International Journal of Image and Graphics, Vol. 1, No. 1 (2001) 93–113
© World Scientific Publishing Company

FUNDAMENTALS OF BIOMETRIC AUTHENTICATION TECHNOLOGIES

JAMES L. WAYMAN*

Office of Research and Graduate Studies, San Jose State University,
San Jose, California, 95192-0080 USA

Biometric authentication technologies are used for the machine identification of individuals. The human-generated patterns used may be primarily physiological or behavioral, but usually contain elements of both components. Examples include voice, handwriting, face, eye and fingerprint identification. In this paper, we look at these technologies and their applications in general, developing a systematic approach to classifying, analyzing and evaluating them. A general system model is shown and test results for a number of technologies are considered.

Keywords: Biometrics; Authentication Technologies; Verification and Identification.

1. General Principles

1.1. *The functions of biometric identification devices*

The term "biometric authentication" refers to the automatic identification or identity verification of living individuals using physiological and behavioral characteristics. Biometric authentication is the "automatic", "real-time", "nonforensic" subset of the broader field of human identification. There are two distinct functions for biometric devices,

(i) To prove you are who you say you are.
(ii) To prove you are not who you say you are not.

These functions are "duals" of each other. In the first function, we really mean the act of linking the presenting person with an identity previously registered or enrolled in the system. The user of the biometric system makes a "positive" claim of identity which is "verified" by the automatic comparison of the submitted "sample" to the enrolled "template". If the system requires a "true" identity, this must be established at the time of enrollment with external documentation. Biometric systems do not inherently require knowledge of the user's "true" identity, allowing for the possibility of anonymous verification through biometrics. The purpose of a positive identification system is to prevent the use of a single identity

*E-mail: biomet@email.sjsu.edu

by multiple people. If a positive identification system fails to find a match between an enrolled template and a submitted sample, a "rejection" results. A match between sample and template results in an "acceptance". References 1–5 document several such systems.

The second function, establishing that you are not someone or not among a group of people already known to the system, constitutes the largest current use of biometrics: negative "identification". The purpose of a negative identification system is to prevent the use of multiple identities by a single person. Consequently, social service and driver's licensing systems use negative identification to prevent issuance of multiple documents to the same individual. If a negative identification system fails to find a match between the submitted sample and all the enrolled templates, an "acceptance" results. A match between the sample and one of the templates results in a "rejection". References 3 and 6 document several systems for "negative" identification.

A negative claim to identity (establishing that you are not who you say you are not) can only be accomplished through biometrics. For positive identification, however, there are multiple alternative technologies such as passwords, PINs (Personal Identification Numbers), cryptographic keys, and various "tokens" including identification cards. Both tokens and passwords have some inherent advantages over biometric identification. Security against "false acceptance" of randomly generated impostors can be made arbitrarily high by increasing the number of randomly generated digits or characters used for identification. Further, in the event of a "false rejection", people seem to blame themselves for PIN errors, blame the token for token errors, but blame the system for biometric errors. In the event of loss or compromise, the token, PIN, password or key can be changed and reissued, but a biometric measure cannot. Biometric and alternatively-based identification systems all require a method of "exception handling" in the event of token loss or biometric failure.

However, the use of passwords, PINs, keys and tokens carries the security problem of verifying that the presenter is the authorized user and not an unauthorized holder. Consequently, passwords and tokens can be used in conjunction with biometric identification to mitigate their vulnerability to unauthorized use. Most importantly, properly designed biometric systems can be faster and more convenient for the user and cheaper for the administrator than the alternatives. In our experience, the most successful biometric systems for performing the positive identification have been those aimed at increasing speed and convenience while maintaining adequate levels of security such as those of Refs. 1–5.

1.2. *Robustness, distinctiveness, accessibility, acceptability and availability*

There seems to be virtually no limit to the body parts, personal characteristics and imaging methods that have been suggested and used for biometric identification:

fingers, hands, feet, faces, eyes, ears, teeth, veins, voices, signatures, typing styles, gaits and odors. This author's claim to biometric development fame is a now-defunct system based on the resonance patterns of the human head measured through microphones placed in the users' ear canals. Which characteristic is best? The primary concerns are at least five-fold: the robustness, distinctiveness, accessibility, acceptability and availability of the biometric pattern. By robust, we mean repeatable, not subject to large changes over time. By distinctive, we mean the existence of wide differences in the pattern among the population. By accessible, we mean easily presented to an imaging sensor. By acceptable, we mean perceived as nonintrusive by the user. By available, we mean that some number of independent measures can be presented by each user. The head resonance system scores high on robustness, distinctiveness and availability, and low on accessibility and acceptability.

Let's compare fingerprinting to hand geometry with regard to these measures. Fingerprints are extremely distinctive, but not very robust, sitting at the very end of the major appendages we use to explore the world. Damaging fingerprints requires a few seconds of exposure to household cleaning chemicals. Many people have chronically dry skin and cannot present clear fingerprints. Conversely, hands are very robust, but not very distinctive. To change your hand geometry, you'd have to hit your hand very hard with a hammer. However, many people (a few in every 1000) have hands similar in shape to yours, so hand geometry is not very distinctive. Hands are easily presented without much training required, but most people initially misjudge the location of their fingerprints, assuming them to be on the tips of the fingers. Both methods require some "real-time" feedback to the user regarding proper presentation. Both fingerprints and the hand are accessible, being easily presented. In the 1990 Orkand study,[7] only 8% of customers at Department of Motor Vehicle offices who had just used a biometric device agreed that electronic fingerprinting "invades your privacy". Summarizing the results of a lengthy survey, the study rated the public acceptance of electronic fingerprinting at 96%. To our knowledge, there is no comparable polling of users regarding hand geometry, but we hypothesize that the figures would not be too different. With regard to availability, our studies have shown that a person can repeatably present at least 6 nearly-independent fingerprints, but only one hand geometry (your left hand may be a near mirror image of your right).

