3. This document concentrates on those measures that are generally applicable to all biometric devices. Aspects of testing which are device-specific, for example tests for image quality of fingerprint scanners shall be dealt with elsewhere.

4. Technical testing of both positive and negative identification devices requires assessment of an application and population-dependent "Receiver Operating Characteristic (ROC) curves". Negative ID systems also require error versus penetration rate assessment of any binning algorithms employed.

5. For both negative and positive ID systems, throughput rate estimation is also generally of great interest. In positive ID applications, throughput rate performance is more dependent upon the human factors than upon the technical. In negative ID systems, throughput rate is additionally limited by hardware processing speed. Additional measures of great interest in both positive and negative identification are the "failure-to-enroll" and "failure-to-acquire" rates.

6. We recognize that sometimes it will not be possible to follow best practice completely. However, we hope the guidelines highlight the potential pitfalls, making it easier for testers to explain reasons for any deviation and the likely effect on results.

2. Scope

7. This report will focus primarily on "best practices" for application and population-dependent ROC assessment in technical, scenario and operational testing. ROC curves are established through the enumeration of experimentally derived "genuine" and "impostor" distances (or scores)[1] . So the primary task is to establish "best practices" for the reasonable assessment of these distances and the "failure-to-enroll" and "failure-to-acquire" rates.

8. This best practice is intended to be applicable across the full range of biometric identification systems: i.e. both negative and positive ID systems, all biometric technologies, and all application and test types.

9. We recognize that ROC measures alone do not fully determine the performance of a biometric system. Usability, security vulnerability etc. of biometric devices are important too, but these issues are outside the scope of this best practice document.

3. Some Definitions

3.1 "Positive" and "Negative" Identification

10. Biometric authentication has traditionally been described as being for the purpose of either "verification" or "identification". In "verification" applications, the user claims an enrolled identity. In "identification" applications, the user makes no claim to identity. In "verification" systems, the user makes a "positive" claim to an identity, requiring the comparison of the submitted "sample" biometric measure to those measures previously "enrolled" (stored) for the claimed identity. In

[1] Hereafter, to simplify the text and with no loss in generality, scores will be referred to as "distances", even though we acknowledge that they will not always be distance measures in the mathematical meaning of the term.

"identification" systems, the user makes either no claim or an implicit "negative" claim to an enrolled identity, thus requiring the search of the entire enrolled database. The inversion of the hypotheses to be tested leads to a reversal in the meanings of "false acceptance" and "false rejection" rates and a reversal of their governing system equations for the two systems. We find the terms "positive" and "negative" identification to be richer descriptions of these same functions, emphasizing their conceptual and mathematical duality.

3.2 Three Basic Types of Evaluation[2]

11. The three basic types of evaluation of biometric systems are: 1) technology evaluation; 2) scenario evaluation; and 3) operational evaluation.

12. The goal of a technology evaluation is to compare competing algorithms from a single technology. Testing of all algorithms is done on a standardized database collected by a "universal" sensor. Nonetheless, performance against this database will depend upon both the environment and the population in which it was collected. Consequently, the "three bear" rule might be applied, attempting to create a database that is neither too difficult nor too easy for the algorithms to be tested. Although sample or example data may be distributed for developmental or tuning purposes prior to the test, the actual testing must be done on data which has not been previously seen by algorithm developers. Testing is done using "off-line" processing of the data. Because the database is fixed, results of technology tests are repeatable.

13. The goal of scenario testing is to determine the overall system performance in a prototype or simulated application. Testing is done on a complete system in an environment that models a "real-world" application of interest. Each tested system will have its own acquisition sensor and so will receive slightly different data. Consequently, care will be required that data collection across all tested systems is in the same environment with the same population. Depending upon data storage capabilities of each device, testing might be a combination of "off-line" and "live" comparisons. Test results will be repeatable only to the extent that the modelled scenario can be carefully controlled.

14. The goal of operational testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population. Depending upon data storage capabilities of the tested device, "off-line" testing might not be possible. In general, operational test results will not be repeatable because of unknown and undocumented differences between operational environments.

