NATIONAL BIOMETRIC TEST CENTER

COLLECTED WORKS

1997-2000

San José State
UNIVERSITY

Edited by:

James L. Wayman, Director

Version 1.2

August, 2000

Prepared under DoD Contract MDA904-97-C-03

and FAA Award DTFA0300P10092

Table of Contents

Forward

The National Biometric Test Center was created at San Jose State University in April, 1997, by the Biometric Consortium, which is the U.S. government interest group on biometric authentication. At the 7th Biometric Consortium meeting held in March of 1995 at the FBI training facility in Quantico, VA, each of 5 competing proponents was required to state their concept for the Test Center. San Jose State proposed that the Test Center, rather than simply testing commercial products, should seek to advance the fundamental scientific understanding of biometric identification. This directional emphasis became important when, after the award, it was determined that the Test Center would not be allowed to directly collect biometric data from U.S. citizens because of its affiliation with the Department of Defense. We ultimately came to view the data collection prohibition as a good thing, as through it, we were forced to obtain data from operational installations with populations far broader than our students and from environments far more challenging than our laboratory spaces

The purpose to this "Collected Works" is to document the writings the many researchers working with or at the National Biometric Test Center during the three-and-one-half years between its inception in 1997 and the completion of its mission at the end of September, 2000. During this period, the Test Center received a total of $1.2M in federal government funding. Certainly, the Test Center was tasked with far more than simply writing technical papers. Archiving and disseminating information on biometrics, communicating findings to other groups and government agencies, participating on standards committees, and supporting a broad range of projects across many agencies were also fundamental activities of the Test Center. The writings, however, document concretely the scientific and mathematical progress made, and allow for the contributions of the Test Center to be reviewed by others in the field. Far from being the "last word" on any of these issues, these papers are intended only to stimulate academic and scientific debate on the nature of biometric identification. It is hoped that their publication in this volume will lead to active discussion and the discovery of new approaches and concepts.

I would like to thank particularly Dr. Joseph P. Campbell for his original efforts at organizing the Test Center concept and setting the research agenda, Dr. John Colombi for his excellent contribution as the government's director of the Test Center from 1997-1999, Jeff Dunn for assuming leadership after Dr. Colombi's departure, Dr. Jim Maar for allowing us to submit statistical problems to the government's Statistical Advisory Group, Dr. Tony Mansfield of the U.K. National Physical Laboratory and Philip Statham for their invaluable collaboration on nearly everything, and Deans Donald Kirk and Nabil Ibrahim for their efforts in creating the Test Center at San Jose State University in the first place.

James L. Wayman, Director
National Biometric Test Center
College of Engineering
San Jose State University
San Jose, CA 05102-0080

July 31, 2000

Introduction

The scientific agenda for the National Biometric Test Center was established by the Biometric Consortium in the 1995 Request for Proposal and in a series of questions posed at that time to the community by the Consortium Chair, Dr. Joseph P. Campbell. Why have biometric device tests failed to adequately predict "real-world(s)" performance? What operational factors affect error rates? Should testing results be reported as ROC curves or as rank order statistics? How big should tests be and can confidence intervals be placed on test outcomes?

In this compilation, we document our inquiry into these issues. This work follows the two volume "Biometric Identification Standards Development Final Report", submitted by San Jose State University to the Federal Highway Administration in 1997 to culminate the two-year study done on the use of biometrics in commercial driver's licensing. In that study, we focused our attention on a single application of biometrics, that of commercial driver identification. As the National Biometric Test Center (NBTC), however, we were concerned with a much broader range of potential applications. This compilation documents the ideas, methods and data developed. It is intended to be in a fairly logical order, but is not related to the chronology of our thinking. Each of the papers was intended to be a "stand alone" effort. Therefore, a considerable amount of repetition occurs when the papers are collected.

These are "Collected Works", not "Complete Works". Not all of our papers have been included—some omitted because of duplication of content, some because they contain sensitive commercial data, and some because of their limited focus. However, all scientific findings of general interest have been included in this compilation. In most cases, there were multiple versions of each paper written for slightly different applications. Many of the papers were published in peer-reviewed journals or conference proceedings. In the case of reviewed publications, outside redactors often made significant changes to the submitted text. This compilation represents the "Director's Cut", in that it includes those versions which are my personal favorites and not necessarily those that were ultimately published.

The first paper, "Fundamentals of Biometric Authentication Technologies", was posted on our web site for a couple of years. It appeared in the current form in the Proceedings of CardTech/SecurTech (CTST) 1999, having been revised from an original work appearing in the Proceedings of CTST the previous year as, 'Testing and Evaluating Biometric Technologies: What the Customer Needs To Know". The second and third papers step back a bit to consider more basic concepts. "A Definition of Biometrics" was first done in 1998 as unsolicited input for the International Biometric Industry Association as they were creating their initial policies on biometrics. "A Generalized Biometric Identification System Model" appeared in the Proceedings of the 31st Asilomar Conference on Signals, Systems and Computing in 1997 and presented for the first time the system model used in many of the subsequent writings.

One of the first tests conducted by the NBTC was on hand geometry data provided by the Immigration and Naturalization Service. That report is published here for the first time as "Report on the Evaluation of the INSPASS Hand Geometry System". This study pointed to the problem of biased estimation of the impostor distribution from template data. We asked the Department of Defense Statistical Advisory Group (SAG) for help in understanding the relationship between the genuine, impostor and inter-template

distributions. Prof. Peter Bickel of UC Berkeley responded with "SAG-97-2-1", showing how the three distributions were related under simplifying assumptions of isentropicity in the underlying vector space. In "Convolution Methods for Mathematical Problems in Biometrics", Chris Frenzen of the U.S. Naval Postgraduate School pushed the problem farther, showing how convolution methods could be used to develop the impostor distribution when only the inter-template and genuine distributions were available.

High on our agenda from the DoD was the issue of confidence interval estimation for test results. The "INSPASS" project was also problematic in this area. For independent comparisons from a population with a homogeneous error rate, the cumulative binomial distribution serves as a model for determining confidence intervals. Biometric Consortium Chair Joe Campbell brought to our attention the 1997 work of Jack E. Porter of ITT Industries. That previously unpublished paper, "On the '30 error' criterion", although not done at the NBTC, is included here (with permission of the author) to make it more accessible to the biometrics community.

In attempting to apply the binomial model to calculating confidence intervals, we ran into numerical problems with the inversion of the cumulative binomial distribution. NBTC Laboratory Director Prof. William Barrett provided methods in "Notes on the Numerical Solution of the Cumulative Binomial Probability Distribution".

The binomial distribution, however, does not serve as a model for errors under the more usual conditions of variable error rates ("goats", "wolves", "lambs" and "sheep") and non-independent comparisons. In response to a request to SAG, we received an excellent personal letter from Prof. Bickel showing how to calculate confidence intervals when non-independent (cross) comparisons are used. This letter is not included here, but the content is included in "Technical Testing and Evaluation of Biometric Identification Devices", which appeared as Chapter 17 in A. Jain, R. Bolle, and S. Pankanti (eds.) Biometrics: Personal Identification in Networked Society (Kluwer Academic Publishers, 1999). Prof. Bickel's equations were validated empirically in "Confidence Interval and Test Size Estimation for Biometric Identification Performance Data", published in the proceedings of the 1999 IEEE AutoID Conference

"Error Rates for the General Biometric System" was written as a companion paper to "Technical Testing….". While the latter discusses the statistical assessment of device performance in a specific environment with a specific population, the "Error Rates..." paper shows how to predict system performance probabilistically from the measured device performance. It presents scaling equations for large and small-scale systems under the assumption of measurement independence. Some of the equations approximate non-identically as identically distributed Bernoulli processes. How bad is this approximation? This problem is closely related to that of computing confidence intervals when error rates are not uniform across the population. Turning again to SAG, we received "A Memo on Non-Identically Distributed Bernoulli Model Problems for System Performance Prediction" from Hani Doss of Ohio State University and "Non-identically distributed Bernoulli sums" from Satish Iyengar of University of Pittsburg.

The issue of the scaleablility of large biometric systems became controversial in 1997. "Large-Scale Civilian Biometric Systems—Issues And Feasibility", included in the proceedings of CTST Government (1997), discusses some of the broader issues of large-scale systems. This topic was taken up again in "Continuing Controversy Over the Technical Feasibility of Large-Scale Systems", which appeared in November,1998, issue

of the "Biometrics In Human Services User Group (BHSUG**)** Newsletter #11", an online publication of the Department of Social Services of the State of Connecticut. www.dss.state.ct.us/digital.htm

Much of what we learned about large-scale systems was motivated by the work done with the Philippine Social Security System. We used the BHSUG newsletter as a journalistic forum for discussing the project as it developed. "The Philippine AFIS Benchmark Test Results", published in BHSUG Newsletter #8 in May, 1998. The 1999 submission to BHSUG Newsletter #12 "Philippine Social Security System Inaugurates Huge Civilian ID Card/AFIS System" reported on the big inaugural party held for the system.

The evaluation of large-scale systems raised many difficult questions. We turned again to the SAG to aid our understanding of the relationship between penetration rate and bin distributions. Kang James and Barry James of the University of Minnesota at Duluth responded with the paper "The 'Penetration Rate' in Automatic Fingerprint Identification Systems". Some systems tested returned only the top K scores when a sample was matched against many stored templates. Under these conditions, what can be said about the underlying "impostor" distribution? James and James contributed "Samples of the k Largest Order Statistics".

After concluding the Philippine benchmark tests, the NBTC tested other volunteering vendors using the same database. Analysis of this additional data led to the paper "Multi-Finger Penetration Rate and ROC Variability for Automatic Fingerprint Identification Systems" which looks at the question, "Should fingerprint systems use thumbs or forefingers?"

Prof. Barrett worked in the NBTC laboratory with many student groups on facial and iris recognition algorithm analysis and testing. His paper, "A Survey of Face Recognition Technologies and Testing Results", in included here.

The online publication "AVANTI" was used as a forum in 1999 to discuss the question of "'Degrees of Freedom' as Related to Biometric Device Performance".

The development of application programming interface standards was outside of our tasking. However, our government director, Dr. John Columbi, worked actively on the original Human Authentication Application Programming Interface (HAAPI), which eventually evolved into the Biometric Application Programming Interface (BioAPI) standard. "Engineering Tradeoffs in Biometric API Design" was published in the Proceedings of CTST in 1998.

In 1999, the Communications Electronic Security Group and the Central Information Technology Unit of the U.K. government organized a Biometrics Working Group. We participated with this group in producing a document summarizing a philosophy of testing biometric devices titled "Best Practices in Testing and Reporting Performance of Biometric Devices". Version 1.0 is included here.

Our mission statement required us to consider the "social and political implications of biometric identification". In "Reconciling Biometric Technologies with Government Due Process and Individual Privacy", we considered the Constitutional issues involved with biometric identification. In the BHSUG #xx submission, "When Bad Science Leads To Good Law: The Disturbing Irony Of The *Daubert* Hearing In The Case Of *U.S. V. Byron C. Mitchell"*, we looked at arguments presented to defend the scientific basis of fingerprinting. In "Federal Biometric Technology Legislation",

published in IEEE Computer, Mar. 2000, we looked at recent U.S. legislation promoting the application of biometrics to government activities. This paper was a condensation of a longer work done for CTST Government 1998. "Biometric Identification Technologies in Election Processes" was commissioned by the Federal Election Commission (FEC).

In 1998, we were asked to testify to the House Committee on Banking and Financial Services hearing on "Biometrics and the Future of Money". That testimony, given on May 20 1998, is reprinted here under the title, "Biometric Identification and the Financial Services Industry". Miss Shanin Leeming of Merritt Island, FL, also testified at this same hearing. Her Junior High School science project titled, "Picture ID: Help or Hindrance? Do People Really Look at the Picture on a Picture ID?", came to the attention of the House Committee after she sent us a copy to be posted on our web site. That project, although not done by the NBTC, is included here.

In 2000, we teamed with U.K. Biometrics Working Group member Ms. Lisa Alyea to present "Picking the Best Biometric for Your Applications" at CardTech/SecurTech.

Finally, a 2000 submission to the American Association of Motor Vehicle Administrators' MOVE magazine, "Biometric Authentication Standards Development ", completes this volume.

Fundamentals of Biometric Authentication Technologies

James L. Wayman, Director
U.S. National Biometric Test Center

I. General Principles

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", "non-forensic" subset of the broader field of human identification. There are two distinct functions for biometric devices:

1. To prove you are who you say you are.
2. 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". Clearly, establishing a "true" identity at the time of enrollment must be done with documentation external to any biometric system. The purpose of a positive identification system is to prevent the use of a single identity by multiple people. If a positive identification system fails to find a match between an enrollment template and a submitted sample, a "rejection" results. A match between sample and template results in an "acceptance".

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. 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".

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 postive identification have been those aimed at increasing speed and convenience, while maintaining adequate levels of security, such as those of references [1-5].

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. 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 non-intrusive 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 you use to explore the world. Damaging your fingerprints requires less than a minute of exposure to household cleaning chemicals. Many people have chronically dry skin and cannot present clear prints. 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 (somewhat less than 1 in 100) have hands much like 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 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, 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.

II. Classifying Applications

Each technology has strengths and (sometimes fatal) 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.

Cooperative versus Non-cooperative

The first partition is "cooperative/non-cooperative". This refers to the behavior of the "wolf", (bad guy or 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 s/he is not. This we call a "cooperative" application. In applications verifying a negative claim to identity, the bad guy is attempting to deceptively not cooperate with the system in an attempt not to be identified. This we call a "non-cooperative" 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 non-cooperative 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.

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 non-forensic 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 non-cooperate unless the application is overt.

Habituated versus Non-habituated

The third partition, "habituated/non-habituated", 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 "non-habituated". 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 "non-habituated", the application is considered "non-habituated". In general, all applications will be "non-habituated" during the first week of operation, and can have a mixture of habituated and non-habituated users at any time thereafter. Access control to a secure work area is generally "habituated". Access control to a sporting event is generally "non-habituated".

Attended versus Non-attended

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. Non-cooperative 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].

Standard Environment

A fifth partition is "standard/non-standard operating environment". If the application will take place indoors at standard temperature (20^{o} 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 "non-standard environment" applications.

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.

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.

This list 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.

III. 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, Vancouver and Toronto airports for rapidly admitting frequent travelers into the United States, can be classified as a cooperative, overt, non-attended, non-habituated, 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 will be overt

because all will be aware that they are required to give a biometric measure as a condition of enrollment into this system. It will be non-attended 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 will be non-habituated 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 non-cooperative, overt, attended, non-habituated, standard environment, public, open application. It is non-cooperative because those wishing to defeat the system attempt not to be identified as someone already holding a license. It is be 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[1]. It is non-habituated 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.

IV. Classifying Devices

In last year's papers at this meeting, I argued that biometric devices were based primarily on either behavioral or physiological measures and could be classified accordingly. 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, 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 non-cooperative applications to the extent that the measured characteristic can be controlled by behavior.

---

[1] Five States currently collect fingerprints from driver's license applicants: California, Colorado,, Georgia, Hawaii, and Texas. Michigan has made the practice illegal and similar legislation is pending in Alabama. A review of the use of biometrics in U.S. drivers' licensing can be found in Wayman [38]. Currently, the ANSI B10.8 committee is considering standards for biometric identification for drivers' licensing.

V. 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.

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[2]. 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, non-cooperative application, the user must not be able to willfully change the biometric or its presentation sufficiently to avoid being matched to previous records.

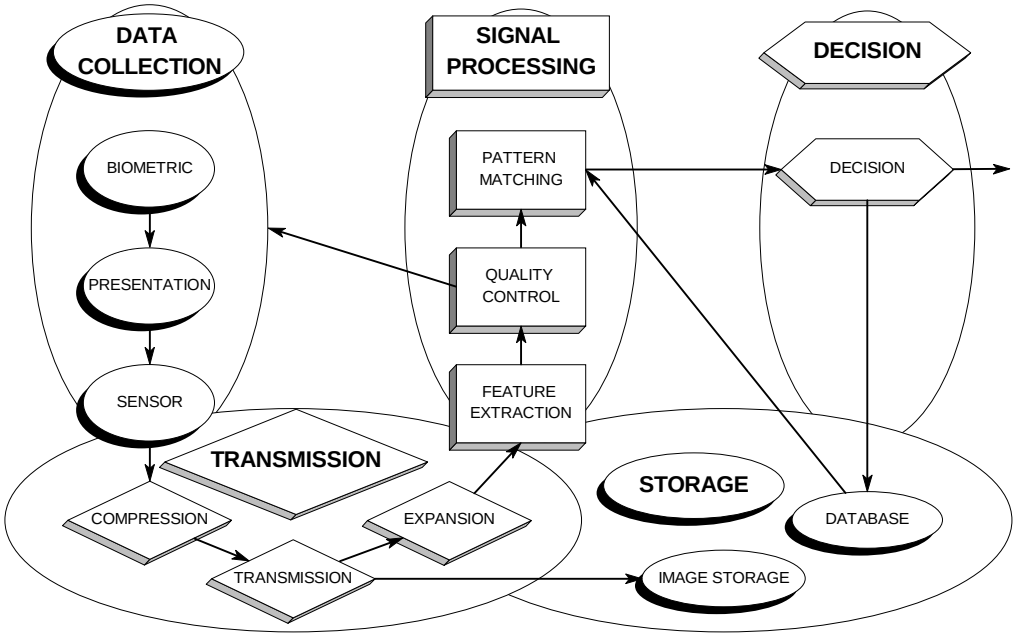


FIGURE 1: GENERIC BIOMETRIC SYSTEM

[2] The mathematical basis for this somewhat surprising statement linking distinctiveness to input variability is found in reference [9].

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 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).

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 deconvolve the true biometric pattern from the presentation and sensor characteristics also coming from the data collection subsystem, in the presence of the noise and signal losses imposed by the transmission process. 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 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 non-reversible 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 requirement for bandwidth.

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 and 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 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.

Decision

The decision subsystem implements system policy by directing the database search, determine "matches" or "non-matches" 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 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 or multiple "candidate" matches. 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 non-matches 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 non-match within the system. In any case, in the testing of biometric devices, it is necessary to decouple the performance of the signal processing subsystem from the policies implemented by the decision subsystem.

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 non-match 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 [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.

I. Testing

Testing of biometric devices requires repeat visits with multiple human subjects. Further, the generally low error rates mean that many human subjects are required for statistical confidence. Consequently, biometric testing is extremely expensive, generally affordable only by government agencies. Few biometric technologies have undergone rigorous, developer/vendor-independent testing to establish robustness, distinctiveness, accessibility, acceptability and availability in "real-world" (non-laboratory) applications. Over the last four years, the U.S. National Biometric Test Center has been focusing on developing lower cost testing alternatives, including testing methods using operational data and methods of generalizing results from a single test for performance prediction over a variety of application-specific decision policies.