What about eye-based methods such as iris and retinal scanning? Eyes are very robust. Humans go to great effort, though both the autonomic and voluntary nervous system, to protect the eye from any damage which heals quickly when it does occur. The eye structure, further, appears to be quite distinctive. On the other hand, the eye is not easy to present although the Orkand study showed that the time required to present the retina was slightly less than that required for the imaging of a fingerprint. No similar studies exist for iris scanning, but our experience indicates that the time required for presentation is not much different from retinal scanning. Proper collection of an iris scan requires a well-trained operator, a cooperative subject, adjusted equipment and well-controlled lighting conditions.

Regarding acceptability, iris scanning is said to have a public acceptance rate of 94%. The Orkand study[8] found a similar rate of acceptability for retinal scanning. The human has two irises for presentation. The question of retina availability is complicated by the fact that multiple areas of the retina can be presented by moving the eye in various directions.

The question of "which biometric device is best?" is very complicated. The answer depends upon the specifics of the application.

2. Classifying Applications

Each technology has strengths and (sometimes terminal) weaknesses depending upon the application in which it is used. Although each use of biometrics is clearly different, some striking similarities emerge when considering applications as a whole. All applications can be partitioned according to at least seven categories. This list of application categories is open, meaning that additional partitions might also be appropriate. We could also argue that not all possible partition permutations are equally likely or even permissible.

2.1. *Cooperative versus noncooperative*

The first partition is "cooperative/noncooperative". This refers to the behavior of the deceptive user. In applications verifying the positive claim of identity such as access control, the deceptive user is cooperating with the system in the attempt to be recognized as someone he/she is not. This we call a "cooperative" application. In applications verifying a negative claim to identity, the deceptive user is attempting to not cooperate with the system in an attempt not to be identified. This we call a "noncooperative" application. Users in cooperative applications may be asked to identify themselves in some way, perhaps with a card or a PIN, thereby limiting the database search of stored templates to that of a single claimed identity. Users in noncooperative applications cannot be relied on to identify themselves correctly, thereby requiring the search of a large portion of the database. Cooperative but so-called "PIN-less" verification applications also require search of the entire database.

2.2. *Overt versus covert*

The second partition is "overt/covert". If the user is aware that a biometric identifier is being measured, the use is overt. If unaware, the use is covert. Almost all conceivable access control and nonforensic applications are overt. Forensic applications can be covert. We could argue that this second partition dominates the first in that a wolf cannot cooperate or noncooperate unless the application is overt.

2.3. *Habituated versus nonhabituated*

The third partition, "habituated/nonhabituated", applies to the intended users of the application. Users presenting a biometric trait on a daily basis can be considered habituated after short period of time. Users who have not presented the trait recently can be considered “nonhabituated”. A more precise definition will be possible after we have better information relating system performance to frequency of use for a wide population over a wide field of devices. If all the intended users are “habituated”, the application is considered a “habituated” application. If all the intended users are “nonhabituated”, the application is considered “nonhabituated”. In general, all applications will be “nonhabituated” during the first week of operation and can have a mixture of habituated and nonhabituated users at any time thereafter. Access control to a secure work area is generally “habituated”. Access control to a sporting event is generally “nonhabituated”.

2.4. *Attended versus nonattended*

A fourth partition is “attended/unattended” and refers to whether the use of the biometric device during operation will be observed and guided by system management. Noncooperative applications will generally require supervised operation while cooperative operation may or may not. Nearly all systems supervise the enrollment process although some do not.[4]

2.5. *Standard environment*

A fifth partition is “standard/nonstandard operating environment”. If the application will take place indoors at standard temperature (20°C), pressure (1 atm.) and other environmental conditions particularly where lighting conditions can be controlled, it is considered a “standard environment” application. Outdoor systems and perhaps some unusual indoor systems are considered “nonstandard environment” applications.

2.6. *Public versus private*

A sixth partition is “public/private”. Will the users of the system be customers of the system management (public) or employees (private)? Clearly attitudes toward usage of the devices which will directly effect performance vary depending upon the relationship between the end-users and system management.

2.7. *Open versus closed*

A seventh partition is “open/closed”. Will the system be required, now or in the future, to exchange data with other biometric systems run by other management? For instance, some State social service agencies want to be able to exchange biometric information with other States. If a system is to be open, data collection, compression and format standards are required.

3. Examples of the Classification of Applications

Every application can be classified according to the above partitions. For instance, the positive biometric identification of users of the Immigration and Naturalization Service's Passenger Accelerated Service System (INSPASS),[3] currently in place at Kennedy, Newark, Los Angeles, Miami, San Francisco, Detroit, Dulles (Washington, D.C.), Vancouver and Toronto airports for rapidly admitting frequent travelers into the United States can be classified as a cooperative, overt, nonattended, nonhabituated, standard environment, public, closed application. The system is cooperative because those wishing to defeat the system will attempt to be identified as someone already holding a pass. It is overt because all will be aware that they are required to give a biometric measure as a condition of enrollment into this system. It is nonattended and in a standard environment because collection of the biometric will occur near the passport inspection counter inside the airports, but not under the direct observation of an INS employee. It is nonhabituated because most international travelers use the system less than once per month. The system is public because enrollment is open to any frequent traveler into the United States. It is closed because INSPASS does not exchange biometric information with any other system.

The biometric identification of motor vehicle drivers for the purpose of preventing the issuance of multiple licenses can be classified as a noncooperative, overt, attended, nonhabituated, standard environment, public, open application. It is noncooperative because those wishing to defeat the system attempt not to be identified as someone already holding a license. It is overt because all are aware of the requirement to give a biometric measure as a condition of receiving a license. It is attended and in a standard environment because collection of the biometric occurs at the licensing counter of a State Department of Motor Vehicles.[a] It is nonhabituated because drivers are only required to give a biometric identifier every four or five years upon license renewal. It is public because the system will be used by customers of the Departments of Motor Vehicles. All current systems are closed as States are not presently exchanging biometric information.