3.3 "Genuine" and "Unknown Impostor" Transactions

15. The careful definition of "genuine" and "impostor" transactions forms an important part of our test philosophy and can be used to resolve unusual test situations. These definitions are independent of the type of test being performed. A "genuine" transaction is a good faith attempt by a user to match their own stored template. An "impostor" transaction is a "zero effort" attempt, by a person unknown to the system, to match a stored template. Stored templates, used in both "impostor" and "genuine" transactions, are acquired from users making good faith attempts to enroll properly, as explicitly or implicitly defined by the system management.

[2] From P.J. Phillips, A. Martin, C. Wilson, M Przybocki, "Introduction to Evaluating Biometric Systems", IEEE Computer Magazine, January 2000

16. A person is "known" to the system if: 1) the person is enrolled; **and** 2) the enrollment affects the templates of others in the system. An enrolled person can be considered "unknown" with reference to others in the system only if the other templates are independent and not impacted by this enrollment. Eigenface systems using all enrolled images for creation of the basis-images and "cohort" based speaker recognition systems are two examples for which templates are not independent. Such systems cannot treat any enrolled person as "unknown" with reference to the other templates.

17. An impostor attempt is classed as "zero-effort" if the individual submits their own biometric feature as if they were attempting successful verification against their own template[3].

3.4 "False Match" and "False Non-Match" Rates

18. To avoid ambiguity with systems allowing multiple attempts, or having multiple templates we define (a) the false match rate and (b) the false non-match rate, to be the error rates of the matching algorithm from a **single** attempt-template comparison in the case of (a) an impostor attempt and (b) a genuine attempt. If each user is allowed one enrollment template and one verification attempt, the reported error rates will be the expected error rates for a single user, as opposed to a single attempt. Expected error rates of a single attempt are weighted by the varying activity levels across all users and consequently are not as fundamental a measure as the expected error rates of a single user.

3.5 "Receiver Operating Characteristic" Graphs

19. Receiver Operating Characteristic (ROC) curves are an accepted method for showing the performance of pattern matching algorithms over a range of decision criteria. They are commonly used (in a slightly modified form[4]) to show biometric system performance, plotting the false non-match rate against the false match rate as the decision threshold varies. Just as the error rates vary between different applications, populations and test types, so will the ROC graphs.

3.6 "Failure to Enroll" and "Failure to Acquire"

20. 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.

21. The "failure to enroll" rate is the proportion of the population for whom the system is unable to generate repeatable templates. This will include those unable to present the required biometric feature, those unable to produce an image of sufficient quality at enrollment, and those unable to match reliably against their template following an enrollment attempt. The failure to enroll rate will depend on the enrollment policy. For example in the case of failure, enrollment might be re-attempted at a later date.

---

[3] In the case of dynamic signature verification, an impostor would sign their own signature in a zero-effort attempt! In this and similar cases, where impostors may easily imitate aspects of the required biometric, for example through copying or tracing another static signatures, a second impostor measure will be needed. However such measures are outside the scope of this document.

[4] In the case of biometric systems the true ROC would plot the true match rate (i.e. 1 - the false non-match rate) against the false match rate. The modified ROC graph is also sometimes referred to as the "Detection Error Tradeoff (DET) graph".

22. The "failure to acquire" rate is the proportion of attempts for which the system is unable to capture or locate an image of sufficient quality. It measures problems in image capture of a transient nature: permanent problems will prevent enrollment resulting in no further attempts.

3.7 "Live" and "Off-line" Transactions

23. Testing a biometric system will involve collection of input images or data, which are used for template generation at enrollment, and for calculation of distance scores at later attempts. The images collected can either be used immediately for "live" enrollment or identification attempt, or may be stored and used later for "off-line" enrollment or identification. Technology testing will always involve data storage for later, "off-line" processing, but scenario and operational testing might not. Scenario and operational tests may make immediate use of the data only, not storing raw images for later, "off-line" transactions.

24. In both scenario and operational testing "live" transactions can be simpler for the tester: the system is operating in its usual manner, and (although recommended) storage of images is not absolutely necessary. "Off-line" testing allows greater control over which attempts and template images are to be used in any transaction, and, regardless of test type, is more appropriate than live testing in several circumstances mentioned later in this best practice document.