Application Dependency of Test Results

All test results must be interpreted in the context of the test application and cannot be translated directly to other applications. 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 suited to decision policies yielding low error rates and high user acceptability. 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. Habituated volunteers, often "incentivized" employees (or students) of the testing agency, may be the most apt to see biometric systems as acceptable and non-intrusive.

Performance of a device at an outdoor amusement park [4] to assure the identity of non-transferable season ticket holders, for instance, cannot be expected to be the same

as in the laboratory. This use constitutes a cooperative, overt, non-habituated, unattended, non-standard environment, public, closed application. Performance in this application can only be predicted from measures on the same device in the same application. Therefore, as a long-term goal in biometric testing, we should endeavor to establish error rates for devices in as many different application categories as possible.

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[3]. 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 distance between samples to non-like 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 [9].

These distributions are shown as Figure 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) 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 [10].

[3] 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.

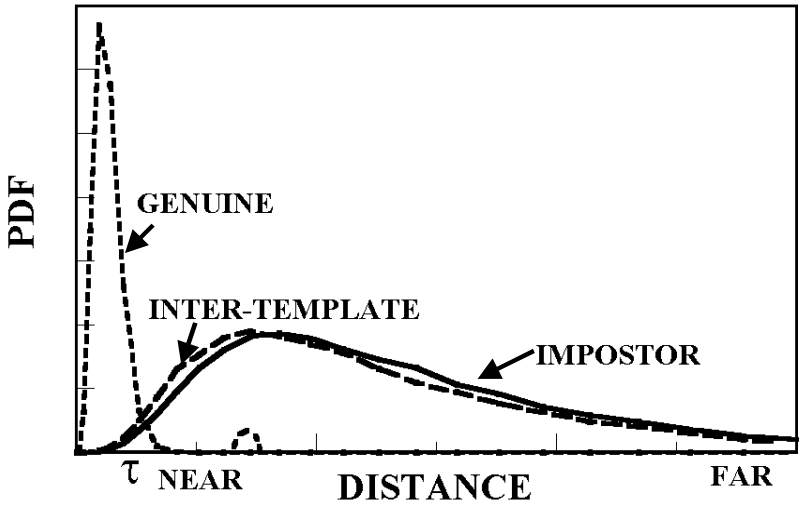


FIGURE 2: DISTANCE DISTRIBUTIONS

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 "non-match", 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 (non-overlapping) from the impostor distribution. Clearly, decreasing the difficulty of the application category will effect 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 non-cooperative applications, it is the goal of the deceptive user ("wolf") 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 non-cooperative system the extent to which "wolves" 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 non-match 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 non-dimensional number, if we are to

compare two unrelated biometric systems using a common and basic technical performance measure.

Non-Dimensional Measures of Comparison

The most useful method for removing the dimensions from the results shown in Figure 2 is to integrate the "impostor" distribution from zero to an upperbound, 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 non-match" at this threshold.

These two values, "false match" and "false non-match", for every τ, can be displayed as a point on a graph with the false match on the abscissa (x-axis) and the false non-match on the ordinate (y-axis). We have done this in Figure 3 for four Automatic Fingerprint Identification System (AFIS) algorithms tested against a standard database. For historic reasons, this is called the "Receiver Operating Characteristic" or ROC curve [11-13]. Mathematical methods for using these measured false match and false non-match rates for "false acceptance" and "false rejection" prediction under a wide range of system decision policies have been established in [8].

Other measures have been suggested for use in biometric testing [19], such as "D-prime"[20,21] and "Kullback-Leibler" [22] 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 non-match error rates, as displayed in the ROC curve, seem to be the only appropriate test measures allowing for even rudimentary system error performance prediction.

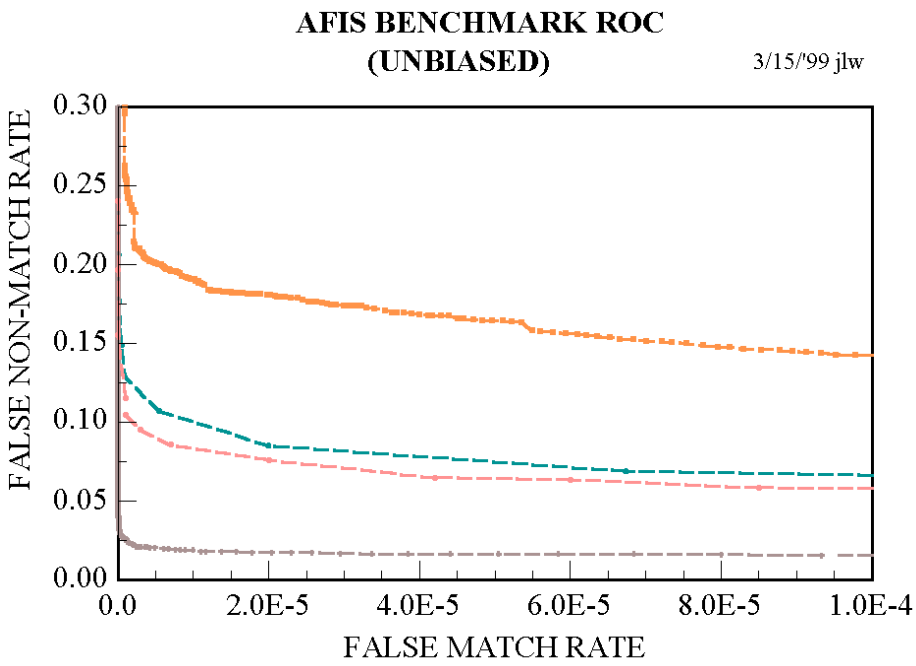


Figure 3

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 [14], error bounds for the ROC at each threshold, τ, have been found through a summation of the binomial distribution. The confidence, β, given a **non-varying** probability p, of K sample/template comparison scores, or fewer, out of N **independent** comparison scores being in the region of integration would be

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

Here, the exclamation point, called "factorial", indicates that we multiply together all integers from 1 to the number indicated. For instance, 3!=1x2x3=6. This number gets so huge so fast that 120! is too big for precise computation on most PCs. In most biometric tests, values of N and K are too large to allow N! and K! in equation (1) to be computed directly. The general procedure is to substitute the "incomplete Beta function" [15,16] 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 Law[4]" 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 [17].

Equation (1) will not be applicable to biometric systems if: 1) trials are not independent; 2) the error probability varies across the population. 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 (1) will not apply. An equation for error bounds in this case has been given by Bickel [18]. The varying error probability across the population ("goats" with high false non-match errors and "sheep" with high false match errors) similarly invalidates (1) as an appropriate equation for developing error bounds. Developing appropriate equations for error bounds under "real-world" conditions of non-independence of the comparisons and non-stationarity of the error probabilities is an important part of our current research.

The real tragedy in the break-down of equation (1) is in our inability to predict even approximately how many tests will be required to have "statistical confidence" in our results. We currently have no way of accurately estimating how large a test will be necessary to adequately characterize any biometric device in any application, even if error rates are known in advance.

In any case, we jokingly refer to error bounds as the "false sense of confidence interval" to emphasize that they refer to the statistical inaccuracy of a particular test owing to finite test size. The bounds in no way relate to future performance expectations for the tested device, due to the much more significant uncertainty regarding user population and overall application differences. We do not report error bounds or "confidence levels" in our testing.

Operational Testing

Given the expense of assembling and tracking human test subjects for multiple sample submissions over time, and the limited, application-dependent nature of the resulting data, we are forced to ask, "Are there any alternatives to laboratory-type testing?" 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 Figure 2.

The problem with operational data is in creating the impostor distribution. Referring to Figure 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 non-like 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

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

the inter-template distribution of Figure 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. 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. 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 [9].

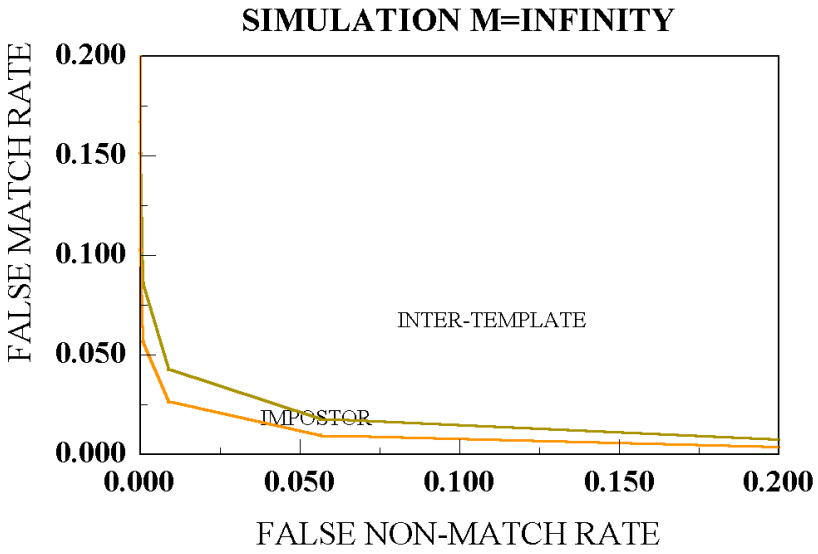


FIGURE 4: INTER-TEMPLATE ROC BIAS

VI. Test Design

So how can we design a test to develop a meaningful ROC and related measures for a device in a chosen application for a projected population? We need to start by collecting "training" and "test" databases in an environment that closely approximates the application and target population. This also implies taking training and test samples at different times to account for the time-variation in biometric characteristics, presentations and sensors. 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 effecting 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.

As explained above, we have no way to really determine the number of distance measures needed for the required statistical accuracy of the test. Suppose that, out of desperation, we accept equation (1) as an applicable approximation. One interesting question to ask is "If we have no errors, what is the lowest false non-match 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 equation 1 using X=0,

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

This reduces to

$$\ln(0.05) > N \ln(1-p) \qquad (3)$$

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

$$N > \frac{3}{p} \qquad (4)$$

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 non-match error rates to be less than one in one hundred (0.01), we will need to conduct 300 independent tests with no errors (3/300 = 0.01). Conducting 300 independent tests of will require 300 samples and 300 templates, a total of 600 patterns. Again, all of this analysis rests upon the questionable validity of the assumptions used to create equation (1).

We might ask, at this point, if it is necessary to have that many test users, or if a small number of users, each giving many samples, might be equivalent. Unfortunately, we require statistically "independent" samples, and no user can be truly independent from him/herself. Technically, we say that biometric data is "non-stationary", meaning that a data set containing one sample from each of one thousand users has different statistical properties than a data set containing one thousand samples from a single user. Ideally, we would require for our tests N different users, each giving one sample. In practice, we may have to settle for as many users as practicable, each giving several samples separated by as much time as possible. The impact of this on system performance prediction is also not known.

VII. Available Test Results

Most past tests have reported "false acceptance" and "false rejection" error rates based on a single or variable system policy. The U.S. National Biometric Test Center has advocated separating biometric performance from system decision policy, by reporting device "false match/ false non-match" rates, allowing users to estimate rejection/acceptance rates from these figures. We point out that some systems (access control) will "accept" a user if a match is found, while other systems (social service and driver's licensing) will "reject" a user if a match is found (during enrollment). Device false match/ false non-match performance may be the same in each system, but the decision policy will invert the measures of "false acceptance" and "false rejection". The reporting of results as a dimension-less Receiver Operating Characteristic (ROC) curve is becoming standard.

Results of some excellent tests are publicly available. The most sophisticated work has been done on speaker verification systems. Much of this work is extremely mature, focusing on both the repeatability of sounds from a single speaker and the variation between speakers [24-30]. The scientific community has adopted general standards for speech algorithm testing and reporting using pre-recorded data from a standardized "corpus" (set of recorded speech sounds), although no fully satisfactory corpus for speaker verification systems currently exists. Development of a standardized database is possible for speaker recognition because of the existence of general standards regarding speech sampling rates and dynamic range. The testing done on speech-based algorithms and devices has served as a prototype for scientific testing and reporting of biometric devices in general.

In 1991, the Sandia National Laboratories released an excellent and widely available comparative study on voice, signature, fingerprint, retinal and hand geometry systems [31]. This study was of 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 evaluation of the IriScan prototype [32] in an access-control environment.

A major study of both fingerprinting and retinal scanning, using people unacquainted with the devices and in a non-laboratory setting, was conducted by the California Department of Motor Vehicles and the Orkand Corporation in 1990 [7]. This report measured the percentage of acceptance and rejection errors against a database of fixed size, using device-specific decision policies, data collection times, and system response times. Error results cannot be generalized beyond this test. The report includes a survey of user and management acceptance of the biometric methods and systems.

The Facial Recognition Technology (FERET) Program has produced a number of excellent papers [33-36] since 1996, comparing facial recognition algorithms against standardized databases. This project was initially located at the U.S. Army Research Laboratory, but has moved now to NIST. This study uses as data facial images collected in a laboratory setting. Earlier reports from this same project included a look at infrared imagery as well [37].

In 1998, San Jose State University released the final report to the Federal Highway Administration [38] 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.

The existence of CardTech/SecurTech, in addition to other factors such as the general growth of the industry, has encouraged increased informal reporting of test results. Recent reports have included the experiences of users in non-laboratory settings [1-5].

VIII. 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.

IX. Bibliography

[1] Gail Koehler, "Biometrics: A Case Study – Using Finger Image Access in an Automated Branch", Proc. CTST'98, Vo. 1, pg. 535-541

[2] J.M. Floyd, "Biometrics at the University of Georgia", Proc. CTST '96, pg 429-230

[3] Brad Wing, "Overview of All INS Biometrics Projects", Proc. CTST'98, pg. 543-552

[4] Presentation by Dan Welsh and Ken Sweitzer, of Ride and Show Engineering, Walt Disney World, to CTST'97 , May 21, 1997.

[5] Elizabeth Boyle, "Banking on Biometrics", Proc.CTST'97, pg. 407-418

[6] D. Mintie, "Biometrics for State Identification Applications – Operational Experiences", Proc. CTST'98, Vol. 1, pg. 299-312

[7] Orkand Corporation, "Personal Identifier Project: Final Report", April 1990, State of California Department of Motor Vehicles report DMV88-89, reprinted by the U.S. National Biometric Test Center.

[8] J.L. Wayman, "Error Rate Equations for the General Biometric System", IEEE Automation and Robotics Magazine, March 1999

[9] J.L. Wayman, "Technical Testing and Evaluation of Biometric Identification Devices" in A. Jain, etal (eds), Biometrics: Personal Identification in a Networked Society, (Kluwer Academic Press, 1998)

[10] C. Frenzen, "Convolution Methods for Mathematical Problems in Biometrics", Naval Postgraduate School Technical Report, NPS-MA-99-001, January 1999

[11] Green, D.M. and Swets, J.A., Signal Detection Theory and Psychophysics (Wiley, 1966),

[12] Swets, J.A.(ed.), Signal Detection and Recognition by Human Observers (Wiley, 1964)

[13] Egan, J.P., Signal Detection Theory and ROC Analysis, (Academic Press,1975)

[14] W. Shen, etal, "Evaluation of Automated Biometrics-Based Identification and Verification Systems", Proc. IEEE, vol.85, Sept. 1997, pg. 1464-1479.

[15] M. Abromowitz and I. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", (John Wiley and Sons, New York, 1972)

[16] W.H. Press, et al, Numerical Recipes, 2nd ed., (Cambridge University Press, Cambridge, 1988)

[17] J. E. Porter, "On the '30 error' criterion", ITT Industries Defense and Electronics Group, April 1997, available from the National Biometric Test Center

[18] P. Bickel, response to SAG Problem #97-23, University of California, Berkeley, Department of Statistics.
[19] J. Williams, "Proposed Standard for Biometric Decidability", Proc. CTST'96, pg. 223-234
[20] Peterson, W.W. and Birdsall, T.G., "The Theory of Signal Detectability", Electronic Defense Group, U. of MI., Tech. Report 13 (1954)
[21] Tanner, W.P. and Swets, J.A., "A Decision-Making Theory of Visual Detection", Psychological Review, Vol. 61, (1954), pg. 401-409
[22] S. Kullback and R. Leibler, "On Information and Sufficiency", Annals of Mathematical Statistics, vol.22, (1951), pg. 79-86
[23] P. Bickel, response to SAG Problem #97-21, University of California, Berkeley, Department of Statistics.
[24] B. Atal, "Automatic Recognition of Speakers from Their Voices", Proc. IEEE, 64, (1976), pg 460-475
[25] A. Rosenberg, "Automatic Speaker Verification", Proc. IEEE, 64, (1976), pg. 475-487
[26] N. Dixon and T. Martin, Automatic Speech and Speaker Recognition (IEEE Press, NY, 1979)
[27] G. Doddington, "Speaker Recognition: Identifying People by Their Voices", Proc. IEEE, 73, (1985), pg 1651-1664
[28] A. Rosenberg and F. Soong, "Recent Research in Automatic Speaker Recognition" in S. Furui and M. Sondhi, eds, Advances in Speech Signal Processing (Marcel Dekker, 1991)
[29] J. Naik, "Speaker Verification: A Tutorial", IEEE Communications Magazine, (1990), pg. 42-48
[30] J.P.Campbell, Jr.,"Speaker Recognition: A Tutorial", Proc. IEEE, vol.85, September 1997, pg. 1437-1463
[31] J.P. Holmes, et al, "A Performance Evaluation of Biometric Identification Devices", Sandia National Laboratories, SAND91-0276, June 1991.
[32] F. Bouchier, J. Ahrens, and G. Wells, "Laboratory Evaluation of the IriScan Prototype Biometric Identifier", Sandia National Laboratories, SAND96-1033, April 1996
[33] P.J. Phillips, et al, "The FERET Evaluation Methodology for Face-Recognition Algorithms", Proc. IEE Conf.on Comp.Vis.and Patt. Recog., San Juan, Puerto Rico, June 1997
[34] S.A. Rizvi, etal, "The FERET Verification Testing Protocol for Face Recognition Algorithms", NIST, NISTIR 6281, October 1998
[35] P.J. Phillips, etal, "The FERET Evaluation" in H. Wechsler, etal (eds) Face Recognition: From Theory to Applications (Springer-Verlag, Berlin, 1998)
[36] P.J. Phillips, "The FERET Database and Evaluation Procedure for Face-Recognition Algorithms", Image and Vision Computing Journal, Vol. 16, No.5, 1998, pg. 295-306
[37] P.J. Rauss, et al, "FERET (Face-Recognition Technology) Recognition Algorithms", Proceedings of ATRWG Science and Technology Conference, July 1996
[38] J.L. Wayman, "Biometric Identifier Standards Research Final Report", October, 1997, sponsored by the Federal Highway Adminstration, downloadable from our web site at www.engr.sjsu.edu/biometrics/fhwa.htm

A Definition of "Biometrics"

James L.Wayman, Director
U.S. National Biometric Test Center

The purpose of this paper is to establish the meaning of "Biometric" as used by the National Biometric Test Center (NBTC) -- particularly important as our use of the term differs from the more common and historical meaning: "application of mathematical-statistical theory to biology[1]". We take "Biometric" as an abbreviation for "biometric authentication", a sub-field of the larger area of human identification science. Specifically, biometric authentication is

"the automatic identification or identity verification of an individual based on physiological and behavioral characteristics".