4. Classifying Devices

The consensus among the research community today is that all biometric devices have both physiological and behavioral components. Physiology plays a role in all technologies even those, such as speaker and signature recognition, previously classified as "behavioral".

The underlying physiology must be presented to the device. The act of presentation is a behavior. For instance, the ridges of a fingerprint are clearly physiological,

[a] Five States currently require fingerprints from driver's license applicants: California, Colorado, Georgia, Hawaii, and Texas. Michigan has made the practice illegal. A review of the use of biometrics in U.S. driver's licensing can be found in Ref. 47. The American Association of Motor Vehicle Administrators has recently developed standards for biometric identification in driver's licensing.

but the pressure, rotation and roll of the finger when presented to the sensor is based on the behavior of the user. Fingerprint images can be influenced by past behavior such as exposure to caustic chemicals as well. Clearly, all biometric devices have a behavioral component and behavior requires cooperation. A technology is incompatible with noncooperative applications to the extent that the measured characteristic can be controlled by behavior.

5. The Generic Biometric System

Although these devices rely on widely different technologies, much can be said about them in general. Figure 1 shows a generic biometric authentication system divided into five sub-systems: data collection, transmission, signal processing, decision and data storage. We will consider these subsystems one at a time.

5.1. *Data collection*

Biometric systems begin with the measurement of a behavioral/physiological characteristic. Key to all systems is the underlying assumption that the measured biometric characteristic is both distinctive between individuals and repeatable over time for the same individual. The problems in measuring and controlling these variations begin in the data collection subsystem.

The user's characteristic must be presented to a sensor. As already noted, the presentation of any biometric to the sensor introduces a behavioral component to every biometric method. The output of the sensor which is the input data upon which the system is built is the convolution of: (1) the biometric measure; (2) the way the measure is presented; and (3) the technical characteristics of the sensor. Both the repeatability and the distinctiveness of the measurement are negatively impacted by changes in any of these factors.[b] If a system is to be open, the presentation and sensor characteristics must be standardized to ensure that biometric characteristics collected with one system will match those collected on the same individual by another system. If a system is to be used in an overt, noncooperative application, the user must not be able to willfully change the biometric or its presentation sufficiently to avoid being matched to previous records.

5.2. *Transmission*

Some, but not all, biometric systems collect data at one location but store and/or process it at another. Such systems require data transmission. If a great amount of data is involved, compression may be required before transmission or storage to conserve bandwidth and storage space. Figure 1 shows compression and transmission occurring before the signal processing and image storage. In such cases, the

[b] The mathematical basic for this somewhat surprising statement linking distinctiveness to input variability is found in Ref. 18.

transmitted or stored compressed data must be expanded before further use. The process of compression and expansion generally causes quality loss in the restored signal with loss increasing with increasing compression ratio. The compression technique used will depend upon the biometric signal. An interesting area of research is in finding, for a given biometric technique, compression methods with minimum impact on the signal processing subsystem.

If a system is to be open, compression and transmission protocols must be standardized so that every user of the data can reconstruct the original signal. Standards currently exist for the compression of fingerprint (WSQ), facial images (JPEG), and voice data (CELP).

5.3. *Signal processing*

Having acquired and possibly transmitted a biometric characteristic, we must prepare it for matching with other like measures. Figure 1 divides the signal processing subsystem into three tasks: feature extraction, quality control, and pattern matching.

Feature extraction is fascinating. Our first goal is separate, in the presence of the noise and signal losses imposed by the transmission process, the true biometric pattern from the presentation and sensor characteristics also coming from the data collection subsystem. Our second, related goal is to preserve from the biometric pattern those qualities which are distinctive and repeatable, and to discard those which are not or are redundant. In a text-independent speaker recognition system, for instance, we may want to find the features, such as the frequency relationships in vowels, that depend only upon the speaker and not upon the words being spoken. And, we will want to focus on those features that remain unchanged even if the speaker has a cold or is not speaking directly into the microphone. There are as

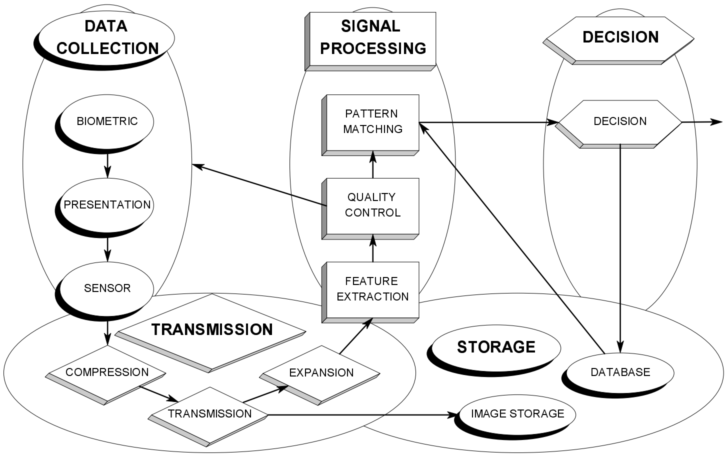

Fig. 1. Generic biometric system.

many wonderfully creative mathematical approaches to feature extraction as there are scientists and engineers in the biometrics industry. You can understand why such algorithms are always considered proprietary. Consequently, in an open system, the "open" stops here.

In general, feature extraction is a form of nonreversible compression, meaning that the original biometric image cannot be reconstructed from the extracted features. In some systems, transmission occurs after feature extraction to reduce the bandwidth required.