4. Prerequisites

25. Performance figures can be very application, environment and population dependent. These aspects should therefore be decided in advance of testing. For technical testing, a "generic" application and population might be envisioned, applying the "three-bears" rule. For scenario testing, a "real-world" application and population might be imagined and modeled in order that the biometric device can be tested on representative users, in a realistic environment. In operational testing, the environment and the population are determined "in situ" with little control over them by the experimenter.

26. In scenario and operational testing any adjustments to the devices for optimal performance (including quality and decision thresholds) will need to take place prior to data collection. This should be done in consultation with the vendor. For example, stricter quality control can result in fewer false matches and false non-matches, but a higher failure to acquire rate. The vendor is probably best placed to decide the optimal trade-off between these figures. The decision threshold also needs to be set appropriately if matching results are presented to the user: positive or negative feedback will affect user behavior.

27. "Off-line" generation of genuine and impostor distance measures will require use of software modules from the vendors Software Developer's Kits (SDK): for generation of enrollment templates from enrollment images; for extracting sample features from the test images; and for generating the distance measures between sample features and templates. Even in cases where "live" testing is permissible, the ability to generate distance measures "off-line" is recommended to allow repeatability of the results for checking etc.

5. The Volunteer "Crew"

28. Both the enrollment and transaction functions require input signals or images[5]. These input images must come originally from a test population, or "crew". We do not accept as "best practice" the generation of artificial images (or the generation of new images by changing data from real images). For scenario evaluation, this crew should be demographically similar to that of the target application for which performance will be predicted from test results. This will be the case if the test population can be randomly selected from the potential users for the target application. In other cases we must rely on volunteers. In the case of operational testing, the experimenter may have no control over the users of the system.

29. For technical and scenario evaluation, enrollment and testing will be done in different sessions, separated by days, weeks, months or years, depending upon the "template aging" anticipated in the target application. A test crew with stable membership over time is so difficult to find, and our understanding of the demographic factors affecting biometric system performance is so poor, that target population approximation will always be a major problem limiting the predictive value of our tests. In operational testing, the enrollment-test time interval generally be outside the control of the experimenter.

30. Further, as we have no statistical methods for determining the required size of the test, no statements can be made about the required size of this crew or the required number of operational uses. Application of the approximate "Doddington's Rule" of collecting data until 30 errors are recorded will not tell us in advance how may trials will be required. The best we can say is that the crew should be as large as practicable[6]. The measure of practicality could be the expense of crew recruitment and tracking.

31. Data developed from test populations is not statistically "stationary", meaning that 10 enrollment-test sample pairs from each of 100 people is not statistically equivalent to 1 enrollment-test sample pair from each of 1000 people. The number of people tested is more significant than the total number of attempts in determining test accuracy. Consequently as a "best practice", we prefer to design tests where there are relatively few (perhaps just one) enrollment-test sample pairs from each user. Of course, this adds to the expense of the test. In operational testing, it is necessary to "balance" the uses of the system over the users so that results are not dominated by a small group of excessively frequent users. Further, if we wish to strictly enforce our definition that error rates are expected values over users, not uses, data must be edited to allow one user per operational user.

32. Recruiting the crew from volunteers may bias the tests. People with unusual features, the regularly employed, or the physically challenged, for instance, may be under-represented in the sample population. Those with the strongest objections to the use of the biometric technology are unlikely to volunteer. The volunteer crew must be fully informed as to the required data collection procedure, must be aware of how the raw data will be used and disseminated, and must be told how many sessions of what length will be required. Regardless of the use of the data, the identities of the crew are never released. A consent form acknowledging that each volunteer

[5] Hereafter, with no loss of generality, we will refer to all input signals as "images", regardless of dimension.

[6] We also note that "the law of diminishing returns" applies to the improvement of confidence intervals with test size. A point will be reached where errors due to bias in the environment used, or in volunteer selection, will exceed those due to size of the crew and number of tests.

understands these issues must be signed, then maintained in confidence by the researchers. A sample consent form is included as Figure 2.

33. Volunteers in technical and scenario evaluations should be appropriately motivated so that their behavior follows that of the target application. If volunteers get bored with routine testing, they may be tempted to experiment, or be less careful. This must be avoided.