Several notable concepts seem to follow directly from this definition:

"automatic":

1) Identification methods requiring substantive levels of human intervention (beyond the simple act of the user in supplying the measured pattern) are not biometric authentication methods under this definition. At the current level of technology, this would exclude DNA, latent fingerprint, body fluid, hair, fiber and all other forms of crime scene analysis as biometric authentication techniques.
2) Computer-based pattern matching is at the core of all biometric authentication technologies. We know of no case in law where the courts have admitted the decision of a computer as an authoritative judge of human identity. Therefore, biometric authentication is not a forensic technology and is not designed to support forensic use. Forensic technologies are not within the scope of the NBTC. However, just as telephone and credit card records can be the subject of a court subpoena, we acknowledge that biometric logs and data could potentially be expropriated by the courts. Biometric authentication technologies are inherently no more of a forensic tool than telephones and credit cards. Concern and caution over the potential secondary forensic use of these technologies is within the scope of the NBTC.
3) The automatic pattern matching is always probabilistic and so decisions are always made with some level of uncertainty. Errors are made by these technologies. In applications where a machine error can result in the denial of service to a user, a method for human adjudication is always available. Human intervention for exception handling is within the definition of biometric authentication.
4) Implicit in the definition, preceding the word "automatic", is the phrase "non-physically invasive". Techniques involving the withdrawal of blood, for instance, even if fully automatic, would not be considered as a biometric authentication technique under this definition.

"identification or identity verification":

1) Biometric authentication has the capacity to connect a person through the measured characteristics to an identity as previously enrolled in a database. This technology cannot link a person to any identity outside the system.

---

[1] The Random House College Dictionary (Unabridged), revised edition, 1984

2) "Verification" means to assess the probable truth of the claim to an identity made by a system user. Voluntary use of biometric authentication technology by individuals to meet public and private requirements for identity verification is never invasive of privacy. Further, if offered as an alternative to more commonly used and privacy invasive techniques (such as recitation of mother's maiden name, date or location of birth, or production of a passport or standard driver's license), these technologies can be considered privacy enhancing. The use of biometric authentication technologies for privacy enhancement is of interest to the NBTC.
3) Not all biometric authentication methods are appropriate verification techniques for every individual, due to across-individual and within-individual variations in the measure. In application, non-biometric identity verification methods must always be available for those who cannot reliably use a chosen technique. Such capabilities will be considered as within the scope of the overall biometric system.
4) "Identification" means to probabilistically link a person to an enrolled record without an identity claim. A biometric system can be created for either "identification" or "identity verification". A single system cannot do both, but systems with each of these capabilities can be combined in an application. The scope of biometric authentication covers both verification and identification.
5) Although "identification" linkages can only be to enrolled records, such linkages can lead to loss of anonymity if the enrollment record contains information with meaning outside the system (legal or common name, for instance). Biometric technologies do not inherently require any information (even name) to be given at time of enrollment. Therefore, biometric technologies are neutral with regard to anonymity, but can be used in systems with enrollment protocols designed to either promote or prevent anonymity. Biometric authentication applies to both promoting and preventing anonymity in ways that are beneficial to societies and individuals.

"individual":

1) Implicit in this term is the word "human". Although automatic identification technologies can be used on animals, vegetables, and all range of objects, both man-made and natural, biometric authentication is only concerned with human identity.

2) Also implicit in this term is the word "living". Biometric authentication is not concerned with the identification of dead individuals.

3) The object of biometric authentication is the linkage of a singular human with an enrollment record from a singular human. Biometric authentication cannot identify groups or populations and it cannot identify individuals as members of groups or populations. These activities are not within the scope of biometric authentication.

"based on behavioral and physiological characteristics":

1) Biometric authentication devices rely on measures that are both behavioral and physiological. As both human behavior and physiology change over time and collection environment, biometric characteristics are not exactly repeatable.

2) Biometric authentication is "based" on these characteristics, but does not use them directly. Rather, the characteristics are generally sampled in some way[2], then acted on by a mathematical process to create of a series of numbers, called "features". Generally, the process cannot be inverted to reconstruct the raw sampled data from these numerical features[3]. The numerical features generally do not have a clear and direct mapping to any physiological characteristics and, consequently, contain no personal information. The purpose of the numerical "features" is only to allow a quantitative comparison of enrollment and sample record. Systems designed specifically to measure physiological features are outside the scope of biometric authentication.

3) The identification of medical conditions, age, race or gender is not possible from the numerical features[4] and is not an inherent part of biometric authentication. Systems which perform these functions are outside of this definition.

---

[2] We acknowledge the difficulty presented here by holographic fingerprint and other correlation techniques, both analog and digital.

[3] We acknowledge the difficulty presented here by the principal component analysis method of "eigenface" facial recognition systems.

[4] We acknowledge the difficulty presented here by spectrally-based speaker verification systems and the possibility of gender estimation from the spectral features.

Generalized Biometric Identification System Model

James L. Wayman, Director
U.S. National Biometric Test Center

Abstract

Commonly-held knowledge and oft-repeated descriptions of biometric identification systems are more complicated than necessary because of the lack of a generally accepted system model. A good example is the supposed difference between "identification" and "verification". In this paper, we will present a system model applicable to all biometric systems, showing the relationship between "verification" and "identification" and "enrollment" and "operation", illustrating where commonalities exist between seemingly disparate systems, and suggesting where interface standards might be useful. Further, we will develop a general mathematical formulation to show "verification" to be the degenerate case of M-to-N matching, with N = 1. Similarly, "PIN-less verification" will be recognized as the degenerate case with small N with M=1, and "identification" can be seen as the more general case with small M and large N.

Introduction

Despite the increasing literature on biometric identification technologies [1-18], including taxonomies of both applications and technologies [1,3], there has been no general description of the biometric system. Primary to the development of interface standards, standardized test methodologies and generalized performance equations, is an understanding of the normative system model. Such a model can illuminate the common structures and parallelisms between seemingly disparate methodologies. Certainly not all biometric technologies will fit any single model, but tremendous insight can be gained by noting where individual systems differ from the norm.

The General System Diagram

Figure 1 shows a system diagram of the general biometric system. Five sub-systems are shown: data collection, transmission, signal processing, storage and decision. To first order approximation only, these sub-systems can be considered independent, with errors introduced at each to be independent and additive. At a more comprehensive level of analysis, these sub-systems will not be independent, with errors impacting sub-system performance downstream. In testing biometric devices, it will generally be easiest to test sub-systems independently, when possible. In the following sections, we will describe each sub-system in detail.

The Data Collection Sub-System

The data collection sub-system samples the raw biometric data and outputs a one- or multi-dimensional signal. The input biometric pattern needs to be fairly stable over the time period of interest. Even if the fundamental, targeted patterns (finger ridges or iris patterns or retinal vasculation) are stable for life, injury or disease may cause changes in the observed patterns. Some patterns, such as voice, face or signature, simply drift, although perhaps reversibly.

Figure 2 shows general "genuine" and "impostor" distributions, widely discussed in the recent literature [1-4] and now part of the general body of knowledge of biometrics. The proximity to the origin of the "genuine" distance distribution is an

indication of the degree to which the biometric pattern is stable across the user population.

The biometric pattern is "presented" to a sensor, which transduces the pattern into an electronic signal. Most biometric systems expect a "standardized", predetermined presentation of the pattern. For instance, fingerprint systems expect the presentation to the scanner of the centered core of the print, with minimal rotation and with moderate pressure. Facial systems require a full facial front image; hand geometry requires fingers placed firmly against the post or posts; dynamic signature requires the habituated hand movement with a pen, perhaps on an imaging pad; eye scanning requires the presentation of the iris or retinal with minimal external lighting.

The location of the "genuine" distance distribution, shown in Figure 2, results from the convolution of both biometric pattern changes and changes in the presentation, a higher degree of standardization in the presentation resulting in greater proximity to the origin. We expect habituated users in office environments, for instance, to present the biometric with greater standardization than non-habituated users outdoors or difficult environmental conditions. Supervision of the presentation would also be expected to produce greater adherence to the presentation standard.

Voice systems and keystroke systems are the exception to the requirement for a predetermined presentation. Voice systems may prompt the user for the recitation of a random word or number, or may allow the user to select his/her own utterance. Keystroke systems work by statistical analysis of time-sampled input patterns determined by the user.

The sensors must be as stable as possible, meaning that we hope that all sensors in the system will have similar performance measures and, for a single sensor, re-calibration will not often be required. Sensor variability moves the "genuine" distribution away from the origin, increasing error rates. The Federal Bureau of Investigation (FBI) and the National Institute of Standards and Technology (NIST) established finger scanner "Image Quality Standards" (FBI/NIST Appendix G", IAFIS-IC-0010(V2), April 1993) precisely for the purpose of limiting sensor variation. Speaker recognition using telephones is particularly difficult because frequency response functions are so variable across handsets.

Some sensors have the capability of determining whether or not a signal of acceptable quality has been acquired. In other systems, quality control is exercised by the signal processing sub-system. The presence of automatic quality control at some stage of the process is not currently present in all biometric systems, at the expense of higher than necessary false non-match rates.

So, we can see that the output of the data collection sub-system is impacted by a number of sources of change, ultimately resulting in performance error. More specifically, the location of the "genuine" distribution results from the convolution of all error sources across the application. For all systems, the signal processing software is designed to compensate for these changes, as will be discussed. Minimization of data collection errors and the relationship of these errors to system error rate are completely uncharted areas of biometric research.

The Transmission Sub-System

Many biometric systems collect data in one physical location and process it in another. Fingerprints, voice data and facial images, submitted to a centralized system

may come from widely distributed sources. To minimize required transmission bandwidth, data compression may be required. The signal processing sub-systems downstream are designed to process the original image, so expansion of the compressed image is always required prior to processing. The compressed image may be most suitable for storage, however. Compression/expansion may be "lossy" or "loss-less" with regard to the quality of the expanded signal, but lossy compression algorithms generally result in much greater bandwidth reduction. Compression standards for "lossy" techniques exist for fingerprinting (Wavelet Scalar Quantization, FBI standard IAFIS-IC-0110v2, 1993), facial imaging (JPEG, as in Data Format Standard ANSI/NIST CSL-1-1993) and speech (Code Excited Linear Prediction, as in FPS-1014, for example), with each method designed specifically for the biometric signal of interest.

Compression/expansion algorithms are designed and tested to work well on signals of an expected quality. If the sensors transducing the original signal do not meet the expected technical quality requirement (lower sampling rate, for instance), losses during the compression/expansion process may be unpredictable. For instance, the quality loss caused by standard WSQ compression on fingerprints taken with non-"Appendix G" compliant scanners is currently unknown. Further complicating the question of quality loss is that technical measures of degradation (mean squared differences, greatest difference) may not directly relate to increases in system error rates. The relationship of sensor quality and compression to system error rates is completely unknown for any biometric system.

The transmission channel may also add noise, particularly if analog signals are used. Transmission channel response is particularly important in telephone-based speaker identification systems. Speaker identification algorithms exploit the stability of the channel, allowing deconvolution of the time-invariant channel characteristics.

The Signal Processing Sub-System

The signal processing sub-system takes the now degraded image of the original biometric pattern and converts it into a "feature vector[1]". The intent is to distill into this vector the information in the original biometric data that is time-, presentation-, sensor-, compression- and transmission-invariant. Clearly, these algorithms are always highly proprietary. They all share the property of being non-invertible, meaning that the original pattern cannot be reconstructed from the feature vector.

Systems not using quality control at the sensor can perform a quality analysis in the signal processing sub-system on the feature vector or on the received pattern. For instance in fingerprinting systems, if the processing results in an insufficient number of minutiae, a signal can be sent back to the sensor that a second sample should be collected.

The feature vector is passed to the pattern matching module, where it is compared to some number of stored feature vectors, one by one. This comparison results in a

[1] Generally, these features are indeed "vectors" in the mathematical sense, fingerprint minutiae being the notable exception.

numerical measure quantifying the degree of similarity or difference between the compared patterns. This value is sent on to the decision sub-system.

At this point, there emerges a difference between enrollment and operation in the activity of the signal processing subsystem. For enrollment, if the policy is to accept any enrolled pattern, the feature vector is passed directly into storage with no pattern matching. If the enrollment policy is to accept only enrolled patterns that do not closely match any already in the database, the pattern matching module will be called on to compare the feature vector with some limited portion of the database. All similarity measures developed during this process will be passed to the decision module. This process is also known as "identification" or sometimes, if operated over a small database, as "PIN-less verification". The only difference between this activity and the so-called "verification" operation is that for verification the searched portion of the da

tabase is confined to a single pattern, so only a single comparison measure is passed on to the decision module.

Additional information is often available to the system that can be used to limit the number of required comparisons. This information path is not indicated in Figure 1, as it can be highly variable between systems. In the case of "verification", the additional information may be a "claimed identity" in the form of a name or an identification number narrowing the comparison to a single stored pattern; or the database may be on a "smart card" containing but a single enrolled template. In the case of "identification", the number of required comparisons may be limited by external information, such as the age of the customer, or by information endogenous to the biometric sample itself, such as the fingerprint pattern type. In any case, the actual activity of the signal processing system is exactly the same: extraction of the feature vector, checking of quality, and comparison of the feature vector to some number of enrolled vectors.

The Storage Sub-System

We have already discussed the two types of data that might be stored in the storage sub-system: compressed biometric patterns and enrolled templates. Because the template extraction process is non-invertible, the original patterns must be stored if their recovery will ever be required. Also as previously discussed, to limit the number of comparisons, the templates may be partitioned in the database on the basis of additional information collected with the biometric data, such as gender.

The Decision Sub-System

The inputs to the decision sub-system are the measures resulting from each comparison of feature vector to stored template. The purpose of the decision sub-system is to invoke the system decision policy. If the measures indicate a close relationship between the feature vector and one of the compared templates, a "match" is declared. If none of the measures indicate similarity, a "non-match" is declared. Beyond this, the differences between system decision policies are so great that little, in general, can be said. Some systems used a fixed threshold for making the decision, shown as τ in Figure 2. Comparison measures less than the threshold lead to the declaration of a "match". Some systems use a variable threshold, with a different value stored with each enrolled template. Some systems require the submission of multiple biometric measures for all customers. Some require submission of additional samples from the same biometric measure if a "non-match' is declared, some changing the required threshold for the additional samples.

The system activities resulting from a "match" and "non-match" may also vary greatly according to application and decision policy. For instance, during enrollment in social service applications, a "non-match" results in the "acceptance" of an applicant into the system. For electronic benefits transfer in the same social service system, however, a successive number of "non-matches" results in the "rejection" of the customer requested funds withdrawal. In systems using biometric patterns subject to significant drift, the "matching" of a sample with an enrolled template may result in the updating of the template.

System Error Rate Equations

Based on the available comparison measures, the system decision policy occasionally results in errors, falsely matching or falsely non-matching samples and enrolled templates. Equations governing the system error rates are as varied as the system policies. Some general equations can be developed for the more common policies, however.

Consider the case where M independent biometric measures are used (in social services systems for instance, two index fingers are used). If the system decision policy is that a match will be declared only if all samples are matched to all enrolled templates (call this the "record") from the same person, than the probability of a false match , FMR_{SR}, for the comparison of samples against any single record can be given by

$$FMR_{SR} = \prod_{j=1}^{M} FMR_j(\tau_j) \tag{1}$$

where $FMR_j(\tau_j)$ is the single-comparison false match rate for the j^{th} biometric method, each of which might have a separate threshold, τ. Assuming that both threshold and single-comparison false match rate are independent of the enrolled record, the probability of not making any single-record false matches in comparing against multiple records can be expressed as

$$1 - FMR_{SYS} = (1 - FMR_{SR})^{P*N} \tag{2}$$

where FMR_{SYS} is the system false match rate, N is the number of records in the database, P is the percentage of the database to be searched on average (this is called the "penetration rate") based on the database partitioning. Explicit dependence on the threshold, τ , has been dropped for notational simplicity.

In the verification case, where N=1 and P=100%, equation (2) degenerates to

$$FMR_{SYS} = FMR_{SR} \tag{3}$$

meaning that the system false match rate is equal to the single-record false match rate, just as expected. In the doubly degenerate case where M=1, the system false match rate is just the single comparison false match rate.

Under the system decision policy being discussed, a single record false non-match does not occur if all biometric patterns are not falsely non-matched. This awkward syntax best expresses the relationship where $FNMR_{SR}$ is the single-record false non-match rate,

$$1 - FNMR_{SR} = \prod_{j=1}^{M} [1 - FNMR_j(\tau_j)] \quad (4)$$

$FNMR_j(\tau_j)$ are the single-comparison false non-match rates for the j^{th} biometric measure, which may have its own threshold, τ. Under the assumption that the database is "clean", meaning only one enrolled record for each customer, the system false non-match rate is the single-record false non-match rate. The absence of the variable N in (4), means that this equation can be used for verification (N=1), "PIN-less verification" (N>1) and identification (N=very large) cases.

We can develop in a similar fashion system error rate equations for a decision policy requiring only Q<M matches for a system "match" to be declared

Conclusions

In this paper, we presented a general system model intending to illustrate the commonality of most biometric systems, whether used for "identification" or "verification", "operation" or "enrollment". Generalized error rate equations were presented supporting the contention that verification is a degenerate case of the identification problem.