After feature extraction, or maybe even before or during, we will want to check to see if the signal received from the data collection subsystem is of good quality. If the features "don't make sense" or are insufficient in some way, we can conclude quickly that the received signal was defective. This is called a "failure to acquire" and we can request a new sample from the data collection subsystem while the user is still at the sensor. The development of this "quality control" process has greatly improved the performance of biometric systems in the last few short years. On the other hand, some people seem never to be able to present an acceptable signal to the system. If a negative decision by the quality control module cannot be over-ridden, a "failure to enroll" error results.

The feature "sample", now of very small size compared to the original signal, will be sent to the pattern matching process for comparison to one or more previously identified and stored features. The term "enrollment" refers to the placing of that feature "sample" into the database for the very first time. Once in the database and perhaps associated with an identity by external information (provided by the enrollee or others), the feature sample is referred to as the "template" for the individual to which it refers.

The purpose of the pattern matching process is to compare a presented feature sample to a stored template, and to send to the decision subsystem a quantitative measure of the comparison. An exception is enrollment in systems allowing multiple enrollments. In this application, the pattern matching process can be skipped. In the cooperative case where the user has claimed an identity or where there is but a single record in the current database (which might be a magnetic stripe card), the pattern matching process only makes a comparison against a single stored template. In all other cases, the pattern matching process compares the present sample to multiple templates from the database one-at-a-time as instructed by the decision subsystem, sending on a quantitative "distance" measure for each comparison.

For simplification, we will assume closely matching patterns to have small "distances" between them. Distances will rarely, if ever, be zero as there will always be some biometric, presentation, sensor or transmission related difference between the sample and template from even the same person.

5.4. *Decision*

The decision subsystem implements system policy by directing the database search determine "matches" or "nonmatches" based on the distance measures received from the pattern matcher and ultimately make an "accept/reject" decision based on the system policy. Such a policy could be to declare a match for any distance lower than a fixed threshold and "accept" a user on the basis of this single match, or the policy could be to declare a match for any distance lower than a user-dependent, time-variant, or environmentally-linked threshold and require matches from multiple measures for an "accept" decision. The policy could be to give all users, good-guys and bad-guys (deceptive users) alike, three tries to return a low distance measure and be "accepted" as matching a claimed template. Or, in the absence of a claimed template, the system policy could be to direct the search of all, or only a portion, of the database and return a single match, multiple "candidate" matches, or declare that no match was found.

The decision policy employed is a management decision that is specific to the operational and security requirements of the system. In general, lowering the number of false nonmatches can be traded against raising the number of false matches. The optimal system policy in this regard depends both upon the statistical characteristics of the comparison distances coming from the pattern matcher and upon the relative penalties for false match and false nonmatch within the system. In any case, in the testing of biometric devices, it is necessary to de-couple the performance of the signal processing subsystem from the policies implemented by the decision subsystem.

5.5. *Storage*

The remaining subsystem to be considered is that of storage. There will be one or more forms of storage used depending upon the biometric system. Feature templates will be stored in a database for comparison by the pattern matcher to incoming feature samples. For systems only performing "one-to-one" matching, the database may be distributed on magnetic stripe cards carried by each enrolled user. Depending upon system policy, no central database need exist although in this application, a centralized database can be used to detect counterfeit cards or to reissue lost cards without re-collecting the biometric pattern.

The database will be centralized if the system performs one-to-N matching with N greater than one as in the case of identification or "PIN-less" verification systems. As N gets very large, system speed requirements dictate that the database be partitioned into smaller subsets such that any feature sample need only be matched to the templates stored in one partition. This strategy has the effect of increasing system speed and decreasing false matches at the expense of increasing the false nonmatch rate owing to partitioning errors. This means that system error rates do not remain constant with increasing database size and identification systems do not linearly scale. Consequently, database partitioning strategies represent a complex policy decision. Scaling equations for biometric systems are given in Ref. 8.

If it may be necessary to reconstruct the biometric patterns from stored data, raw (although possibly compressed) data storage will be required. The biometric pattern is generally not reconstructable from the stored templates. Further, the templates themselves are created using the proprietary feature extraction algorithms of the system vendor. The storage of raw data allows changes in the system or system vendor to be made without the need to re-collect data from all enrolled users.

6. Testing

The principles of biometric testing have been detailed in Ref. 9. As originally suggested in Ref. 10, there are three basic types of tests: technology, scenario, and operational. Technology tests evaluate the performance of algorithms on pre-collected databases. References 10–18 describe such tests, often held in a competitive environment. Scenario evaluations seek to test a prototype biometric system in an environment that models some proposed application. Operational tests use data collected directly from an application (perhaps a "pilot project"). References 19 and 20 give examples of such a test. All tests generally focus on repeatability and distinctiveness of the measures, as reflected by the system error rates. Scenario and operational tests also usually measure user throughput rates, as well.

All type of testing require repeat visits with multiple human subjects. Further, the generally low error rates mean that many human subjects are required to detect even a few errors. The number of subjects required to produce "statistically significant" results, however, is not well understood.[9,20,21] The consensus in the test community is that systematic errors owing to uncontrolled variables are far more significant than random errors owing to small sample sizes.

Biometric testing is extremely expensive, generally affordable only by government agencies. Few biometric technologies have undergone rigorous, developer/ vendor-independent testing in scenario or operational environments. These points will be explored in more detail in this section.

6.1. *Distance distributions*

The most basic technical measures which we can use to determine the distinctiveness and repeatability of the biometric patterns are the distance measures output by the signal processing module.[c] Through testing, we can establish three application-dependent distributions based on these measures. The first distribution is created from distance measures resulting from comparison of samples to like templates. We call this the "genuine" distribution. It shows us the repeatability of measures from the same person. The second distribution is created from the distance measures resulting from comparison of templates from different enrolled individuals. We call this the "inter-template" distribution. The third distribution is created from the

[c] Strictly speaking, these are "scores" and may not represent distances in what mathematicians call a "metric space". We can assume without loss of generality that the larger the measure, the greater the difference between sample and template or template and template.