6. Collecting Enrollment Data

34. Collected biometric images are properly referred to as a "corpus". The information about those images and the volunteers who produced them is referred to as the "database". Both the corpus and the database can be corrupted by human error during the collection process. In fact, error rates in the database collection process may easily exceed those of the biometric device. For this reason, extreme care must be taken during data collection to avoid both corpus (mis-acquired image) and database (mislabeled volunteer ID or body part) errors. Data collection software minimizing the amount of data requiring keyboard entry, multiple collection personnel to double-check entered data, and built-in data redundancy are required Any unusual circumstance surrounding the collection effort must be documented by the collection personnel. Even with these precautions, data collection errors are likely to be made and will add uncertainty to the measured test results. "After-the-fact" database correction will be based upon whatever redundancies are built into the collection system.

35. Each volunteer may enroll only once (though an enrollment may generate more than one template, and multiple attempts at enrollment may be allowed to achieve one good enrollment). Care must be taken to prevent accidental multiple enrollments. In scenario and operational evaluations, images may be recorded as a corpus for "off-line" testing or may be input directly into the biometric system for "live" enrollment. In the latter case we recommend that the raw images used for the enrollment be recorded. In all evaluations, it is acceptable to perform "practice" tests at the time of enrollment to ensure that the enrollment images are of sufficient quality to produce a later match. Scores resulting from such "practice" tests must not be recorded as part of the "genuine" comparison record.

36. In scenario evaluations, enrollment must model the target application enrollment. The taxonomy of the enrollment environment will determine the applicability of the test results. Obviously, vendor recommendations should be followed and the details of the environment should be completely noted. The "noise" environment requires special care. Noise can be acoustic, in the case of speaker verification, or optical, in the case of eye, face, finger or hand imaging systems. Lighting "noise" is of concern in all systems using optical imaging, particularly any lighting falling directly on the sensor and uncontrolled reflections from the body part being imaged. Lighting conditions should reflect the proposed system environment as carefully as possible. It is especially important to note that test results in one noise environment will not be translatable to other environments.

37. In technical evaluations, every enrollment must be carried out under the same general conditions. Many data collection efforts have been ruined because of changes in the protocols or equipment during the extended course of collection[7]. The goal should be to control presentation and

[7] The most famous example is the "great divide" in the Switchboard speech corpus. During the course of data collection a power amplifier failed and was replace by another unit. Unfortunately, the frequency response characteristics of the new amplifier did not match that of the old, creating a "great divide" in the data and complicating the scientific analysis of algorithms based on the data.

transmission channel effects so that such effects are either: 1) uniform across all enrollees; or 2) randomly varying across enrollees.

38. Regardless of evaluation type, the quality control module may prevent acceptance of some enrollment attempts. Quality control modules for some systems requiring multiple images for enrollment will not accept images that vary highly between presentations, other quality control modules will reject single poor quality images. If these modules allow for tuning of the acceptance criteria, we recommend that vendor advice be followed. Multiple enrollment attempts should be allowed, with a pre-determined maximum number of attempts or maximum elapsed time. All quality scores and enrollment images should be recorded. Advice or remedial action to be taken with volunteers who fail an enrollment attempt should be predetermined as part of the test plan. The percentage of volunteers failing to enroll at the chosen criteria must be reported.

39. All quality control may not be automatic. Intervention by the experimenter may be required if the enrollment measure presented was inappropriate according to some pre-determined criteria[8]. For instance, enrolling volunteers may present the wrong finger, hand or eye, recite the wrong enrollment phrase or sign the wrong name. This data must be removed, but a record of such occurrences should be kept. In technical and scenario evaluations, enrollment data should not be removed simply because the enrolled template is an "outlier". In operational evaluations, no information regarding appropriate presentation may be available. Data editing to remove inappropriate biometric presentations may have to be based on removal of outliers, but the effect of this on resulting performance measures should be fully noted.

7. Collecting Test Data

40. For technical evaluations, test data should be collected in an environment that anticipates the capabilities of the algorithms to be tested: test data should be neither too hard nor too easy to match to the enrollment templates. For scenario evaluations , test data must be collected in an environment, including noise, that closely approximates the target application. For all types of tests, the test environment must be consistent throughout the collection process. Great precaution must be taken to prevent data entry errors and to document any unusual circumstances surrounding the collection. It is always advisable to minimize keystroke entry on the part of both volunteers and experimenters.