References

[1] J.L. Wayman, "A Scientific Approach to Evaluating Biometric Systems Using a Mathematical Methodology", Proc. CTST'97, pg. 477-492

[2] J.L. Wayman, "Benchmarking Large-Scale Biometric System: Issues and Feasibility", Proc. CTST Government'97, Sept. 1997

[3] J.L. Wayman, "The Science of Biometric Technologies: Testing, Classifying, Evaluating", Proc. CTST'97, pg. 385-394

[4] J.P. Campbell, "Speaker Recognition", Proc. of IEEE, vol. 85 no.9, Sept. 1997, pg.1437-1463

[5] Proc. of IEEE, Special Issue on Automated Biometric Systems, Sept. 1997

[6] R. Hopkins, "Benchmarking Very Large-Scale Identity Systems", Proc. CTST'97, pg. 314-332

[7] D.C. Bright, "Examining the Reliability of a Hand Geometry Identity Verification Device for Use in Access Control", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1987

[8] M. Fuller, "Technological Enhancements for Personal Computers", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1992

[9] S.C. Geshan, "Signature Verification for Access Control", Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1991

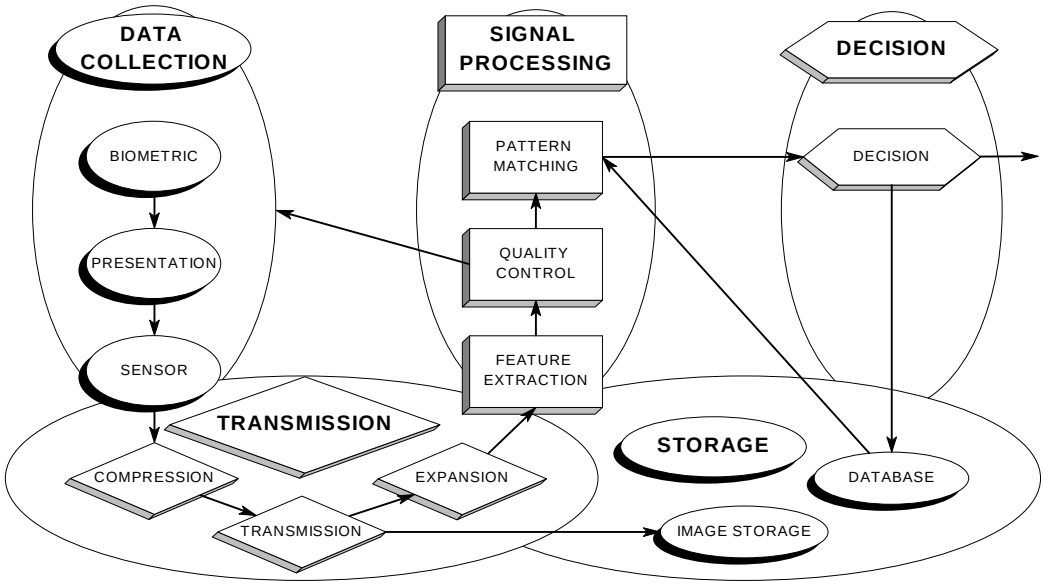
[10] D. Helle, "Examination of Retinal Pattern Threshold Levels and Their Possible Effect on Computer Access Control Mechanisms", Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1985

[11] G. Poock, "Fingerprint Verification for Access Control", Naval Postgraduate School Report NPSOR-91-12, Monterey, CA, April 1991

[12] G. Poock, "Voice Verification for Access Control", Naval Postgraduate School Report NPSOR-91-01, Monterey, CA, October 1990

[13] H. Kuan, "Evaluation of a Biometric Keystroke Typing Dynamics Computer Security System", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1992
[14] L. Tirado, "Evaluation of Fingerprint Biometric Equipment", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1991
[15] Orkand Corporation, "Personal Identifier Report", California Department of Motor Vehicles, DMV 88-89, May 1990
[16] J.P. Holmes, et al, "A Performance Evaluation of Biometric Identification Devices", Sandia National Laboratories, SAND91-0276, June 1991.
[17] F. Bouchier, et al., "Laboratory Evaluation of the IriScan Prototype Biometric Identifier", Sandia National Laboratories, SAND96-1033, April 1996.
[18] P.J. Phillips, et al, "FERET (Face-Recognition Technology) Recognition Algorithm Development and Test Results", Army Research Laboratory, ARL-TR-995, October 1996

Figure 1: The General Biometric System



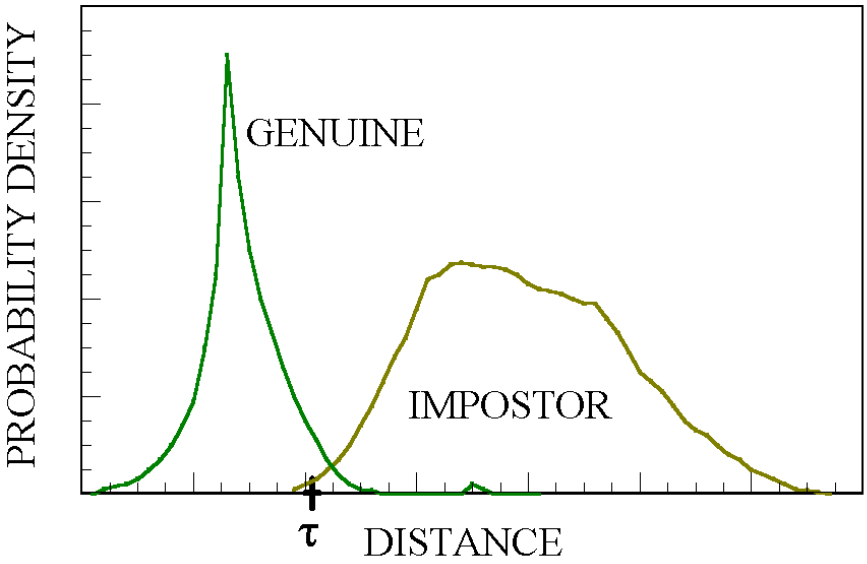


Figure 2: Genuine And Impostor Distance Distribution

Evaluation of the INSPASS Hand Geometry Data

James.L. Wayman, Director
U.S. National Biometric Test Center

I. INTRODUCTION

We have evaluated hand geometry biometric access templates and transaction records collected by the Immigration and Naturalization Service (INS) Passenger Accelerated Service System (INSPASS) at a test installation. The goal of this study was to establish false match and false non-match error rates (FMR and FNMR, respectively) and, ultimately, system false accept and false rejection rates as a function of decision threshold and to determine the effects of the threshold on system operation.

The INSPASS biometric system, operated by the contractor EDS, uses the "ID3D" hand geometry reader produced by Recognition Systems, Inc. (RSI). This application is classified as "cooperative, non-habituated, non-attended, standard environment" using the Test Center's taxonomy scheme and is the first study we know of for hand geometry in this application. This application is considered one of the most difficult for any biometric device. Consequently, we hypothesize that the FMR/FNMR will be far higher than reported in previous, laboratory studies [1].

INS, EDS and RSI cooperated tremendously with this effort: INS and EDS supplying the data and RSI supplying source code for evaluating that data.

II. DATA SELECTION AND "CONDITIONING"

The data supplied by INS/EDS were recorded between January, 1996 and February, 1997, at a INSPASS test kiosk at the Toronto International Airport. In all, 2946 templates and 9862 transactions were used from a much larger set of each supplied. A template is created by automatically measuring the geometry of the right hand of a customer with the ID3D unit three times during enrollment into the INSPASS system. The hand geometry is represented by a 9-byte vector of 9 integers in the range 0 to 255. This vector is, in turn, reduced in character size by encoding into a single 14 character glyph.

At the time of enrollment, each customer is given an INSPASS number. The templates for this study were selected such that no single person, as identified by this INSPASS number, is represented by more than one template. The purpose of this rule was to create a "clean" template base which will allow us to assess the statistics of the variation between hand geometries over the population.

A "transaction" may occur when a customer, already enrolled in the INSPASS system, presents him/herself at the test kiosk with an INSPASS card. A transaction is said to have occurred if there is a successful reading of the INSPASS card magnetic stripe, followed by the successful acquisition of the hand geometry by the ID3D unit. The successful acquisition is defined by a "Biofail" score of 0 on the transaction record. A successful acquisition only indicates that the ID3D unit received a signal; it does not indicate that the signal was successfully matched.

The transactions were selected such that no person, as identified by the INSPASS number, is represented by more than one transaction. The purpose of this rule is to avoid correlation problems caused by multiple sequential attempts by the same person and is in recognition of the fact that general biometric data is "non-ergodic". By non-ergodic, we mean that the biometric usage statistics (such as those of matching and non-matching

probabilities of presented sample and template patterns) of an entire population are not represented by the statistics of any single user. By accepting only one transaction per user, we obtain an "ergodic" database, divisible into subsets of only trivial statistical differences. When multiple transactions or templates were available, the first one chronologically was accepted.

Implicit in this analysis are the untestable assumptions that: 1) the database creation software accurately records customer transactions; 2) all usage is by "genuine" customers with no impostor transactions; 3) no customers are enrolled under multiple identification numbers.

III. THE TEMPLATES

The template data was made available by EDS in Microsoft (MS) Access format. Using MS-ACCESS, we ordered the column of templates. This column contains those passed from the central storage location to the hand geometry unit upon a successful card swipe by a customer. We make the unproven assumption that if the templates are received at all from central storage, they are received correctly. These centrally stored templates are adaptively updated after every successful use of the system by the customer. The number of times each template has been updated is not recorded by the system. This will strongly limit our ability to estimate the ROC, as will be seen.

Ordering of the templates allowed us to immediately remove any that were missing or listed as null. With these removed, the data was ordered according to the INSPASS identification number of the card holder. This allowed us retain the first record and to remove from the database all multiple templates of the same customer. This process left us with 2946 templates.

IV. THE TRANSACTIONS

Ideally, we would like the 9-byte feature vector resulting from each hand sample given during a transaction. Unfortunately, this information is not stored by the system and, consequently, was not available to us. The available information from each transaction is the "BioScore", indicating the extent to which the presented hand geometry sample matched the centrally stored template for the customer. We assumed that no customer was attempting to defraud the system by using an INSPASS card not genuinely issued to that customer and that each customer legitimately desired to be correctly recognized by the system. Again using MS-ACCESS, we ordered the data by the "CardFail" column. All records resulting from the failure of the card reader to acquire a signal were eliminated.

The data was then ordered by the "BioFail" column. Approximately 7% of the attempted transactions resulted in the failure of the biometric system to acquire a signal. This may be due to the system "timing out" after a customer does not present a hand, or it may result from the hand being incorrectly placed in the ID3D unit. The ID3D unit is created to indicate to a customer when the hand has been correctly placed in the unit. Therefore, a "BioFail" signal is considered for this study to be "user error", not "biometric identification error". Once ordered by the "BioFail" column, all records connected with a failure to acquire the biometric signal were removed from the database.

Finally, the data was ordered by INSPASS identification number, allowing the retention of the first transaction and removal of multiple transactions by the same customer. The "BioScore" of these remaining transactions were extracted from the database.

V. ANALYSIS METHODOLOGY

The first task was to create a histogram from the 1769 transaction "BioScore" measures, appropriately normalized to unit area. These scores are related by RSI-proprietary software to the distance in 9-space between a customer's presented hand geometry sample and the template stored for that customer. The data was "low pass filtered" by creating the histogram with about 50 bins of width 2 units, then linearly interpolating to create about 100 bins of width 1 unit. This histogram is a model of the underlying distance probability distribution, which we call the "genuine" distribution when the distances measured are between samples and their true templates

The next task was to convert the study template database from the 14 character glyph to the 9 integer vector. This was accomplished using the source code supplied by RSI. At this point, we would like to create an "impostor" score histogram by comparing all transaction feature vectors to all non-self templates. Unfortunately, this is not possible because the transaction feature vectors were not stored.. However, we can establish an inter-template histogram by computing the score between each template pair using the RSI-supplied source code. We modified the RSI code to remove a time-saving provision aborting calculation of very large distances. The score function is symmetric (the score from template A to B is the same as the score from template B to A), allowing us to use all 2946 templates in ½*2946*2946=4,379,850 comparisons. The data was binned into 256 bins of 1 unit width each. No further low pass filtering was deemed necessary. Figure1 shows the genuine and inter-template histograms.

Evaluation of the false non-match error rate as a function of threshold can be done entirely with the genuine probability distribution function as approximated by the "BioScore" histogram. Evaluation of the false match error rate as a function of threshold requires the unavailable "impostor" score histogram. Under the assumption of isotropic distribution (the spread of the data assumed the same in each of the 9 template components) of the 9-dimensional distance data, we believe that the "impostor" probability distribution can be obtained from the convolution of the radial genuine and inter-template distributions as estimated from their histograms. (editor's note: Convolution of these radial distributions (functions only of a one-dimensional distance) was later discussed by Franzen [2]). This was not attempted in this study. Rather, we looked at the question, "How well can the inter-template histogram be used as a proxy for the impostor distribution?".

VI. SIMULATION MODEL

To determine the effect of using the inter-template score histogram as a proxy for the impostor histogram, we created a simulation model. We started with the random selection from the experimental data set of 300 nine-dimensional template vectors. We took these as our "anchors". Around each of these anchors in 9-space, we created a 150 simulated samples by adding a gaussian variable to each component using the isotropic assumption (identical variance of each component gaussian error model). We have no information available upon which to evaluate these distributional assumptions, but the error variance was set so that the "genuine" histogram would look approximately like that of the INSPASS histograms of Figure 1. Three of these samples at each anchor were added to create a simulated template. The "genuine" score distribution was calculated by comparing with the RSI algorithm the simulated samples to simulated "self" templates. The inter-template score histogram was also computed by comparing simulated

templates. The sample vector (genuine) histogram was computed using the RSI algorithm for score assessment. The comparison scores between the sample vectors and randomly chosen "non-self" templates were also computed and used to create an "impostor" histogram.

This simulation was repeated, this time using the "anchors" as the templates, in effect simulating the addition of an infinite number of randomly generating samples in the template creation process. Figure 2 shows the genuine, impostor and inter-template histograms resulting from this latter simulation. The difference between the impostor and inter-template histograms is readily apparent.

Figure 3 shows the two ROC curves resulting from use of the simulated "impostor" and "inter-template" histograms for the first simulation and Figure 4 shows those for the second simulation. These figures clearly show that the "inter-template" not a good proxy for the "impostor" histogram, grossly overestimating error rates. The overestimate is worse when an infinite number of samples are used for the histogram than when three are used.

Returning to the real INSPASS data, the ROC estimated using the "inter-template" distribution is given in Figure 5. Based on the analysis of the simulation results, Figure 5 should be presumed to represent a gross upper bound on the true ROC, which is certainly lower by some unknown amount which will depend, upon other things, on the number of samples used in each updated template. The "equal error rate" of about 3% shown in Figure 5 is an overestimate of the true value.

VII. SYSTEM PERFORMANCE

The INSPASS system makes "accept/reject" decisions on the basis of a maximum of three entry attempts by the customer. What effect might this decision policy have on the system? A false rejection by the system requires three consecutive false non-matches. If the scores over multiple sequential attempts are independent, then the system false rejection rate (FRR) can be given by

$$FRR = FNMR^3 \quad (1)$$

Are additional samples following false non-matches independent? In Figure 6, we show the "genuine" histograms for all first attempts, all second attempts by those that failed the first attempt, and all third attempts by those that failed the first two attempts. If the attempts are independent, than the distributions will be nearly identical (within limitations of the small sample size). We were rather surprised by the similarity of the three distributions, although their movement to the right indicates an increasing false non-match rate with subsequent tries after failures. Consequently, equation 1 will underestimate the system false rejection rate for a "three-strikes" policy, but can be used as a lower bound on the true system false rejection rate. Because the biased ROC represents an upper bound on the system false non-match rate, we cannot combine these upper and lower bounds in any meaningful way. It is clear, however, that the false non-match rate will over estimate the "three strikes" system false rejection rate.

How is the system false acceptance rate affected by a "three strikes" policy, if impostors have up to three attempts to obtain a false acceptance? Being correctly rejected on three attempts requires being correctly not falsely matched on all three tries. Assuming independence of attempts, the probability of being correctly not matched three times is $(1\text{-}FMR)^3$. Consequently, the system false acceptance rate (FAR) can be given

$$1 - \mathrm{FAR} = (1 - \mathrm{FMR})^3 \tag{2}$$

Sequential attempts by the same impostor may not be independent, particularly if the impostor can learn from the attempt. Consequently, this value will serve as a lower-bound on the system false acceptance rate. Again, we can't combine the upper bound of the biased false match rate with the lower bound of equation (2) in any meaningful way. As an unsubstantiated guess, we might conclude that the FMR might be used as the system FAR.

VIII. OTHER PERFORMANCE MEASURES

Two other performance measures were applied to the biased data.

The first is "d-prime" defined as

$$\mathrm{d}' = \frac{\mu_1 - \mu_2}{\left((\sigma_1^2 + \sigma_2^2)/2\right)^{1/2}} \tag{3}$$

where μ_1 and μ_2 are the means of the genuine and inter-template distributions and σ_1 and σ_2 are their standard deviations. This measure was computed to be 2.1 for this data.

The second is the Kullback-Leibler measure, defined as

$$\mathrm{KL} = \int_0^\infty p_{\mathrm{IMP}}(\tau) \ln \frac{p_{\mathrm{GEN}}(\tau)}{p_{\mathrm{IMP}}(\tau)} \, \mathrm{d}\tau \tag{4}$$

where $p_{IMP}(\tau)$ is the impostor distribution, $p_{GEN}(\tau)$ is the genuine distribution and is τ the distance threshold. This measure was computed to be -6.4 for this data.

IX. CONCLUSIONS

This paper has presented a biased estimate of the INSPASS ROC for a test system, overestimating the true false match and false non-match error rates by an unknown amount. The estimate was the best that can be done from the available data, because the 9-byte transaction feature vectors were not saved by the system. Simulation shows that the extent of the bias will depend upon the number of times the templates have been updated. Consequently, Figure 5 will be seen as an overestimate of errors. Computation of the system false rejection and false acceptance rates require additional data or assumptions not available to us. The best we can guess with the given data is that the equal error rate of the INSPASS test system may be less than 3%.

X. REFERENCES

[1] J.P. Holmes, et al, "A Performance Evaluation of Biometric Identification Devices", Sandia National Laboratories, SAND91-0276, June 1991.