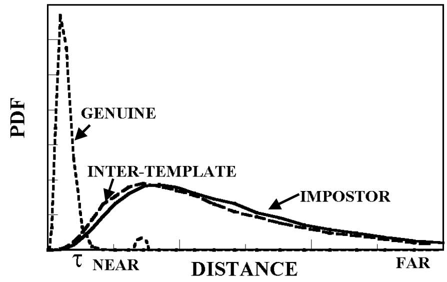

Fig. 2. Distance distributions.

distance between samples to nonlike templates. We call this the "impostor" distribution. It shows us the distinctiveness of measures from different individuals. A full mathematical development of these concepts is given in Ref. 18.

These distributions are shown as Fig. 2. Both the impostor and inter-template distributions lie generally to the right of the genuine distribution. The genuine distribution has a second "mode" (hump). We have noticed this in all of our experimental data. This second mode results from match attempts by people that can never reliably use the system (called "goats" in the literature[22–24]) and by otherwise biometrically-repeatable individuals that cannot use the system successfully on this particular occasion. All of us have days that we "just aren't ourselves". Convolution of the genuine and inter-template curves in the original space of the measurement, under the template creation policy, results in the impostor distribution. The mathematics for performing this convolution is discussed in Refs. 25 and 26.

If we were to establish a decision policy by picking a "threshold" distance, then declaring distances less than the threshold as a "match" and those greater to indicate "nonmatch", errors would inevitably be made because of the overlap between the genuine and impostor distributions. No threshold could cleanly separate the genuine and impostor distances. In a perfect system, the repeatability (genuine) distribution would be disjoint (nonoverlapping) from the impostor distribution. Clearly, decreasing the difficulty of the application category will affect the genuine distribution by making it easier for users to give repeatable samples, thus moving the genuine curve to the left and decreasing the overlap with the impostor distribution. Movement of the genuine distribution also causes secondary movement in the impostor distribution as the latter is the convolution of the inter-template and genuine distributions. We currently have no quantitative methodology or predicting movement of the distributions under varying applications.

In noncooperative applications, it is the goal of the deceptive user not to be identified. This can be accomplished by willful behavior, moving a personal distribution to the right and past a decision policy threshold. We do not know for any noncooperative system the extent to which deceptive users can move genuine measures to the right.

Some systems have strong quality-control modules and will not allow poor images to be accepted. Eliminating poor images by increasing the "failure to enroll" rate can decrease both false match and false nonmatch rates. Two identical devices can give different ROC curves based on the strictness of the quality-control module.

We emphasize that with the exception of arbitrary policies of the quality control module, these curves do not depend in any way upon system decision policy, but upon the basic distinctiveness and repeatability of the biometric patterns in this application. This leads us to the idea that maybe different systems in similar applications can be compared on the basis of these distributions. Even though there is unit area under each distribution, the curves themselves are not dimensionless owing to their expression in terms of the dimensional distance. We will need a nondimensional number if we are to compare two unrelated biometric systems using a common and basic technical performance measure.

6.2. *Nondimensional measures of comparison*

The most useful method for removing the dimensions from the results shown in Fig. 2 is to integrate the "impostor" distribution from zero to an upper bound τ. The value of the integral represents the probability that an impostor's score will be less than the decision threshold τ. Under a threshold-based decision policy, this area represents the probability of a single comparison "false match" at this threshold.

We can then integrate the "genuine" distribution from this same bound τ to infinity, the value of this integral representing the probability that a genuine score will be greater than the decision threshold. This area represents the probability of a single comparison "false nonmatch" at this threshold.

These two values, "false match" and "false nonmatch" for every τ can be displayed as a point on a graph with the false match on the abscissa (x-axis)

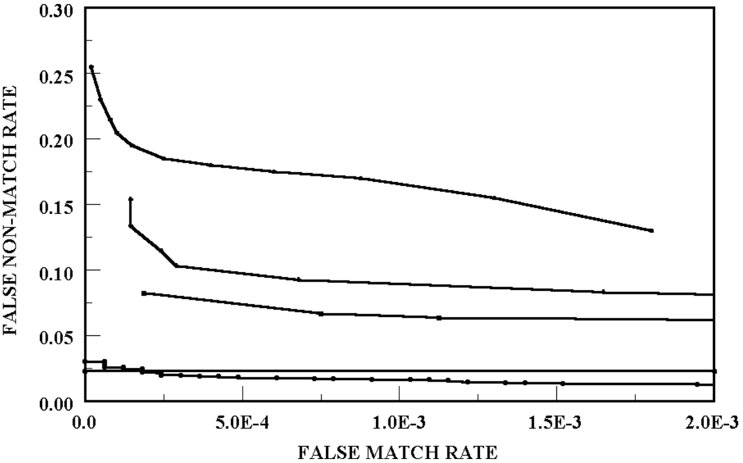

Fig. 3. AFIS benchmark ROC.

and the false nonmatch on the ordinate (y-axis). We have done this in Fig. 3 for five Automatic Fingerprint Identification System (AFIS) algorithms tested against a standard database. For historic reasons, this is called the "Receiver Operating Characteristic" or ROC curve.[27–29] Mathematical methods for using these measured false match and false nonmatch rates for "false acceptance" and "false rejection" prediction under a wide range of system decision policies have been established in Ref. 8.

Other measures have been suggested for use in biometric testing such as "D-prime"[30–32] and "Kullback–Leibler"[33] values. These are single, scalar measures, however, and are not translatable to error rate prediction.