41. In technical and scenario evaluations, test data should be added to the corpus independently of whether or not it matches an enrolled template. Some vendor software will not record a measure from an enrolled user unless it matches the enrolled template. Data collection under such conditions will be severely biased in the direction of underestimating false non-match error rates. Data should be rejected only for predetermined causes independent of comparison scores.

42. In operational evaluations, it may not be possible to detect data collection errors. Data may be corrupted by impostors or genuine users who intentionally misuse the system. Although every effort must be made by the researcher to discourage these activities, data should not be removed from the corpus unless external validation of the misuse of the system is available.

43. For technical evaluations, the time interval between the enrollment and the test data will be determined by the desired difficulty of the test. Longer time intervals generally make for more

[8] As the tests progress, an enrollment supervisor may gain additional working knowledge of the system which could affect the way later enrollments are carried out. To guard against this, the enrollment process and criteria for supervisor intervention should be determined in advance.

difficulty in matching samples to templates due to the phenomenon known as "template aging". Template aging refers to the increase in error rates caused by time related changes in the biometric pattern, its presentation, and the sensor.

44. For scenario evaluations, test data must be separated in time from enrollment by an interval commensurate with "template ageing" of the target system. For most systems, this interval may not be known. In such cases, a rule of thumb would be to separate the samples at least by the general time of healing of that body part. For instance, for fingerprints, 2 to 3 weeks should be sufficient. Perhaps, eye structures heal faster, allowing image separation of only a few days. Considering a hair cut to be an injury to a body structure, facial images should perhaps be separated by one or two months. In the ideal case, between enrollment and the collection of test data, volunteers would use the system with the same frequency as the target application. However, this may not be a cost effective use of volunteers. It may be better to forego any interim use, but allow re-familiarization attempts immediately prior to test data collection.

45. Specific testing designed to test either user habituation or template aging will require multiple samples over time. If template aging and habituation occur on different time scales, the effects can be de-convolved by proper exploitation of the time differences. In general, however, there will be no way to de-convolve the counteracting effects of habituation (improving distance scores) and aging (degrading scores).

46. Operational evaluations may allow for the determination of the effects of template aging from the acquired data if the collected data carries a time stamp.

47. In both technical and scenario evaluations, the collection must ensure that presentation and channel effects are either: 1) uniform across all volunteers; or 2) randomly varying across volunteers. If the effects are held uniform across volunteers, then the same presentation and channel controls in place during enrollment must be in place for the collection of the test data. Systematic variation of presentation and channel effects between enrollment and test data will obviously lead to results distorted by these factors. If the presentation and channel effects are allowed to vary randomly across test volunteers, there must be no correlation in these effects between enrollment and test sessions across all volunteers.

48. Not every member of the test population will be able to test in the system. The "failure to acquire" rate measures the percentage of the population unable to give a usable sample to the system as determined by either the experimenter or the quality control module. In operational tests, the experimenter should attempt to have the system operators acquire this information. As with enrollment, quality thresholds should be set in accordance with vendor advice.

49. All attempts, including failures to acquire, should be recorded. In addition to recording the raw image data, details should be kept of the quality measures for each sample if available and, in the case of "live" testing, the distance score(s).

50. In some scenario evaluations, distance scores may be calculated "live". This is **not** appropriate:

a) if stored templates are not independent; when the impostor distance scores are incorrect;
b) if comparison scores are not reported in full, as may be the case when the system tries matching against more than a single template;
c) if data is not recorded until a matching template is found; or if ranked matches are returned, as occurs in some identification system.

If the experimenter is certain that none of these conditions prevail, live scenario testing can be undertaken, but raw data should be recorded. If "live" testing is deemed appropriate, impostor

testing requires the random assignment (without replacement) of some number of impostor identities (less than or equal to the total number of enrolled identities) to each volunteer. Volunteers should not be told whether the current comparison is genuine or impostor to avoid even unconscious changes in presentation. Resulting impostor scores are recorded.

8. ROC Computation

51. The ROC measures will be developed from the genuine and impostor distances developed from comparisons between single test samples and single enrollment templates. These distances will be highly dependent upon the details of the test and training collection. As previously explained, we have no way to determine the number of distance measures needed for the required statistical accuracy of the test. Further, the distances will be highly dependent upon the quality control criteria in place for judging the acceptability of an acquired image. Stricter quality control will increase the "failure to acquire" rate, but decrease the false match and non-match error rates.