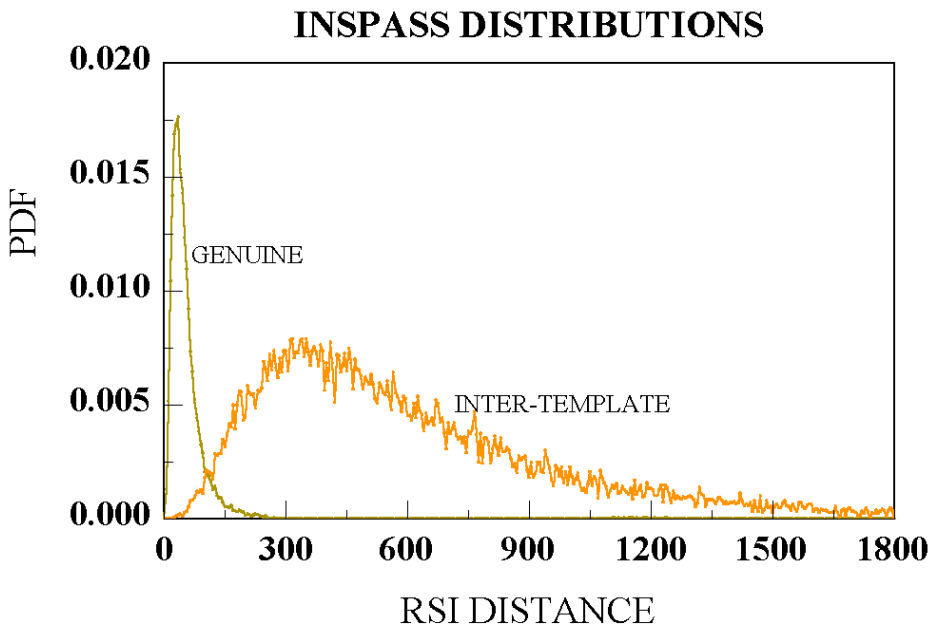


FIGURE 1

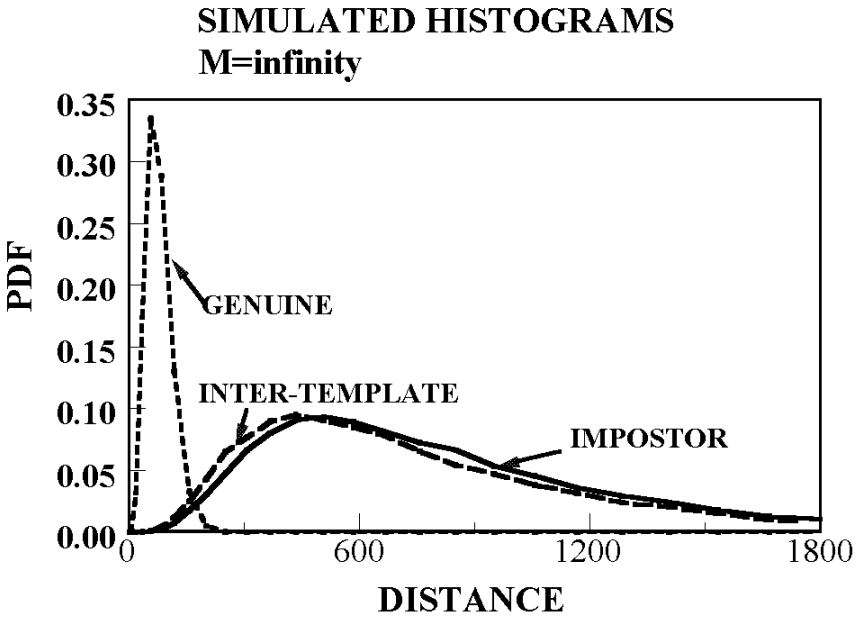


FIGURE 2

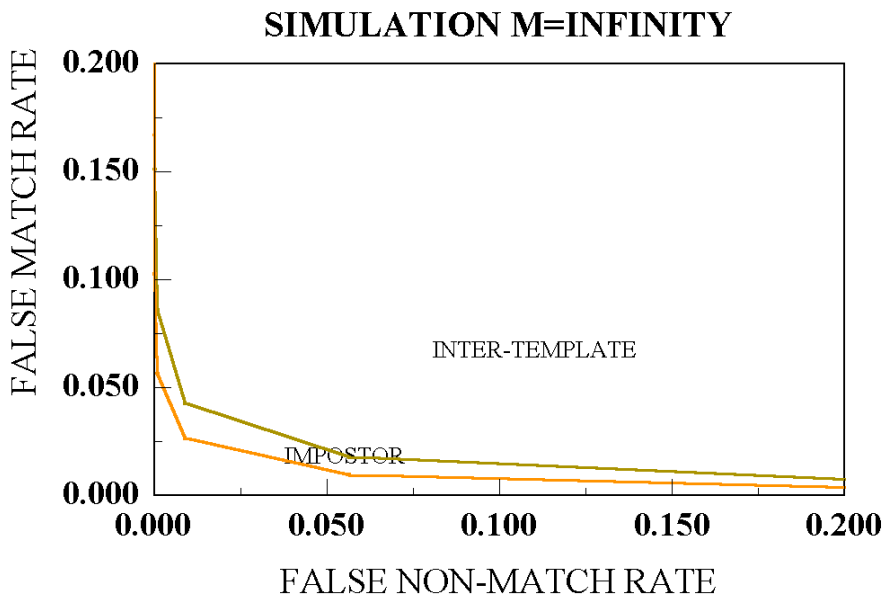


FIGURE 3

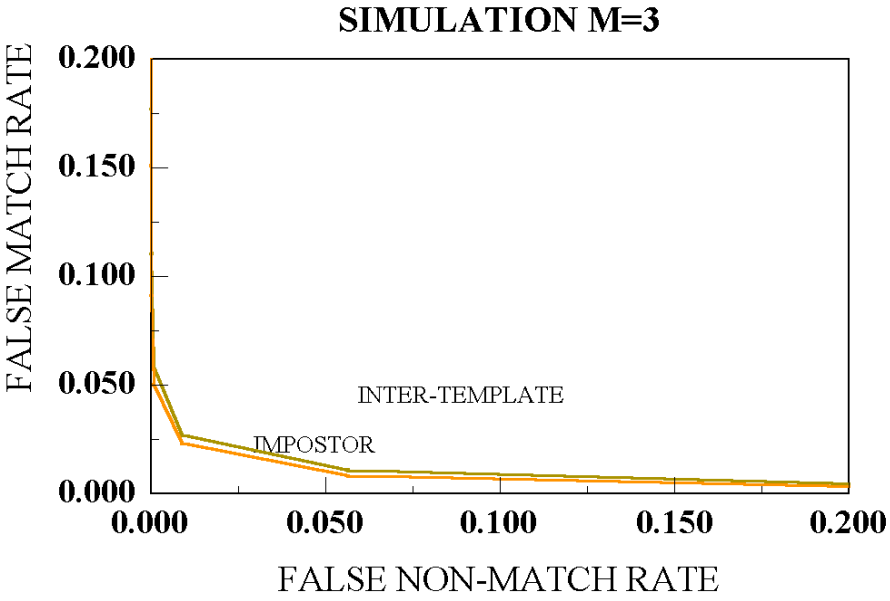


FIGURE 4

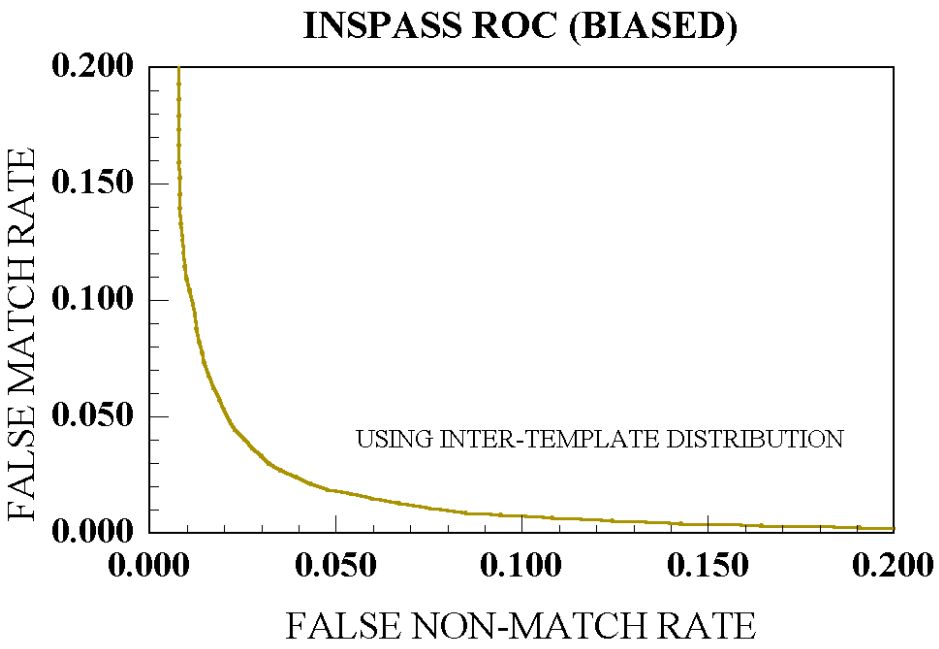


FIGURE 5

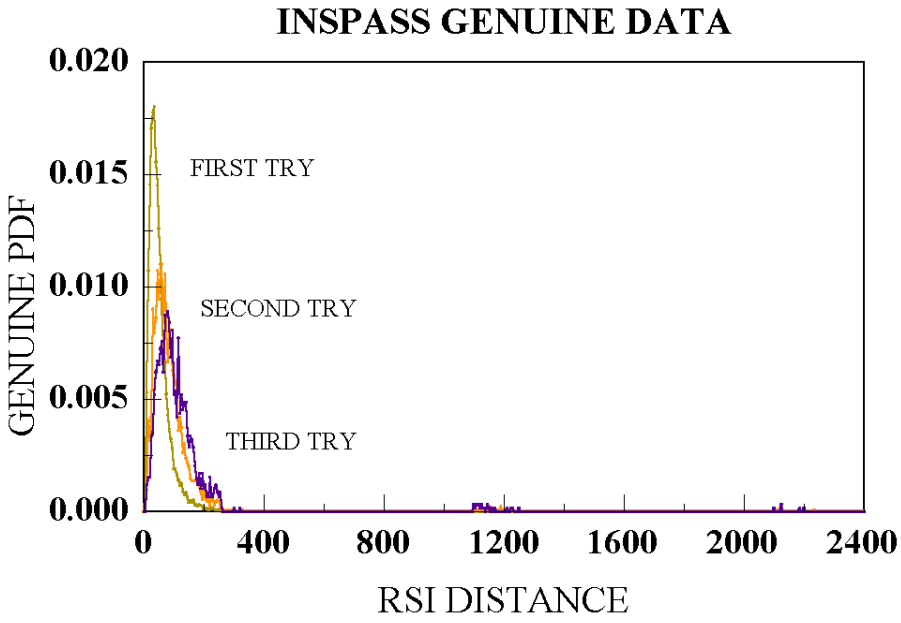


Figure 6

SAG Problem 97-2-1

Peter Bickel
Department of Statistics
University of California, Berkeley

Editor's note: This paper concerns the estimation of the "impostor" distance distribution when only the "genuine" and "inter-template" distributions are known. The problem originally posed at SAG-91-2-1 was:

Given K samples from each of M isotropic (spherically symmetric), identical distributions in N space. Each distribution has a centroid which can be computed from the K samples. Assume that these centroids are isentropically distributed (fill the space like a gas) in a bounded region of the N space. Now assume that we know the one-dimensional distance pdf's of both the "sample" and "centroid" distributions and that K is large. Can we calculate the distance distribution between any single sample and the collection of centroids from this information alone? Convolution of the one-dimensional distance distributions yields an incorrect answer. It appears that Fourier convolution is appropriate if a spherical Bessel function is used in the transform kernel. If each centroid is constructed from a single sample (K=1), then the single sample to centroid distribution is identical to the centroid distribution. How is the number of samples, K, incorporated into the correct answer?

As rephrased we give the following formulation of the problem. For each of M+1 individuals, K measurements measurements X_{ij}, i = 1, …, M+1, j = 1, …, K are taken. The X_{ij} are N vectors. The implicit assumptions seem to be

$$X_{ij} = \mu_i + \varepsilon_{ij} \tag{0}$$

where $\mu_1, \cdots, \mu_{M+1}$ are individual effects i.i.d. N vectors from some distribution spherically symmetric about some $\mu \in R^N$ and ε_{ij} are i.i.d. independent of the μ_i and spherically symmetric about 0. The data retained are $\left|\bar{X}_i - \bar{X}_{i'}\right|$ and $\left|X_{M+1,j} - X_{M+1,\cdot}\right|$, $1 \le i < i' \le M$, $1 \le j \le k$, where $|x|$ is the length of x and $\bar{X}_i = \frac{1}{K}\sum_{j=1}^{K} X_{ij}$. Let p be the density of $\left|\bar{X}_1 - \bar{X}_2\right|$ and q that of $\left|X_{M+1,1} - X_{M+1,\cdot}\right|$, and r the density of $\left|X_{M+1,1} - \bar{X}_1\right|$. The problem as we see it is that of estimating r given estimates $\hat{p}$ and $\hat{q}$. I will not dwell further on estimation of p and q which are densities on $[0,\infty)$. We expect $p^{(j)}(0) = q^{(j)}(0) = 0$, $0 \le j \le N-1$, where (j) denotes the j^{th} derivative – see (6) below. However, kernel estimation taking edge effects into account as discussed in say Fan and Gijbels (1996) Local Polynomial Modelling and its Applications is probably good. We turn to the relation between p,q, and r.

Write

$$X_{M+1,1} - \bar{X}_1 = \varepsilon_{M+1,1} + \mu_{M+1} - \mu_1 - \bar{\varepsilon}_1 \tag{1}$$

$$\bar{X}_{M+1} - \bar{X}_1 = \bar{\varepsilon}_{M+1} + \mu_{M+1} - \mu_1 - \bar{\varepsilon}_1 \tag{2}$$

$$X_{M+1,1} - \bar{X}_1 = \varepsilon_{M+1,1} - \bar{\varepsilon}_{M+1} \tag{3}$$

Evidently, (2) + (3) = (1). Further, (2) and (3) have spherically symmetric distributions. Since $|(2)|$ has density p, $|(3)|$ has density q and $|(1)|$ has density r, it would appear that what is wanted is a formula for the density of the formula of the length of the convolution of two spherically symmetric (about 0) distributions given the density of their lengths. This is not quite right. (2) and (3) are not independent since $\varepsilon_{M+1,1} - \bar{\varepsilon}_{M+1}$ and $\bar{\varepsilon}_{M+1}$ are only uncorrelated, and in general independent only if ε_{11} has a Gaussian distribution. However, we are told that K is large. Write $Z_K = (K-1)^{-1/2} \sum_{j=2}^{K} \varepsilon_{M+1,j}$ and $U_K = \varepsilon_{M+1,1}\left(1 - \frac{1}{K}\right)$. Then, Z_K and U_K are independent and, if K is large,

$$\varepsilon_{M+1,1} - \bar{\varepsilon}_{M+1} = U_K - \frac{(K-1)^{1/2}}{K} Z_K \approx U_K \tag{4}$$

and

$$K^{1/2}\bar{\varepsilon}_{M+1} = \left(\frac{K-1}{K}\right)^{1/2} Z_K + U_K \frac{K^{1/2}}{(K-1)} \approx Z_K \tag{5}$$

It therefore seems reasonable to ignore the dependence between (4) and (5). The solution to the problem lies in the formulae we derive below.

The U be an isotropically distributed N vector with density f on R^N. Let g be the corresponding density on R^+ of the distance length $|U|$ where $|x|^2 \equiv \sum_{i=1}^{N} x_i^2$, $x \equiv (x_1, \cdots, x_N)$. Then it is well known that

$$f(u) = |u|^{-(N-1)} c_N^{-1} g(|u|) \tag{6}$$

where c_N, the surface of the unit sphere in R^N

$$c_N = \frac{2\pi^{N/2}}{\Gamma\left(\frac{N}{2}\right)}$$

This follows, for instance, by equating

$$f(x)dx = h(|x|)|x|^{N-1}\, dS\, d|x|$$

where dS is the surface element on the unit sphere and $f(x) = h(|x|)$ by the isotropicity assumption and then integrating out dS to obtain g.

Now let U,V be independent random N vectors with isotropic distributions having densities f_1, f_2 on R^N respectively. Let g_1, g_2 be the corresponding densities on R^+ of $|U|, |V|$. Let f be the density of U + V and g be the density of $|U+V|$. Then,

$$f(y) = \int_{R^N} f_2(y-x)\, f_1(x)\, dx$$

$$= \int_{R^N} |y-x|^{-(N-1)}\, c_N^{-1} g_2\left(|y-x|\right) |x|^{-(N-1)}\, c_N^{-1} g_1\left(|x|\right) dx \qquad (7)$$

Hence,

$$g\left(|y|\right) = c_N^{-1} |y|^{N-1} \int_{R^N} |y-x|^{-(N-1)} |x|^{-(N-1)}\, g_2\left(|y-x|\right) g_1\left(|x|\right) dx \qquad (8)$$

This formula may be simplified by changing variables orthogonally from x to w, where $w_1 = \dfrac{(x,y)}{|y|}$. Then, (8) becomes

$$g\left(|y|\right) =$$

$$c_N^{-1} |y|^{N-1} \int_{R^N} \left(|y|^2 - 2w_1|y| + \sum_{i=1}^{N} w_i^2 \right)^{-\frac{(N-1)}{2}} g_2\left(\left(|y|^2 - 2w_1|y| + \sum_{i=1}^{N} w_i^2 \right)^{\frac{1}{2}} \right) |w|^{-(N-1)}\, g_1\left(|w|\right) dw \qquad (9)$$

Let

$$A(|y|, v, g_1, g_2) =$$

$$\int_{-\infty}^{\infty} \left(|y|^2 - 2w_1|y| + w_1^2 + v^2 \right)^{-\frac{(N-1)}{2}} g_2\left(\left(|y|^2 - 2w_1|y| + w_1^2 + v^2 \right)^{\frac{1}{2}} \right) g_1\left(\left(w_1^2 + v^2 \right)^{\frac{1}{2}} \right) \left(w_1^2 + v^2 \right)^{\frac{-(N-1)}{2}} dw_1$$

Then, for $N \geq 2$, by changing to spherical coordinates

$$g\left(|y|\right) = |y|^{N-1}\, c_N^{-1}\, c_{N-1} \int_0^{\infty} v^{N-2} A\left(|y|, v, g_1, g_2\right) dv \qquad (10)$$

From (10) we can derive the solution to the problem initially posed. Let g_1 by p and g_2 be q in our original notation. Then r is (neglecting the dependence between $\varepsilon_{M+1,1}$ and $\bar{\varepsilon}_{M+1}$) given by

$$r\left(|y|\right) \simeq |y|^{N-1}\, c_N^{-1}\, c_{N-1} \int_0^{\infty} v^{N-2} A\left(|y|, v, g_1, g_2\right) dv \qquad (11)$$

Note that we need to plug in estimates of $\hat{p}, \hat{q}$ to obtain $\hat{r}$. Although K and M do not appear explicitly in (11), biases of $\hat{p}$ and $\hat{q}$ due to their being density estimates translate into biases of $\hat{r}$ even as an estimate of the right-hand side of (11). These will depend on M, K and the true p and r. Further, variances of $\hat{p}$ and $\hat{q}$ translated by the delta method to $\hat{r}$ will also depend on M, K and the true distribution of X in a complicated way.

References

Fan and Gijbels (1996). Local Polynomial Modelling and its Applications. Chapman and Hall.