We end this section by emphasizing that all of these measures are highly dependent upon the category of the application and the population demographics and are related to system error rates only through the decision policy. Nonetheless, false match and false nonmatch error rates as displayed in the ROC curve seem to be the only appropriate test measures allowing for even rudimentary system error performance prediction.

6.3. *Error bounds*

Methods for establishing error bounds on the ROC are not well understood. Each point on the ROC curve is calculated by integrating "genuine" and "impostor" distributions between zero and some threshold τ. Traditionally, as in Refs. 34 and 35, error bounds for the ROC at each threshold τ have been found through a summation of the binomial distribution. The confidence β given a **nonvarying** probability p of K sample/template comparison scores or fewer, out of N **independent** comparison scores being in the region of integration would be,

$$\beta = \Pr\{i \le K\} = \sum_{i=0}^{K} \frac{N!}{i!(N-i)!} p^i (1-p)^{N-i} \,. \tag{1}$$

In most biometric tests, values of N and K are too large to allow $N!$ and $K!$ in Eq. (1) to be computed directly. The general procedure is to substitute the "incomplete Beta function"[36,37] for the cumulative binomial distribution on the right hand side above, then numerically invert to find p for a given N, K, and β.

This equation can be used to determine the required size of a biometric test for a given level of confidence if the error probability is known *in advance*. Of course, the purpose of the test is to determine the error probability, so, in general, the required number of comparison scores (and test subjects) cannot be predicted prior to testing. To deal with this, "Doddington's Rule[d]" is to test until 30 errors have been observed. If the test is large enough to produce 30 errors, we will be about 95% sure that the "true" value of the error rate for this test lies within about 40% of that measured in Ref. 21.

[d] Named after U.S. Department of Defense speech scientist George Doddington.

Equation (1) will not be applicable to biometric systems if: (1) trials are not independent; (2) the error probability varies across the population with single users involved in more than one trial. If cross-comparisons (all samples compared to all templates except the matching one) are used to establish the "impostor distribution", the comparisons will not be independent and Eq. (1) will not apply. An equation for error bounds in this case has been given in Ref. 38 and has been verified using operational data in Ref. 20. The varying error probability across the population ("goats" with high false nonmatch errors and "lambs/wolves" with high false match errors)[22] similarly invalidates Eq. (1) as a generally appropriate equation for developing error bounds if users are involved in more than one trial.

Reference 20 establishes Eq. (1) as an applicable approximation for the false nonmatch rate under the rather restrictive condition that each enrolled user give but one test sample. Under this protocol, trials are independent and can be treated as coming from a uniform population. One interesting question to ask is "if we have no errors, what is the lowest false nonmatch error rate that can be statistically established for any threshold with a given number of comparisons?" We want to find the value of p such that the probability of no errors in N trials, purely by chance, is less than 5%. This is called the "95% confidence level". We apply Eq. (1) using $X = 0$,

$$0.05 > \Pr(K = 0) = \sum_{i=0}^{0} \frac{N!}{i!(i-N)!} p^i (1-p)^{N-i} = (1-p)^N \,. \tag{2}$$

This reduces to

$$\ln(0.05) > N \ln(1-p) \,. \tag{3}$$

For small p, $\ln(1-p) \approx -p$ and, further, $\ln(0.05) \approx -3$. Therefore, we can write,

$$N > \frac{3}{p} \tag{4}$$

This means that at 95% statistical confidence, error rates can never be shown to be smaller than three divided by the number of independent tests. For example, if we wish to establish false nonmatch error rates to be less than 3%, we will need to conduct 100 independent tests using 100 volunteers with no errors ($3/.03 = 100$).

The real problem with confidence intervals is that they refer to the statistical inaccuracy of a particular test owing to finite test size. The intervals in no way relate to future performance expectations for the tested device due to the much more significant systematic uncertainty regarding user population and overall application differences. The reporting of confidence intervals is not recommended in Ref. 9.

6.4. *Scenario testing*

The concepts of Secs. 6.1–6.3 apply to all tests, whether technical, scenario, or operational. In technical tests, we develop an ROC using the algorithms of a signal processing subsystem using a pre-collected database. The collection subsystem is separated in time and space from the signal processing subsystem. The resulting ROC for the algorithms tested are highly dependent upon the conditions under which the database was collected and the exact collection subsystem used.

To test how a complete system might perform in a real environment, we can construct a "scenario" test, replicating the target operational environment in our test protocols. All such test results must be interpreted in the context of the collection scenario and cannot be translated directly for performance prediction in other conditions. Most prior testing has been done in cooperative, overt, habituated, attended, standard environment, private, closed application of the test laboratory. This is the application most likely to yield low error rates. Clearly, people who are habitually cooperating with an attended system in an indoor environment with no data transmission requirements are the most able to give clear, repeatable biometric measures.

Use of a system at an outdoor amusement park[4] to assure the identity of non-transferable season ticket holders constitutes a cooperative, overt, nonhabituated, unattended, nonstandard environment, public, closed application. Performance in this application will not be well predicted from tests in the previously mentioned habituated, attended application

One of the major considerations in a scenario test is that of "template aging". All biometric measures change to some extent over time. In a target application, enrollment and use may be separated by years. To create a suitable scenario test, we must enroll and test volunteers over a time interval similar to that expected in the operational environment. That may not be possible for systems expected to operate over long time scales. 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.

A test population 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.

The ROC measures will be developed from the distributions of distances between samples created from the test data and templates created from the training data. Distances resulting from comparisons of samples and templates from the same people will be used to form the genuine distribution. Distances resulting from comparison of samples and templates from different people will be used to form the impostor distribution.