52. Each transaction will result in a recorded distance. Distances developed for genuine transactions will be ordered. Impostor distances will be handled similarly. Outliers will require investigation to determine if labeling errors are indicated. Removal of any scores from the test must be fully documented and will lead to external criticism of the test results.

53. In operational testing, development of impostor distances may not be straight forward. Inter-template comparisons will result in biased estimation of impostor distances if more than a single image is collected for the creation of the enrollment template. This is true whether the enrollment template is averaged or selected from the best enrollment image. No methods currently exist for correcting this bias. If the operational system saves sample images or extracted features, impostor distance can be computed "off-line". If this data is not saved, impostor distances can be obtained through "live testing". Because of the non-stationary statistical nature of the data across users, it is preferable to use many volunteer impostors, each challenging one non-self template than to use a few volunteers challenging many non-self templates. If the volunteer is aware that an impostor comparison is being made, changes in presentation behavior may result in unrepresentative results.

54. Distance histograms for both genuine and impostor scores can be instructive but will not be used in the development of the ROC. Consequently, we make no recommendations regarding the creation of the histograms from the transaction data, although this is a very important area of continuing research interest. The resulting histograms will be taken directly as the best estimates for the genuine and impostor distributions. Under no circumstances should models be substituted for either histogram as an estimate of the underlying distribution.

55. "Off-line" development of distance measures must be done with software modules of the type available from the vendors in Software Developer's Kits (SDK). For systems with independent templates, one module will create templates from enrollment images. A second module will create sample features from test samples. These will sometimes be the same piece of code. A third module will return a distance measure for any assignment of a sample feature to a template. If processing time is not a problem, all features can be compared to all templates. If there are N feature-template pairs, N^2 comparisons will obviously be performed. The resulting distances can be thought of or actually arranged into a matrix with the N "genuine" scores on the diagonal and N(N-1) "impostor" scores in the upper and lower triangles. The impostor comparisons will not be statistically independent, but this approach is statistically unbiased and represents a more efficient estimation technique than the use of only N randomly chosen impostor comparisons

56. In the case that only single samples are given for enrollment, and enrollment and test quality control are equivalent, N test (or enrollment) templates can be compared to the remaining (N-1) test (or enrollment) templates. Regardless of whether or not the resulting comparison matrix is symmetric, only the upper or the lower triangle should be used for N(N-1)/2 impostor comparison scores.

57. In addition to the N feature-template pairs, there may be R additional features and Q templates for which there are no mates. This presents no additional problems provided that the additional data was acquired under precisely the same conditions and the same general population as the feature-template pairs. There will still be N "genuine" comparisons. Now there will be (N+R)(N+Q)-N impostor comparisons. If the target operational system uses "binning" or "filtering" as a strategy to decrease the size of the search space, impostor testing should also be done with feature-template comparisons within the same binning set. The use of so-called "background databases" of biometric features acquired from different (possibly unknown) environments and populations cannot be considered "best practice".

58. "Genuine" scores are computed "off-line" in the same way for systems with independent or non-independent templates. All volunteer enrollment samples are processed, then each volunteer test sample is compared to the matching template to produce N distances.

59. For systems with non-independent templates, however, "impostor" distances may require the "jack-knife" approach to create the enrollment templates. The "jack-knife" approach is to enroll the entire crew with a single volunteer omitted. This omitted volunteer can then be used as an unknown impostor, comparing his/her sample to all (N-1) enrolled templates. If this enrollment process is repeated for each of the N volunteers, N(N-1) impostor distances can be generated. This approach may not be possible in operational tests.

60. A second approach for systems with non-independent templates is to sample, under the same conditions, an additional R volunteers who are not enrolled in the system. These R samples can be used as unknown impostors against each enrolled template creating RN impostor distances. This would be the desired approach in operational testing.

61. The ROC curves are established through the accumulation of the ordered "genuine" and "impostor" scores. Each point on the ROC curve represents a false match/ false non-match ordered pair, plotted parametrically with score, as the score is allowed to vary from zero to infinity.. The false match rate is the percentage of impostor scores encountered below the current value of the score parameter. The false non-match rate is the percentage of genuine scores not yet encountered at the score parameter. In other words, the false non-match rate is the complement of the percentage of genuine scores encountered at the score threshold. The curves should be plotted on "log-log" scales, with "False Match Rate" on the abscissa (x-axis) and "False Non-Match Rate" on the ordinate (y-axis). Error bars should not be used.