Convolution Methods for Mathematical Problems in Biometrics

C.L. Frenzen
Department of Mathematics
Naval Postgraduate School

1. Introduction

The problem we shall investigate can be formulated in the following way, due to Peter Bickel [1]: For M+1 individuals, K measurements X_{ij}, i = 1, …, M+1, j = 1, …, K are made. The X_{ij} are assumed to be vectors in $\Re^N$.

We assume, following Bickel [1], that

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

where $\mu_1, \cdots, \mu_{M+1}$ are individual effects vectors from $\Re^N$ which are i.i.d. (independently and identically distributed) with a distribution which is spherically symmetric about some μ in $\Re^N$. The ε_{ij} are i.i.d. independently of the μ_i and their distribution is spherically symmetric about 0.

The K(M+1) X_{ij} s represent K biometric measurements made on M+1 individuals; however the measuring device actually records only the data $\left|\bar{X}_i - \bar{X}_{i'}\right|$ and $\left|X_{M+1,j} - \bar{X}_{M+1}\right|$, $1 \le i < i' \le M$, $1 \le j \le k$, where $|x|$ is the length of x and

$$\bar{X}_i = \frac{1}{K}\sum_{j=1}^{K} X_{ij}$$

is the centroid of the K measurements taken on the the i^{th} individual.

Let p be the density of $\left|\bar{X}_1 - \bar{X}_2\right|$ and q the density of $\left|X_{M+1,1} - \bar{X}_{M+1}\right|$, and r the density of $\left|X_{M+1,1} - \bar{X}_1\right|$. The basic problem is to estimate r given estimates $\hat{p}$, $\hat{q}$ for p and q. As p and q are densities on $[0,\infty)$, the problem of estimating $\hat{p}$ and $\hat{q}$ is a standard problem in estimation theory which we do not consider further.

2. Bickel's Approach

What is the relationship between p, q, and r? Following Bickel [1], we write

$$X_{M+1,1} - \bar{X}_1 = \varepsilon_{M+1,1} + \mu_{M+1} - \mu_1 - \bar{\varepsilon}_1 \tag{1}$$

$$\bar{X}_{M+1} - \bar{X}_1 = \bar{\varepsilon}_{M+1} + \mu_{M+1} - \mu_1 - \bar{\varepsilon}_1 \tag{2}$$

$$X_{M+1,1} - \bar{X}_1 = \varepsilon_{M+1,1} - \bar{\varepsilon}_{M+1} \tag{3}$$

Note that addition of equations (2) and (3) yields equation (1). Further, the quantities on the left sides of (2) and (3) have spherically symmetric distributions. As the lengths of the left sides of (2) and (3) have densities p and q respectively, and the length of the left side of (1) is r, it seems that what is required is a formula for the density of the length of the convolution of two spherically symmetric (about 0) distributions given the densities of their lengths. However, as Bickel pointed out in [1], this is not quite correct

since the use of the convolution assumes the independence of the densities p and q, and these two densities are not generally independent, since in the right sides of (2) and (3) the terms $\varepsilon_{M+1,1}-\bar{\varepsilon}_{M+1}$ and $\bar{\varepsilon}_{M+1}$ are only uncorrelated, and not in general independent unless ε_{11} has a Gaussian distribution.

However, for large K, we follow Bickel's argument in [1] to show that the terms $\varepsilon_{M+1,1}-\bar{\varepsilon}_{M+1}$ and $\bar{\varepsilon}_{M+1}$ are independent. Let

$$Z_K = (K-1)^{-1/2}\sum_{j=2}^{K}\varepsilon_{M+1,j} \tag{4}$$

and

$$U_K = \varepsilon_{M+1,1}\left(1-\frac{1}{K}\right) \tag{5}$$

Then, Z_K and U_K are independent and, if K is large,

$$\varepsilon_{M+1,1}-\bar{\varepsilon}_{M+1} = U_K - \frac{(K-1)^{1/2}}{K}Z_K \approx U_K \tag{6}$$

and

$$K^{1/2}\bar{\varepsilon}_{M+1} = \left(\frac{K-1}{K}\right)^{1/2} Z_K + U_K\frac{K^{1/2}}{(K-1)} \approx Z_K \tag{7}$$

Thus, for large K, to a first approximation it is possible to ignore the dependence between (6) and (7). The terms U_K and Z_K on the right sides of (6) and (7) respectively are the first terms in an asymptotic expansion for large K of the left sides of those equations.

3. Convolution and Fourier Transform

Let U,V be random vectors in $\Re^N$ which are independent and have isotropic distributions with densities f_1, f_2 respectively. We let g_1, g_2 be the corresponding densities on $\Re^+$ of the lengths of U and V, $|U|, |V|$. Further, let f be the density of U + V and g be the density of $|U+V|$. Then by the independence of U,V,

$$f(y) = \int_{\Re^N} f_2(y-x)\, f_1(x)\, dx \tag{8}$$

Our interest is to determine a formula for g in terms of g_1 and g_2. To this end, we introduce the Fourier transform. If f is absolutely integrable on $\Re^N$, then the Fourier transform of f is defined by

$$\hat{f}(t) = \frac{1}{(2\pi)^{N/2}}\int_{\Re^N} f(y)\, e^{it\cdot y}\, dy \tag{9}$$

where the N vector $t=(t_1,t_2,\cdots,t_N)$ and $t\cdot y=t_1y_1+\cdots t_Ny_N$. If f(y) is spherically symmetric, i.e., f(y) is a function of $r=|y|$ only, say f(y) = h®, then its Fourier transform $\hat{f}(t)$ is also spherically symmetric; more specifically, we have

$$\hat{f}(t)=\rho^{(2-N)/2}\int_0^\infty r^{N/2}\,h(r)\,J_{(N-2)/2}(\rho r)dr \tag{10}$$

where $\rho=|t|$ and $J_\nu(r)$ denotes the Bessel function of the first kind of order ν. For a proof of this formula, we refer to Schwartz[2]. (Note that we have introduced the same letter r for $|y|$ as we used for the density of the length of the left side of (1). From the context, there should be no confusion as to which meaning for r is intended.). The Fourier transform of the convolution in (8) yields

$$\hat{f}(t)=\hat{f}_2(t)\,\hat{f}_1(t) \tag{11}$$

and combining (10) and (11) gives

$$\hat{f}(t)=\rho^{2-N}\int_0^\infty r^{N/2}\,h_2(r)\,J_{(N-2)/2}(\rho r)dr\quad\int_0^\infty r^{N/2}\,h_1(r)\,J_{(N-2)/2}(\rho r)dr \tag{12}$$

where the functions h_1, h_2 in (12) are defined by

$$h_1(r)=f_1(y),\quad h_2(r)=f_2(y) \tag{13}$$

with $r=|y|$ since the distributions defined by f_1, f_2 are isotropic.

Now if U is an isotropically distributed N vector with density f on $\Re^N$, and g is the corresponding density on $\Re^+$ of the length $|U|$, where $|x|^2=\sum_{i=1}^N x_i^2$, $x=(x_1,\cdots,x_N)$, then the relationship between the densities f and g is given by

$$f(y)=|y|^{-(N-1)}\,c_N^{-1}g(|y|) \tag{14}$$

where c_N, the surface 'area' of the unit sphere in $\Re^N$, is

$$c_N=\frac{2\pi^{N/2}}{\Gamma\left(\frac{N}{2}\right)} \tag{15}$$

Hence, upon substituting

$$h_1(r)=f_1(y)=r^{-(N-1)}c_N^{-1}g_1(r)$$

$$h_2(r)=f_2(y)=r^{-(N-1)}c_N^{-1}g_2(r) \tag{16}$$

into (12), we have

$$\hat{f}(t)=\rho^{2-N}\,c_N^{-2}\int_0^\infty r^{1-N/2}\,g_2(r)\,J_{(N-2)/2}(\rho r)dr\quad\int_0^\infty r^{1-N/2}\,g_1(r)\,J_{(N-2)/2}(\rho r)dr \tag{17}$$

If the dimension N of the space the measurement vectors X_{ij} belong to is even, say N = 2m, then $J_{(N-2)/2}(\rho r) = J_{m-1}(\rho r)$ is a Bessel function of the first kind of integer order. If N is odd, say N = 2m+1, then $J_{(N-2)/2}(\rho r) = J_{m-1/2}(\rho r)$ is a Bessel function of the first kind of fractional order, and is closely related to the Spherical Bessel function of the first kind $j_n(z)$, defined by

$$j_n(z) = \sqrt{\frac{\pi}{2z}}\, J_{n+1/2}(z) \tag{18}$$

More detailed information about Bessel functions may be in found in Abramowitz and Stegun [3].

4. Convergence of the Integrals

We now discuss convergence of the integrals in (17). Both integrals are functions of the variable $\rho = |t|$, hence $\hat{f}(t)$ is a spherically symmetric function of the transform variable t. This means that the inverse Fourier transform of $\hat{f}(t)$, i.e. f(y), can also be obtained as a one-dimensional integral with a Bessel function kernel.

For fixed ρ and small r, $J_{(N-2)/2}(\rho r) = O\left(r^{(N-2)/2}\right)$. Since we expect $g_i^{(j)}(0) = 0$ for i=1,2 for $0 \le j \le N-1$ (see Bickel [1]), it follows that both integrals in (17) are convergent at the lower limit 0. At the upper limit ∞, $J_{(N-2)/2}(\rho r) = O\left(r^{1/2}\right)$ and this by itself will not be enough to make the integral convergent. However the term $r^{(1-N/2)}$ will also make the integrals converge at the upper limit if N is sufficiently large. In practice the densities g_1, g_2 also tend to zero sufficiently rapidly to make the integrals converge at the upper limit. If these densities have compact support (i.e., are zero outside of a closed bounded subset of $\Re^+$, then the integrals in (17) no longer have infinite upper limits. The integrals in (17) can be evaluated accurately and efficiently by standard numerical quadrature methods.

5. Inversion

Note that the right side of (17) is a function of $\rho = |t|$ only. Hence $\hat{f}(t)$ is spherically symmetric. Let $\hat{f}(t) = G(\rho)$. The inverse Fourier transform of $\hat{f}(t)$, i.e., f(y) from (9), is defined by

$$f(y) = \frac{1}{(2\pi)^{N/2}} \int_{\Re^N} \hat{f}(t)\, e^{-it\cdot y}\, dt \tag{19}$$

It follows by analogy with (10) that

$$f(y) = r^{(2-N)/2} \int_0^\infty \rho^{N/2}\, G(\rho)\, J_{(N-2)/2}(\rho r)\, d\rho \tag{20}$$

where $r = |y|$. The relationship (14) between the density f of an isotropically distributed N vector and the corresponding density g of its length then implies

$$g(r) = r^{N/2} c_N \int_0^\infty \rho^{N/2} G(\rho) J_{(N-2)/2}(\rho r) d\rho \qquad (21)$$

where $G(\rho) = \hat{f}(t)$ is given by (17). Numerical evaluation of the integral in (21) proceeds similarly to the integrals in (17). With g_1, g_2 taken as p,q introduced at the end of section 1, and g taken as r (the density function for $\left|X_{M+1,1} - \bar{X}_{M+1}\right|$, *not* $|y|$), (17) and (21) together give the density r in terms of p and q.

6. Conclusions

We have shown that for large K, the left sides of (2) and (3) are approximately independent. Under this approximation, we derived an expression for the density r of the length of the left side of (1) in terms of the densities p,q of the lengths of the left sides of (2) and (3). This result is contained in equations (17) and (21) of the previous section.

Future work could include determining corrections to the above result for a finite number of measurements K, and practical numerical implementation of the above result.

7. References

[1] Peter Bickel, SAG Problem 97-2-1 Solution

[2] L. Schwartz , Mathematics for Physical Sciences, Addison-Wesley, Reading, MA, 1966, pp. 201-203.

[3] A. Abramowitz and I. Stegun, Handbook of Mathematical Functions, National Bureau of Standards, 1964

On the "30 error" criterion

Jack E. Porter
ITT Industries
April, 1997

Background

In 1985 it became known at ITT Industries Speech Laboratory in San Diego that George Doddington had recommended that speech recognition devices should, as a rule of thumb, be tested until at least thirty errors had been recorded, irrespective of what the error rate may be. This note, dating from that time, presents one rationalization of that rule. Since the same logic applies to testing speaker verifiers, or indeed to testing a wide variety of biometric devices, the 1985 original is reproduced below with some minor editing.

On the "30 error" criterion

Model the recognizer test precess as N independent trials, in which e errors are observed. [In the practice, we're skeptical of the independence assumption, but it is difficult to define or measure reasonable models incorporating dependence.] Our task is to estimate what the probability of error is. Call it p.

We treat p, the error probability, as a property of the device, hence not a random variable. Statements about the probability of p being greater or less than some fixed number are nonsense, because it either is or it isn't; there's no probability about it. But we can make probabilistic statements about , $e/N = \hat{p}$, the maximum likelihood estimate of p, because the number of errors observed in testing, is a random variable arising in the random process of testing. First let's see that e/N is the maximum likelihood estimator. The probability of observing exacting e errors is:

$$\mathrm{Prob}[\text{e errors}] = \binom{N}{e} p^e (1-p)^{N-e},$$

because independence of trials causes e to have a binomial distribution.

The likelihood function is this same function, with e considered fixed and p as the variable. The maximum likelihood estimator of p is the value. $\hat{p}$, which maximizes the likelihood function. Since the logarithm is a monotonically increasing function, it also maximizes the log (of the) likelihood function, L. Computing the derivative,

$$\frac{\partial L}{\partial p} = \frac{\partial}{\partial p}\left[\log\binom{N}{e} + e\,\log(p) + (N-e)\log(1-p)\right]$$

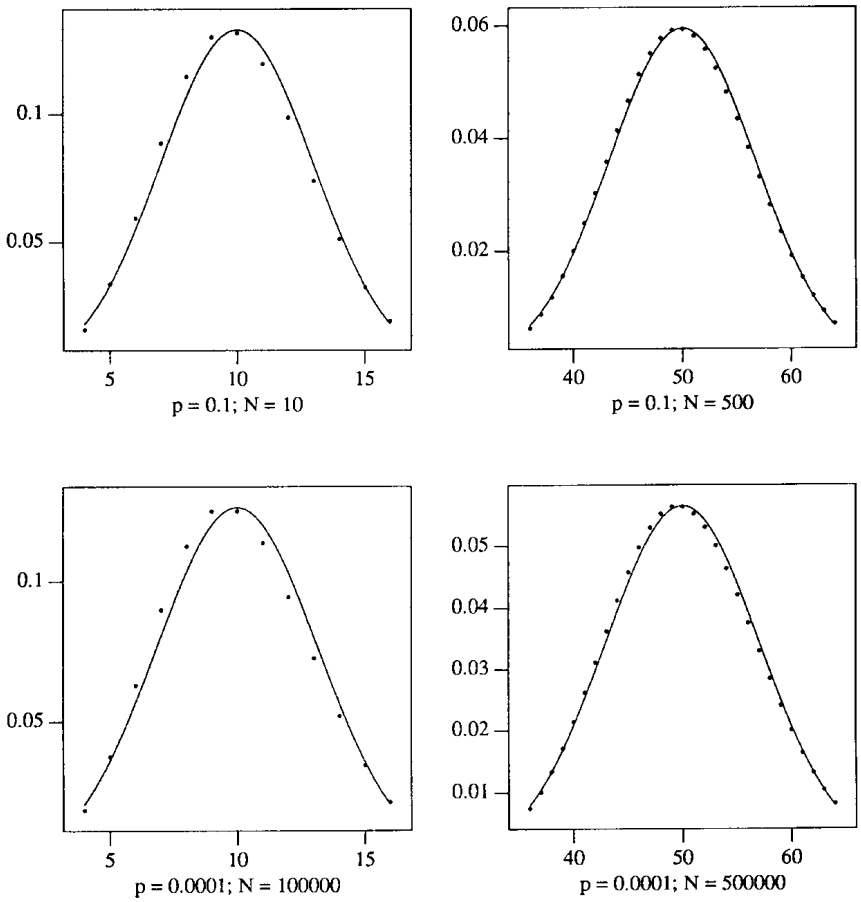
$$= \frac{e}{p} - \frac{N-e}{1-p}$$

which vanishes at

$$\hat{p} = \frac{e}{N}$$

(By taking the derivative again, you will see that the result is necessarily negative, so the unique $\hat{p}$ above is indeed a local maximum.)

Now when N is sufficiently large, the shape of the binomial distribution near its maximum becomes very similar to the Normal distribution with the binomial's mean and variance. How large does N have to be? It depends on p; but if the product Np is more than about ten, the shape is quite nearly Normal. Of course, the binomial probabilities are concentrated at integer values, while the Normal density is spread out smoothly, but the area under the Normal curve over any interval is nearly is almost equal to the sum of the discrete, binomial probabilities in that same interval. Thus if the expected number of errors (Np) is ten or more, we can use a normal approximation to the binomial, in the vicinity of the maximum. The following graphs show just how good this approximation is for various combinations of p and N.



The variance of the number of observed errors (a property of the binomial distribution) is

$$\sigma_e^2 = N\,p\,(1-p),$$

so the standard deviation of the maximum likelihood estimator of p is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}}$$

and using the maximum likelihood of p in place of p,

$$\sigma_{\hat{p}} = \frac{1}{N}\sqrt{e\left(1-\frac{e}{N}\right)}.$$

Since the distribution of e is approximately normal, so is that of $e/N = \hat{p}$.

For any given value of p (the true error probability), the maximum likelihood estimator $\hat{p}$ will usually (about 68% of the time) lie in the interval

$$p - \sigma_{\hat{p}} \le \hat{p} = \frac{e}{N} \le p + \sigma_{\hat{p}},$$

and almost always (about 95% of the time) lie in the interval

$$p - 2\sigma_{\hat{p}} \le \hat{p} = \frac{e}{N} \le p + 2\sigma_{\hat{p}}.$$

The fractions 68% and 95% are the areas under the Normal curve of error, in the intervals of width one and two standard deviations about the mean, respectively. They can be used because of the similarity of the binomial and Normal distributions shown earlier.

Since this relation holds for any value of p, if we make a habit of saying

$$p - \sigma_{\hat{p}} \le \hat{p} \le p + \sigma_{\hat{p}},$$

for this and any similar test of this or any other recognizer, we will be right, in the long run, 68% of the time. But with a little manipulation, we can show (for reasonable values of k)

$$p - k\sigma_{\hat{p}} \le \hat{p} \le p + k\sigma_{\hat{p}} \text{ if and only if } \hat{p} - k\sigma_{\hat{p}} \le p \le \hat{p} + k\sigma_{\hat{p}},$$

and the logically equivalent statement on the right is more interesting as it proclaims the true error probability lies within a certain interval. So we agree to make a habit of proclaiming

$$\hat{p} - k\sigma_{\hat{p}} \le p \le \hat{p} + k\sigma_{\hat{p}}$$

knowing that in any one test the statement is either true or false (not subject to probabilistic interpretation), but that in the long run, in tests of this or any other recognizer, we will be right 68% to 95% of the time, accordingly as we use k=1 or k=2.