6.5. *Operational testing*

Given the expense of assembling and tracking human test subjects for multiple sample submissions over time and the limited, scenario-dependent nature of the resulting data, we are forced to ask, "Are there any alternatives to scenario testing for real-world performance prediction?" Perhaps the operational data from installed systems can be used for evaluating performance. Most systems maintain an activity log which includes transaction scores. These transaction scores can be used directly to create the genuine distribution of Fig. 2.

The problem with operational data is in creating the impostor distribution. Referring to Fig. 1, the general biometric system stores feature templates in the database and, rarely, compressed samples as well. If samples of all transactions are stored, our problems are nearly solved. Using the stored samples under the assumption that they are properly labeled (no impostors) and represent "good faith" efforts to use the system (no players, pranksters or clowns), we can compare the stored samples with nonlike templates, in "off-line" computation, to create the impostor distribution.

Unfortunately, operational samples are rarely stored due to memory restrictions. Templates are always stored, so perhaps they can be used in some way to compute the impostor distribution. Calculating the distance distribution between templates leads to the inter-template distribution of Fig. 2. Figure 2 was created using a simulation model based on biometric data from the Immigration and Naturalization Service Passenger Accelerated Service System (INSPASS) used for U.S. immigration screening at several airports.[19] It represents the relationship between genuine, impostor and inter-template distributions for this 9-dimensional case. Clearly, the inter-template distribution is a poor proxy for the impostor distribution. Figure 4 shows the difference in ROC curves resulting from the two cases.

Currently, we are not technically capable of correcting ROCs developed from inter-template distributions except in the case where the template resulted from a single enrollment sample. The correction factors depend upon the template creation policy (number of sample submissions for enrollment) and more difficult questions such as the assumed shape of the genuine distribution in the original template space.[18,25,26]

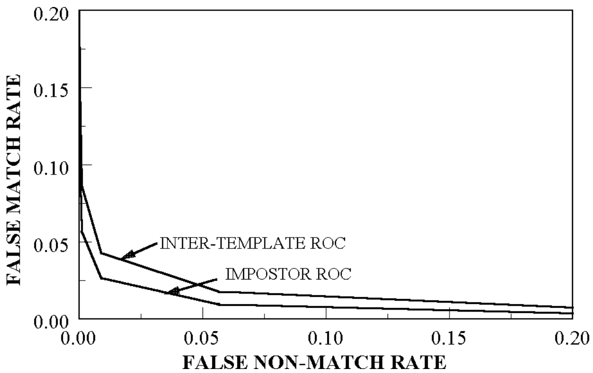

Fig. 4. Inter-template ROC bias.

7. Available Test Results

Results of some excellent tests are publicly available. The most sophisticated work has been done in technical tests of speaker verification algorithms. Much of this work is extremely mature, focusing on both the repeatability of sounds from a single speaker and the variation between speakers.[16,22,23,39–44] The (U.S.) National Institute of Standards and Technology (NIST) holds an annual test of speaker verification algorithms.[16] The scientific community has adopted general standards for biometric testing[9] modeled after the NIST speaker verification protocols. Technical testing for speaker verification algorithms on pre-stored "corpora" is aided by the existence of general standards for speech sampling rates and dynamic range.

In 1991, the Sandia National Laboratories released an excellent and widely available scenario test on voice, signature, fingerprint, retinal and hand geometry systems.[45] Although the results are now dated, this test has served as the prototype for scenario testing in an office environment. The test used data acquired in a laboratory setting from professional people well-acquainted with the devices. Error rates as a function of a variable threshold were reported as were results of a user acceptability survey. In April, 1996, Sandia released an operational evaluation of the IriScan prototype[46] in an access-control environment.

The Facial Recognition Technology (FERET) Program produced a number of excellent papers,[10–15] comparing facial recognition algorithms against standardized databases in technical tests. Earlier reports from this same project included a look at infrared imagery as well.[15]

In 1998, San Jose State University released the final report to the Federal Highway Administration[47] on the development of biometric standards for the identification of commercial drivers. This report includes the results of an international automatic fingerprint identification system benchmark test.

More recently, the Fingerprint Verification Competition 2000[17] reported results from a technical test of 11 algorithms against 4 different fingerprint databases.

8. Conclusion

The science of biometrics, although still in its infancy, is progressing extremely rapidly. Just as aeronautical engineering took decades to catch up with the Wright brothers, we hope to eventually catch up with the thousands of system users who are successfully using these devices in a wide variety of applications. The goal of the scientific community is to provide tools and test results to aid current and prospective users in selecting and employing biometric technologies in a secure, user-friendly, and cost-effective manner.

References

1. G. Koehler, "Biometrics: A case study — Using finger image access in an automated branch," *Proc. CTST'98*, Vol. 1, p. 535.
2. J. M. Floyd, "Biometrics at the University of Georgia," *Proc. CTST'96*, p. 429.