9. Uncertainty Levels

62. Because biometric comparisons at a given threshold do not represent independent "Bernoulli trials", at our current level of understanding, uncertainty levels owing to sample size cannot be calculated on the basis of the number of test attempts or the number of users in the trial.

63. In conducting the trial, many assumptions will have been made. For example in technical or scenario evaluations, we may assume that the volunteer crew is sufficiently representative of the target population, and that under-representation of some types of individual does not bias the

results. We probably assume that difference between the trial environment and that of the real application has little effect on the ROC. The extent to which such assumptions are valid will affect the uncertainty levels.

64. Where it is possible to check that our assumptions are reasonably correct this should be done. For example we might check that the error rates for an under-represented category of individuals are consistent with the overall rates. Or we may repeat some of the trial in different environmental conditions to check that the measured error rates are not unduly sensitive to small environmental changes.

10. Binning Error versus Penetration Rate Curve

65. Full testing of negative identification systems requires the evaluation of any binning algorithms in use. The purpose of these algorithms is to partition the template data into subspaces. An input sample is likewise partitioned and compared only to the portion of the template data that is of like partition(s). The penetration rate is defined as the expected percentage of the template data to be searched over all input samples under the rule that the search proceeds through the entire partition regardless of whether a match is found. Lower penetration rates indicate fewer searches and, hence, are desirable.

66. The process of partitioning the template data, however, can lead to partitioning errors. An error occurs if the enrollment template and a subsequent sample from the same biometric feature on the same user are placed in different partitions. In general, the more partitioning of the database that occurs the lower the penetration rate, but the greater the probability of a partitioning error. These competing design factors can be graphed as a binning error versus penetration rate curve.

67. Fortunately, the testing corpus collected for "off-line" testing can be used in a second test to establish both penetration and bin error rates. Both enrollment templates and test samples are binned using the offered algorithm. Binning errors are assessed by counting the number of matching template-sample pairs that were placed in non-communicating bins and reporting this as a fraction of the number of pairs assessed. The penetration rate is assessed by the brute-force counting of the number of comparisons required under the binning scheme for each sample against the template database. The average number over all input samples, divided by the size of the database, represents the penetration rate. These results can be graphed as a point on a two-dimensional graph.

68. Frequently, the partitioning algorithm will have tunable parameters. When this occurs, the experimenter might graph a series of points (a curve or a surface) expressing the penetration and error rate tradeoffs over the range of each parameter.

11. Reporting of Results and Interpretation

69. Performance measures such as the ROC curve, failure to enroll and failure to acquire rates, and binning penetration and error rates are dependent on test type, application and population. So that these measures can be interpreted correctly additional information should be given.

a) Details of the volunteer crew and test environment are needed. How well these approximate a other target populations and applications can then be judged.
b) The size of the volunteer crew and the number of attempt-template comparisons should be stated. The smaller the number of tests the larger the uncertainty in the results, even if this uncertainty cannot be quantified.

c) Details of the test procedure (for example enrollment policy), especially deviations from this best practice should also be given.

12. Multiple Tests

12.1 Technical Evaluations

70. The cost of data collection is so high that we are tempted to create technical evaluation protocols so that multiple tests can be conducted with one data collection effort. In the case of biometric devices for which image standards exist (fingerprint[9], face[10], voice[11]), it is possible to collect a single corpus for "off-line" testing of pattern matching algorithms from multiple vendors.

71. In effect, we are attempting to de-couple the data collection and signal processing sub-systems. This is not problem-free however, as these sub-systems are usually not completely independent. The quality control module, for instance, which may require the data collection sub-system to reacquire the image, is part of the signal processing sub-system. Further, even if image standards exist, the user interface which guides the data collection process, thus impacting image quality, will be vendor specific. Consequently, "off-line" technical evaluation of algorithms using a standardized corpus may not give a good indication of total system performance.