These numbers, 68% and 95%, are called the "confidence level" of the statement. (Don't say they are the probability that the statement is correct in the presence of a statistician if you don't want to hear a lecture on the subject.) The confidence level to use

in forming confidence interval statements is usually a matter of convention. Values of 90%, 95% and 99% are common. Often a customer or someone interested in ensuring that the testing is adequate will specify what confidence level must be used. You can find the corresponding values of k using a table of areas under the Normal error curve.

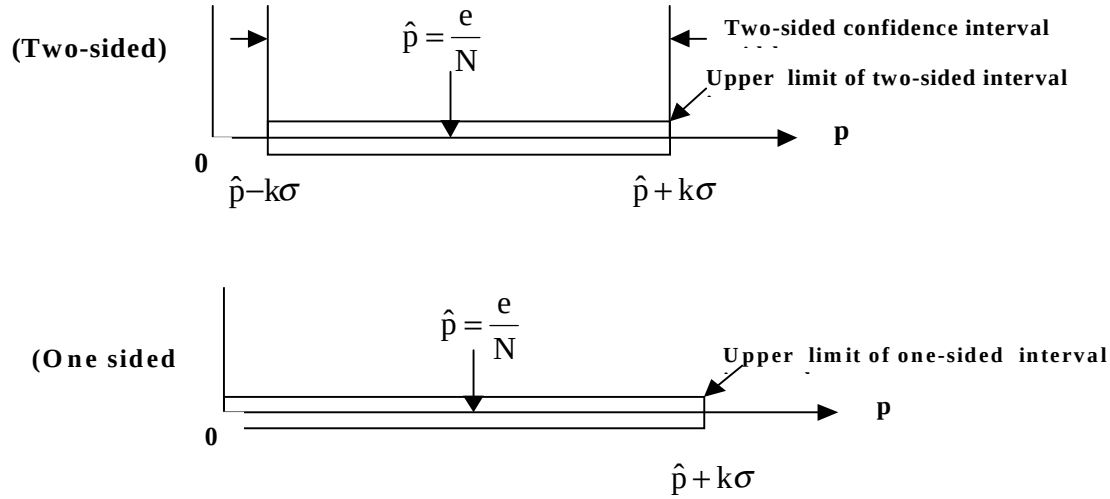
We could also make "one sided" statements of the form:

$$p \le \hat{p} + k\sigma_{\hat{p}}$$

and be right, in the long run, a percentage of the time equal to the area under the Normal curve from $-\infty$ to $k\sigma$ to the right of the mean.

So far we have just reviewed the logic of confidence interval statements applied to recognizer testing, and how areas under the Normal curve of error can be used to find the k corresponding to a given confidence level. This background is necessary because the 30 error criterion is related to being able to make confidence interval statements which are satisfying and sound sensible, as opposed to odd sounding ones.

Now, how satisfying are these confidence interval statements? Pictorially, they locate the true error probability in an interval around its maximum likelihood estimate:



When the end-point or –points of the interval are too far from the estimate, unsatisfactory statements result, like "with ninety percent confidence, the error rate is between 0.1% and1.9%." The lower and upper bounds differ by a factor of 19 in that case, and there are a lot of applications for which performance at the lower error rate bound would be acceptable but performance at upper bound would not. So the statement is roughly the same as "with ninety nine percent confidence, the error rate might possibly be accepatble" – unsatisfactory indeed. Of course the only way to shrink the confidence interval (at the chosen level of confidence) is to do more testing, and that's what the thirty error criterion is all about; just how *much* testing needs to be done to make satisfactory confidence interval statements.

One indicator of how satisfying the statement is the width of the confidence interval relative to the estimated error probability. In the unsatisfactory case cited above, the interval has width 1.8% (1.9%-0.1%), which is 1.8 time the estimate itself (1.0%). A value of one or less for this ratio leads to much more satisfactory confidence interval statements, and one is about the upper limit on this ratio for making reasonable statements. When the ration is exactly one, the upper limit of the confidence interval will be three times the lower limit, and ratios higher than that sugest the testing wasn't thorough enough.

Other indicators of a satisfactory confidence interval statement are the ratio of the upper one- or two-sided confidence interval limit to the maximum likelihood estimate. A value of about 1.5 for these ratios is a rough upper limit for satisfying statements.

The rule of thirty errors comes about because these indicators of satisfactory confidence intervals depend only on the number of errors observed (and not on the number of test trials or the actual error rate) when the error rate is small. To see this is so, you can use the formula given above to show that

$$\frac{\text{two sided interval width}}{\text{ML estimate}} = \frac{2k_2\sqrt{e(1-\frac{e}{N})}}{e} = \frac{2k_2}{\sqrt{e}},$$

$$\frac{\text{two sided upper limit}}{\text{ML estimate}} = \frac{e+k_2\sqrt{e(1-\frac{e}{N})}}{e} = 1+\frac{k_2}{\sqrt{e}},$$

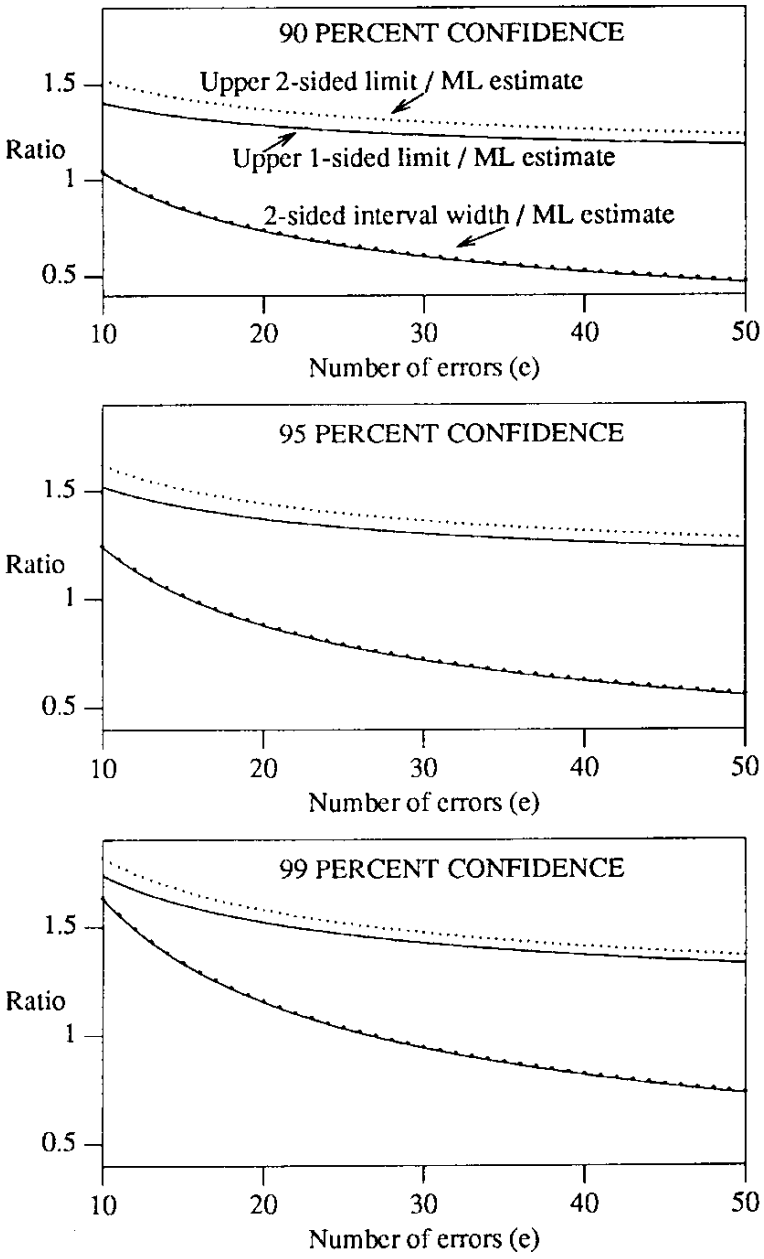
and

$$\frac{\text{one sided upper limit}}{\text{ML estimate}} = \frac{e+k_1\sqrt{e(1-\frac{e}{N})}}{e} = 1+\frac{k_1}{\sqrt{e}},$$

where k_1 and k_2 are the coefficients appropriate to one- or two-sided confidence intervals, respectively, and which may be approximated using the Normal distribution as illustrated earlier. Following is a graph showing these ratios for the number of errors (e) ranging from ten to fifty.

The bottom graph shows that, in order to make confidence interval statements with the two-sided interval width equal to the estimated error rate, one must test until about 30 errors are observed, when 99% confidence is sought. The same rule of thumb results in one-and two-sided upper limits which are about 1.5 times the maximum likelihood estimate. The other graphs show that for ninety or ninety five percent confidence, the rule of thumb would be to test until eleven or sixteen errors are observed, respectively.

So construction of satisfying confidence intervals is intimately related to the number of errors observed, when error probabilities are small. Maybe George Doddington had this sort of criterion in mind when he suggested testing until at least thirty errors have been observed.


90 PERCENT CONFIDENCE
Upper 2-sided limit / ML estimate
Upper 1-sided limit / ML estimate
2-sided interval width / ML estimate
Ratio
1.5
1
0.5
10
20
30
40
50
Number of errors (e)
95 PERCENT CONFIDENCE
Ratio
1.5
1
0.5
10
20
30
40
50
Number of errors (e)
99 PERCENT CONFIDENCE
Ratio
1.5
1
0.5
10
20
30
40
50
Number of errors (e)

Some Observations on the Cumulative Binomial Probability Distribution

W. A. Barrett
U.S. National Biometric Test Center

Abstract

The *cumulative binomial probability distribution* predicts the probability *P* of an event occurring *k* or more times in *n* trials, where the probability of occurrence of each event is *p.* In evaluating biometric systems, it is often of interest to determine the probability *p*, given *P*, for a comparatively large number of events and trials. Values of *P* of interest are also often very small or very nearly 1. It happens that this is a difficult computational task if approached in a *brute force* manner, i.e., performing large numbers of sums for various trial *p.*

We discuss some of the issues involved in such calculations. We also propose a more powerful computational approach to *p*, which makes use of the method of bisection combined with a known algorithm for computing the *incomplete beta function* $I_x(a,b)$, which is closely related to *P*.

We also propose some simple approximation formulas which are valid for very small *p*, by which the tails of the cumulative distribution may be estimated.

Background

The cumulative binomial probability distribution is defined as follows (Eq. 1):

$$P(k,n,p) \equiv \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} \qquad \text{for} \qquad 0 <= p <= 1 \wedge k < n \tag{1}$$

Here, *p* is the probability of a successful occurrence of some event. For example, in tossing a coin, we might say that a successful event is a *head*, with probability 0.5. Given *n* trials of the event, *P(k,n,p)* is the probability of *k* or more successful occurrences. For example, this yields the probability of obtaining *heads* in 5 or more successive tosses within a set of 20 tosses; *n* is 20, and *k* is 5. This particular probability is expected to be nearly 1.0, since the expected number of heads in 20 trials is 10, and we are only asking for 5 or more heads out of 20.

As *k* approaches *n*, *P(k,n,p)* approaches 0; similarly as *k* approaches 0, *P(k,n,p)* approaches 1. The function is clearly monotonic in *k*, and also in *p.*

The incomplete beta function $I_x(a, b)$ is defined as

$$I_x(a,b) = \frac{B_x(a,b)}{B(a,b)} \infty \tag{2}$$

where

$$B_x(a,b) = B_x(b,a) = \int_0^x t^{a-1}(1-t)^{b-1} dt \qquad \text{for} \qquad 0 <= x <= 1 \tag{3}$$

It can then be shown that

$$P(p,k,n) = I_p(k, n-k+1) \tag{4}$$

A plot of $I_x(a,b)$ for various values of (a, b) is given in figure 1. If both *a* and *b* are appreciably greater than one, then $I_x(a,b)$ is very nearly zero for small values of *x*, then rises very sharply at about $x=a/(a+b)$ to very nearly unity. This is suggested by the graph for (a,b) = (8,10) in figure 1.

Computing $I_x(a,b)$

Press *et al* [1] give a computationally efficient representation for the incomplete beta function, as follows:

$$I_x(a,b) = \frac{x^a(1-x)^b}{aB(a,b)}\left[\frac{1}{1+}\frac{d_1}{1+}\frac{d_2}{1+}\cdots\right] \tag{5}$$

where

$$d_{2m+1} = -\frac{(a+m)(a+b+m)x}{(a+2m)(a+2m+1)} \tag{6}$$

and

$$d_{2m} = \frac{m(b-m)x}{(a+2m-1)(a+2m)} \tag{7}$$

Equation (5) is a continued fraction expansion, i.e.

$$\left[\frac{1}{1+}\frac{d_1}{1+}\frac{d_2}{1+}\cdots\right] \equiv \cfrac{1}{1+\cfrac{d_1}{1+\cfrac{d_2}{1+\cdots}}} \tag{8}$$

This continued fraction converges rapidly for x < (a+1)/(a+b+2). For x > (a+1)/(a+b+2), it's better to use the symmetry relation

$$I_x(a,b) = 1 - I_{1-x}(b,a) \tag{9}$$

These considerations result in a pair of functions found in [1], **betai** and **betacf.** Function **betacf(float a, float b, float x)** computes the continued fraction for the incomplete beta function (eqs 5 and 6), returning its value as a **float.** Function **betai(float a, float b, float x)** computes the incomplete beta function $I_x(a,b)$, returning its value as a **float.**

Asymptotic Behavior as $x \to 0$

The incomplete beta function can also be expressed as a series expansion [1]:

$$I_x(a,b) = \frac{x^a(1-x)^b}{aB(a,b)}\left[1+\sum_{n=0}^{\infty}\frac{B(a+1,n+1)}{B(a+b,n+1)}x^{n+1}\right] \tag{10}$$

As $x \to 0$, the term in the brackets approaches 1, assuming that the two B functions are reasonable. Also $(1-x)^b$ approaches 1, so we have

$$\lim_{x\to 0} I_x(a,b) = \frac{x^a(1-x)^b}{aB(a,b)} \tag{11}$$

and therefore

$$\lim_{x \to 0} \ln(I_x(a,b)) = a \ln x - \ln a - \ln B(a,b) \qquad (12)$$

This is a straight line in a log-log plot, with a slope *a*. A study of Figure 2 shows that they are indeed straight lines (except near $x = 1$) and show a slope of *a*. That implies that for very small *x* (and reasonable *a*, *b*), we can easily estimate *x* from Eq. (12) and tabular values for *B(a,b)*.

The Inverse Beta Function

How may we find *x*, given *I(x)*, *a* and *b*, in general? This is an interesting problem. There's no known closed form solution for *x*. Also function *I(x)* behaves in a mischievous way for values of *a* and *b* that are very different from 1 (i.e. << 1 or >> 1). For example, if *a* and *b* are both large, e.g. 50, *I(x)* is extremely small until *x* approaches 0.5, whence it rapidly changes to a value just below 1 (see Figure 3). It tends to converge exponentially to 0 as x approaches 0, and to 1 as x approaches 1.

This behavior makes finding an inverse by such accelerated numerical methods as the secant, *regula falsi*, or Newton-Raphson iteration methods difficult. One is never sure that the method will converge in a finite number of trials. Indeed for certain *I(x)*, *a* and *b*, any of these will either yield trial solutions far beyond the (0,1) bounds, or may converge so slowly as to be impractical. (This problem is discussed at length in [1], chapter 9).

However, the bisection method (reference [1], section 9.1) will always yield a solution for this function, since it only requires monotonicity and continuity, which I_x satisfies. *x* is bounded by 0 and 1, so the bisection can start by evaluating $I_x(a,b)$ for $x = 0.5$, then continue by bisecting one or the other sides, depending on each trial's outcome. For example, given N trials, a solution will always be found to within an absolute precision of 2^{-N}.

A plot of a set of inverse I_x values computed this way is in figure 4, as a log-log plot. Notice that all the curves in figure 4 level off sharply at log(x) = -23. (The log is to base *e*). This is solely due to stopping the bisection process at a finite number of iterations, yielding a maximum absolute precision in x of about 10 decimal places. ($10^{-10} \approx e^{-23}$). By changing the fixed parameter **PRECISION** (desired number of decimal places) in function **getRoot,** one may shift this flattening-off position to any desired level.

The C++ code for our root-finding function is given in figure 5. The function **getRoot** draws upon function **betai** given in [1]. It's written with double-precision floating point. Function **betai** and its companion functions should also be modified by replacing **float** with **double** everywhere.

On a 75 Mhz Pentium, 180 solutions take about 3 seconds for a precision of 10^{-10}.

Reference

[1] Press, Teukolsky, Vettering and Flannery, *Numerical Recipes in C*, second edition, Cambridge University Press. The companion software programs in diskette form may be ordered from the publishers.

Figures

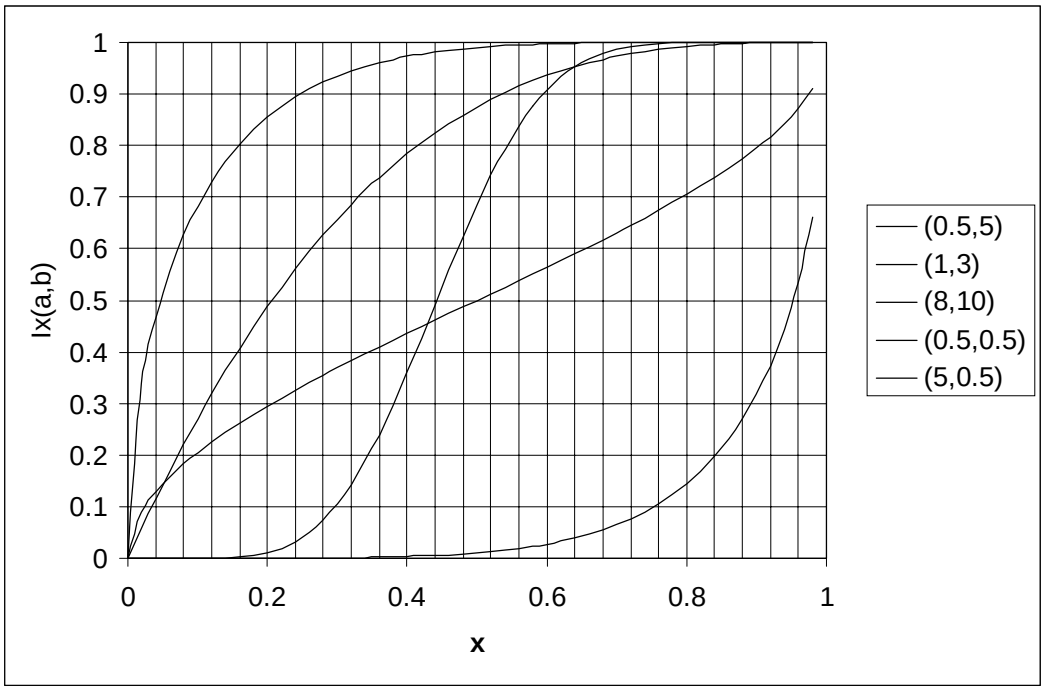


Figure 1. Plots of $I_x(a,b)$ for various (a, b). x is along the horizontal axis, and I along the vertical axis. Along the ordinate $I_x = 0.6$, from left to right, the curves are for (a, b) = (0.5, 5), (1, 3), (8, 10), (0.5, 0.5), (5, 0.5).