3. B. Wing, "Overview of all INS biometrics projects," *Proc. CTST'98*, p. 543.
4. Presentation by D. Welsh and K. Sweitzer, of Ride and Show Engineering, Walt Disney World, to CTST'97, May 21, 1997.
5. E. Boyle, "Banking on biometrics," *Proc. CTST'97*, p. 407.
6. D. Mintie, "Biometrics for state identification applications — Operational experiences," *Proc. CTST'98*, Vol. 1, p. 299.
7. Orkand Corporation, "Personal Identifier Project: Final Report," State of California Department of Motor Vehicles report DMV88-89 (1990), reprinted by the U.S. National Biometric Test Center.
8. J. L. Wayman, *Automation and Robotics* **6**(1), 35 (1999).
   Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.
9. "Best practices in testing and reporting performance of biometric devices," Version 1.0, United Kingdom Biometric Working Group, January, 2000.
   Available www.afb.org.uk/bestprac10.pdf.
10. P. J. Philips *et al.*, *Computer* **33**(2), 56 (2000).
11. P. J. Phillips *et al.*, "The FERET evaluation methodology for face-recognition algorithms," *Proc. IEEE Conf. Comp. Vis. and Patt. Recog.*, IEEE, San Juan, Puerto Rico, 1997.
12. S. A. Rizvi *et al.*, "The FERET verification testing protocol for face recognition algorithms," NIST, NISTIR 6281, October 1998.
13. P. J. Phillips *et al.*, "The FERET evaluation," in *Face Recognition: From Theory to Applications*, eds. H. Wechsler *et al.* (Springer-Verlag, Berlin, 1998).
14. P. J. Phillips, *Image and Vision Comput. J.* **16**(5), 295 (1998).
15. P. J. Rauss *et al.*, "FERET (Face-Recognition Technology) recognition algorithms," *Proc. ATRWG Science and Technology Conf.*, July 1996
16. A. Martin and M. Przybocki, *Digital Signal Processing* **10**, 1 (2000).
17. D. Maio *et al.*, "FVC2000: fingerprint verification competition," available at http://bias.csr.unibo.it/fvc2000.
18. J. L. Wayman, "Technical testing and evaluation of biometric identification devices," in *Biometrics: Personal Identification in Networked Society*, eds. A. Jain *et al.* (Kluwer Academic Press, 1998), p. 345.
    Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.
19. J. L. Wayman, "Evaluation of the INSPASS hand geometry data," in *National Biometric Test Center Collected Works: 1997–2000*,
    www.engr.sjsu.edu/biometrics/nbtccw.pdf.
20. J. L. Wayman, "Confidence interval and test size estimation for biometric data," *Proc. AutoID'99*, IEEE, Murry Hill, NJ, 1999, p. 177.
    Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.
21. J. E. Porter, "On the '30 error' criterion," ITT Industries Defense and Electronics Group, 1997.
    Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.
22. G. Doddington *et al.*, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," *Proc. ICSLP'98*, Sydney, Australia, November 1998.
23. J. P. Campbell Jr., "Speaker recognition: A tutorial," *Proc. IEEE* **85**, 1437 (1997).
24. L. O'Gorman, "Fingerprint Verification" in *op. cit.*, ed. A. Jain, p. 43.
25. C. Frenzen, "Convolution methods for mathematical problems in biometrics," Naval Postgraduate School Technical Report, NPS-MA-99-001, January 1999.
    Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.

26. P. Bickel, response to SAG Problem #97-2-1, University of California, Berkeley, Department of Statistics.
Available www.engr.sjsu.edu/biometrics/nbtccw.pdf.
27. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophys.* (J. Wiley and Sons, New York, 1966).
28. J. A. Swets (ed.), *Signal Detection and Recognition by Human Observers* (J. Wiley and Sons, New York, 1964).
29. J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic Press, 1975).
30. W. W. Peterson and T. G. Birdsall, "The theory of signal detectability," Electronic Defense Group, U. of MI., Tech. Report 13 (1954).
31. W. P. Tanner and J. A. Swets, *Psychological Rev.* **61**, 401 (1954).
32. J. Williams, "Proposed standard for biometric decidability," *Proc. CTST'96*, p. 223.
33. S. Kullback and R. Leibler, *Annals of Mathematical Statistics* **22**, 79 (1951).
34. W. Shen *et al.*, "Evaluation of automated biometrics-based identification and verification systems," *Proc. IEEE* **85**, 1464 (1997).
35. K. V. Diegert, "Estimating performance characteristics of biometric identifiers," *Proc. Biometric Consortium 8*, San Jose, CA, June, 1996.
36. M. Abromowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (John Wiley and Sons, New York, 1972).
37. W. H. Press *et al.*, *Numerical Recipes*, 2nd edition (Cambridge University Press, Cambridge, 1988).
38. P. Bickel, response to SAG Problem #97-2-3, University of California, Berkeley, Department of Statistics, unpublished.
39. B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE* **64**, 460 (1976).
40. A. Rosenberg, "Automatic speaker verification," *Proc. IEEE* **64**, 475 (1976).
41. N. Dixon and T. Martin, *Automatic Speech and Speaker Recognition* (IEEE Press, NY, 1979).
42. G. Doddington, "Speaker recognition: Identifying people by their voices," *Proc. IEEE* **73**, 1651 (1985).
43. A. Rosenberg and F. Soong, "Recent research in automatic speaker recognition", in *Advances in Speech Signal Processing*, eds. S. Furui and M. Sondhi (Marcel Dekker, 1991).
44. J. Naik, "Speaker verification: A tutorial," *Communications*, 42 (1990).
45. J. P. Holmes *et al.*, "A performance evaluation of biometric identification devices," Sandia National Laboratories, SAND91-0276, June 1991.
46. F. Bouchier, J. Ahrens, and G. Wells, "Laboratory evaluation of the IriScan prototype biometric identifier," Sandia National Laboratories, SAND96-1033, April 1996.
47. J. L. Wayman, "Biometric identifier standards research final report," College of Engineering, San Jose State University, San Jose, 1997.
Available www.engr.sjsu.edu/biometrics/fhwa.htm.

James Wayman received the Ph.D. degree in Engineering from the University of California at Santa Barbara in 1980 and joined the Faculty of the Department of Mathematics at the U.S. Naval Postgraduate School in 1981. In 1986, he became a full-time Researcher for the Department of Defense in the areas of technical security and biometrics, inventing and developing a biometric system based on the acoustic resonances of the human head. In 1995, he joined San Jose State University in San Jose, California and served as Founding Director of the U.S. National Biometric Test Center from 1997 to 2000. Dr. Wayman holds two patents in speech processing and is the author of dozens of articles in books, technical journals and conference proceedings on biometrics, speech compression, acoustics and network control. He serves on Editorial Boards of two journals and on several standards committees. He is a Senior Member of the Institute of Electrical and Electronic Engineers.