12.2 Scenario Evaluations

72. Multiple scenario evaluations can be conducted simultaneously by having a volunteer crew use several different devices or scenarios in each session. This approach will require some care. One possible problem is that the volunteers will become habituated as they move from device to device. To equalize this effect over all devices, the order of their presentation to each volunteer must be randomized.

73. A further potential problem occurs where ideal behavior for one device conflicts with that for another. For example some devices work best with a moving image, while others require a stationary image. Such conflicts may result in lower quality test images for one or more of the devices under test.

12.3 Operational Evaluations

74. Operational evaluations do not generally allow for multiple testing from the same collected data set.

13. Conclusions

75. We recognize that the recommendations in this document are extremely general in nature and that it will not be possible to follow best practice completely in any test. However, we hope that these

[9] FBI/NIST "Appendix G: Image Quality Standard for Scanners", although originally written for document scanners used to produce digitized images from inked fingerprint cards, it is held as a specification for fingerprint sensor image quality. The dual use of this standard is problematic, particularly for the non-optical fingerprint sensors.

[10] AAMVA Facial Imaging "Best Practices" Standard

[11] There are at least two de-facto standards for voice collection: the telephone handset standard of 4kHz sample bandwidth and the 22kHz audio CD bandwidth standard.

concepts can serve as a framework for the development of scientifically sound test protocols for a variety of devices in a range of environments.

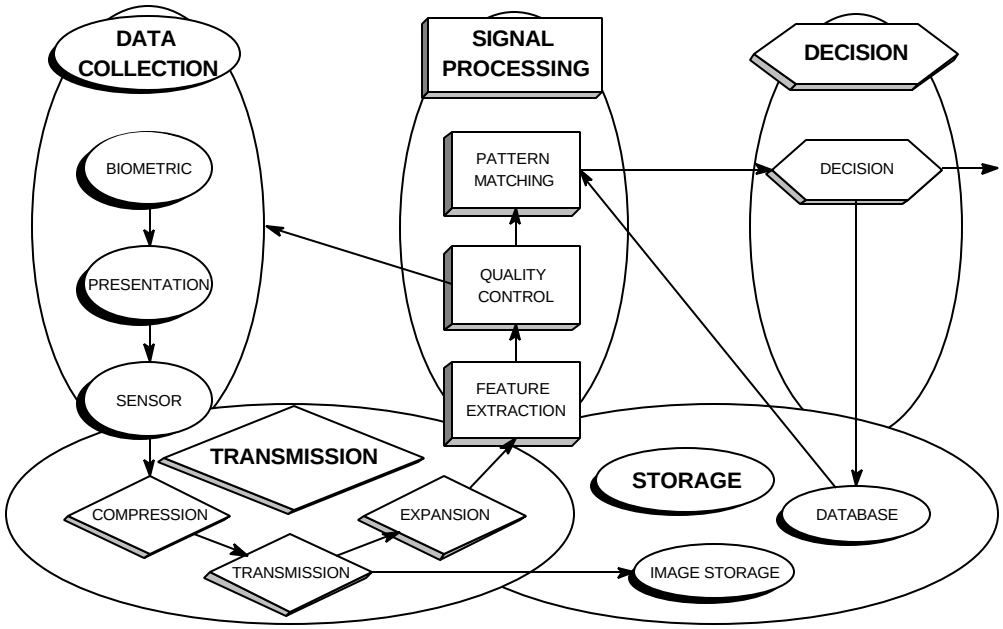

Figure 1 Diagram of General Biometric System

Consent form for Biometric Performance Trial	
Name	<*name*>
Contact Details	<*details*>
Identifier(s) used in Test Corpus	<*identifiers*>
I willingly participate in these trials. I consent to <*images/recordings*> of my <*finger/ face/ iris/ hand/* …> and my questionnaire responses[12] being collected during the trial and stored electronically. I agree to the use of this data by <*testing organization*> and <*list other companies that may use the data*> for the purposes of evaluating performance of biometric systems and identifying problems and improvements. I understand that my name[13]/identity will not be stored or shown in any released database[14]. or report.	
Signature	

Figure 2 Sample Volunteer Consent Form

[12] It can be useful to record other information about the volunteer crew, e.g. age occupation etc.

[13] May need to be changed when testing signature systems.

[14] When the corpus contains images from two types of biometrics, e.g. signatures and face images, it should not be possible to align the different types of images e.g. associating a face with a signature.