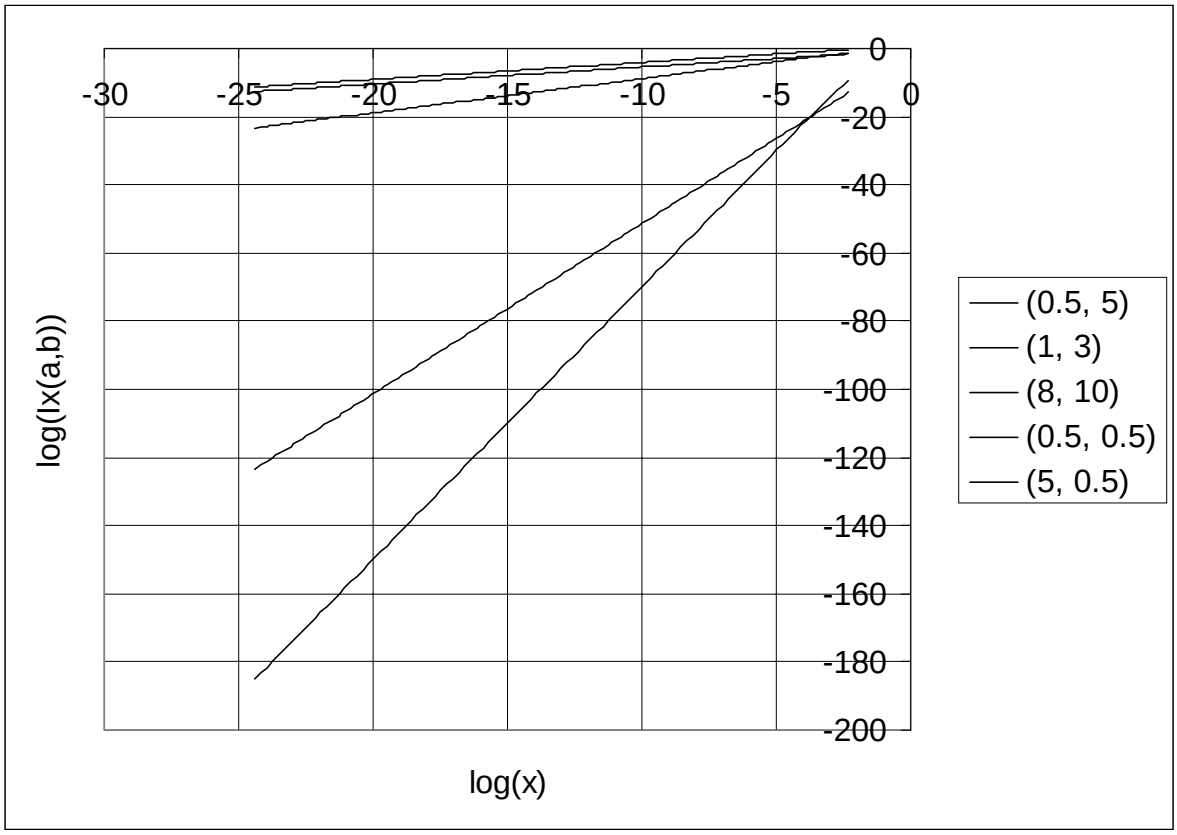


Figure 2. Incomplete beta function for small values of *x*, plotted in *log-log* scales. From top to bottom, (a, b) = (0.5, 5), (0.5, 0.5), (1, 3), (5, 0.5), and (8, 10).

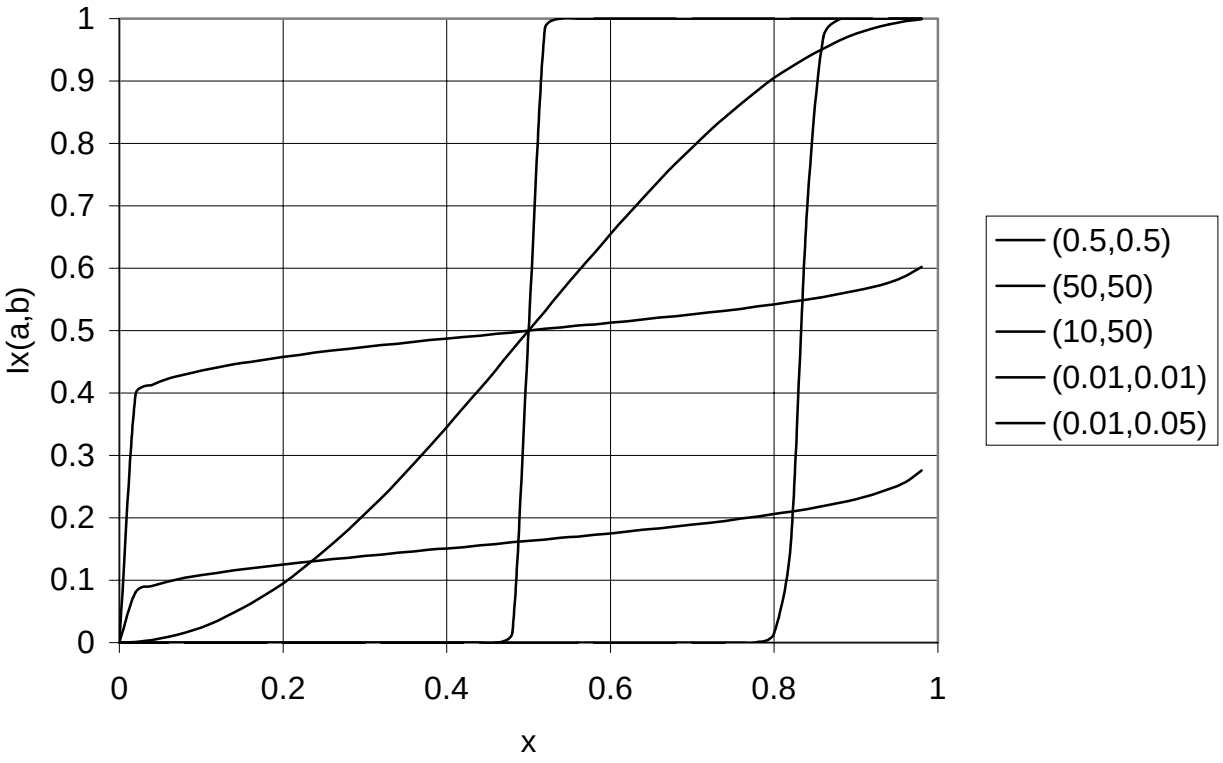


Figure 3. Incomplete beta function for some extreme values of (a, b). At x = 0.6, the curves are, from top to bottom, (a, b)= (50, 50), (0.5, 0.5), (0.01, 0.01), (0.01, 0.05), and (10, 50).

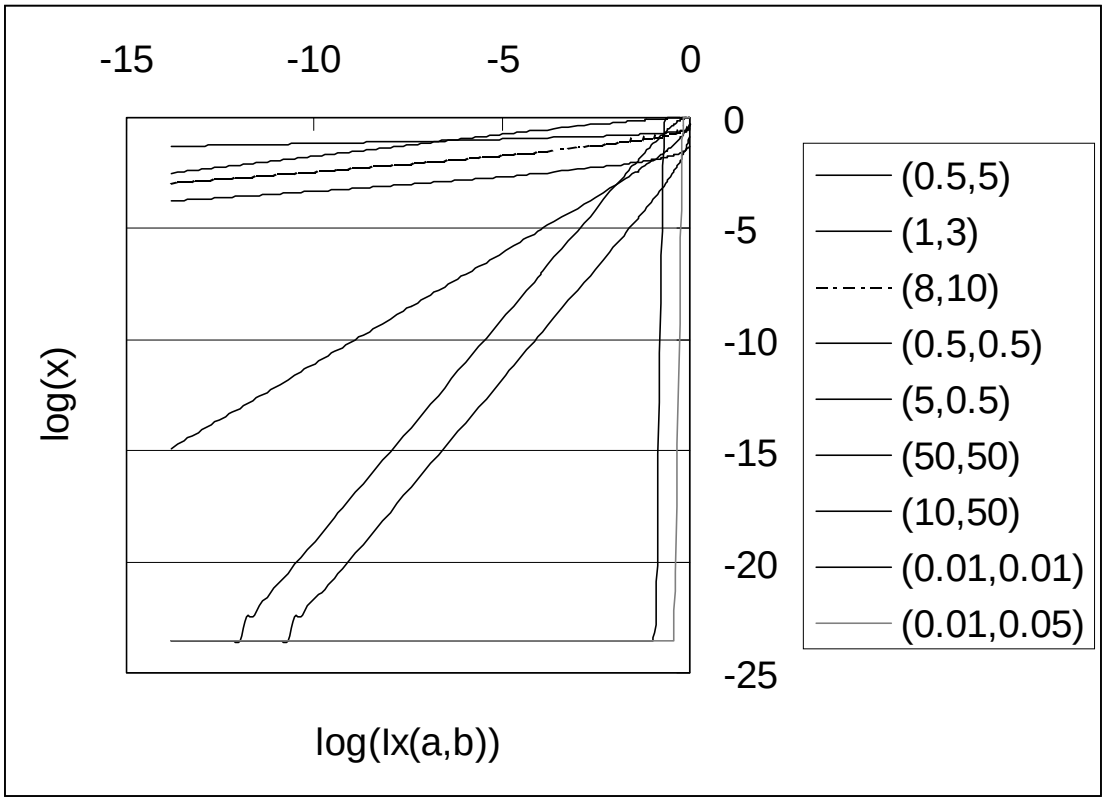


Figure 4. *log-log* plot of the inverse of $I_x(a, b)$ for various (a, b). These were computed using the bisection method, starting with a given value of *I*, to find the corresponding *x*. *I* is along the horizontal axis, and *x* along the vertical axis. All the curves flatten out below *x = 10^-23* due to the fixed number of iterations chosen for the computation.

```
#define PRECISION 1e-10
static const int iterations= (int)(-log(PRECISION)/(log(2.0)))+1;
// note:  not all C++ compilers will evaluate the above function at
// compile time

double
getRoot(double a, double b, double Ix)
{
  // This will always converge to a root since betai is
  //   monotonic increasing in x (0 < x < 1), and is bounded by (0, 1)
  //   for all (a, b).  It behaves properly for values of (a, b) very
  //   different from 1, i.e. very small or very large.
  // The root will be bracketed in (x1, x2) where x1 < x2.
  //   x0 will lie between x1..x2

  double x0= 0.5, x1= 0.0, x2= 1.0;
  double root;
  int    trial;

  for (trial= 0; trial < iterations; trial++) {
    x0= 0.5*(x1+x2);
    root= betai(a, b, x0);
    if (fabs(Ix - root) < PRECISION) break;
    if (Ix > root)
      x1= x0;     // root is bracketed in x0..x2
    else
      x2= x0;   // root is bracketed in x1..x0
  }
  return x0;
}
```

Figure 5. C++ code for computing the inverse beta function using the bisection method. This requires function **betai** found in Press, Flannery, *et al*, reference [1].

Technical Testing and Evaluation of Biometric Identification Devices

James L. Wayman, Director
National Biometric Test Center

Abstract

Although the technical evaluation of biometric identification devices has a history spanning over two decades, it is only now that a general consensus on test and reporting measures and methodologies is developing in the scientific community. By "technical evaluation", we mean the measurement of the five parameters generally of interest to engineers and physical scientists: false match and false non-match rates, binning error rate, penetration coefficient and transaction times. Additional measures, such as "failure to enroll" or "failure to acquire", indicative of the percentage of the general population unable to use any particular biometric method, are also important. We have not included in this chapter measures of more interest to social scientists, such as user perception and acceptability. Most researchers now accept the "Receiver Operating Characteristic" (ROC) curve as the appropriate measure of the application-dependent technical performance of any biometric identification device. Further, we now agree that the error rates illustrated in the ROC must be normalized to be independent of the database size and other "accept/reject" decision parameters of the test. This chapter discusses the general approach to application-dependent, decision-policy independent testing and reporting of technical device performance and gives an example of one practical test. System performance prediction based on test results is also discussed.

Introduction

We can say, somewhat imprecisely, that there are two distinct functions for biometric identification devices: 1) to prove you are who you say you are, and 2) to prove you are not who you say you are not. In the first function, the user of the system makes a "positive" claim of identity. In the second function, the user makes the "negative" claim that she is not anyone already known to the system.

Biometric systems attempt to use measures that are both distinctive between members of the population and repeatable over each member. To the extent that measures are not distinctive or not repeatable, errors can occur. In discussing system errors, the terms "false acceptance" and "false rejection" always refer to the claim of the user. So a user of a positive identification system, claiming to match an enrolled record, is "falsely accepted" if incorrectly matched to a truly non-matching biometric measure, and "falsely rejected" if incorrectly not matched to a truly matching biometric measure. In a negative identification system, the converse is true: "false rejection" occurring if two truly non-matching measures are matched, and "false acceptance" occurring if two truly matching measures are not matched. Most systems have a policy allowing use of multiple biometric samples to identify a user. The probability that a user is ultimately accepted or rejected depends upon the accuracy of the comparisons made and the accept/reject decision policy adopted by the system management. This decision policy is determined by the system manager to reflect the operational requirements of acceptable error rates and transaction times and, thus, is not a function of the biometric device itself.

Consequently, in this chapter we refer to "false matches" and "false non-matches" resulting from the comparison of single presented biometric measure to a single record previously enrolled. These measures can be translated into "false accept" and "false reject" under a variety of system decision policies.

In addition to the decision policy, the system "false rejection" and "false acceptance" rates are a function of five inter-related parameters: single comparison false match and false non-match rates, binning error rate, penetration coefficient, and transaction speed. In this chapter, we will focus on testing of these basic parameters and predicting system performance based on their resulting values and the system decision policy.

Regardless of system function, the system administrator ultimately has three questions: What will be the rate of occurrence of false rejections, requiring intervention by trained staff?; Will the probability of false acceptance be low enough to deter fraud?; Will the throughput rate of the system keep up with demand? The first question might further include an estimate of how many customers might be unable to enroll in or use the system. The focus of this chapter will be on developing predictive tools to allow "real-world" estimates of these numbers from small-scale tests.

Classifying Applications

Technology performance is highly application dependent. Both the repeatability and distinctiveness of any biometric measure will depend upon difficulty of the application environment. Consequently, we must test devices with a target application in mind. Although each application is clearly different, some striking similarities emerge when considered in general. All applications can be partitioned according to at least seven categories:

1. Cooperative versus Non-cooperative: Is the deceptive user attempting to cooperate with the system to appear to be someone she is not, or attempting not to cooperate to not appear to be someone known to the system?
2. Overt versus Covert: Is the user aware that the biometric measure is being taken?
3. Habituated versus Non-habituated: Is the user well acquainted with the system?
4. Attended versus Non-attended: Is the use of the biometric device observed and guided by system management?
5. Standard Environment: Is the application indoors or in an outdoor, or environmentally stressful, location?
6. Public versus Private: Will the users of the system be customers (public) or employees (private) of the system management?
7. Open versus Closed: Will the system be required, now or in the future, to exchange data with other biometric systems run by other management?

This list is incomplete, meaning that additional partitions might also be appropriate. We could also argue that not all possible partition permutations are equally likely or even permissible. A cooperative, overt, habituated, attended, private, application in a laboratory environment will generally produce lower error rates than outdoor applications on a non-habituated, unattended population.

The Generic Biometric System

Although biometric devices rely on widely different technologies, much can be said about them in general. Figure 1 shows a generic biometric identification system, divided into five sub-systems: data collection, transmission, signal processing, decision and data storage. The key subsystems are:

1. Data collection, which includes the imaging of a biometric pattern presented to the sensor.
2. Transmission, which may include signal compression and re-expansion and the inadvertent addition of noise.
3. Signal processing, in which the stable, yet distinctive, "features" are extracted from the received signal and compared to those previously stored.
4. Storage of "templates" derived from the "features" and possibly the raw signals received from the transmission subsystem.
5. Decision, which makes the decision to "accept" or "reject" based upon the system policy and the scores received from the signal processing system.

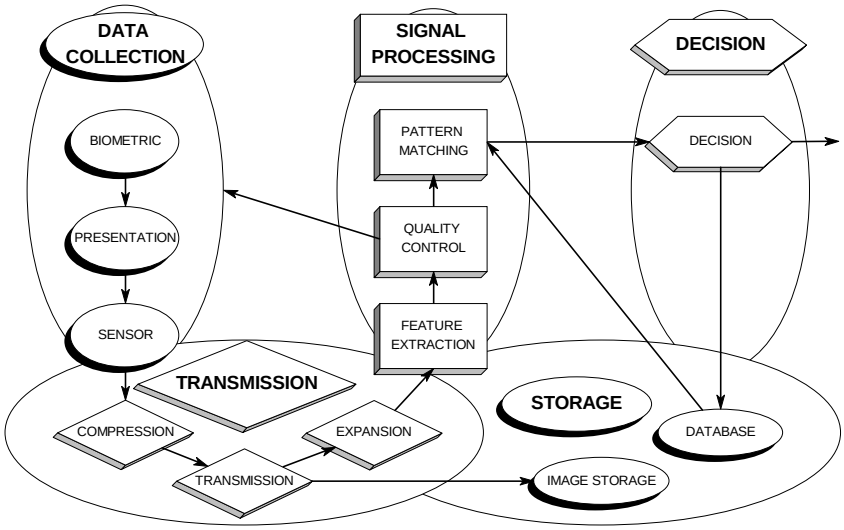


Figure 1: General biometrics system.

Application-Dependent Device Testing

We are now in a position to present a more mathematical development of the above ideas and to explain more precisely three major difficulties in biometric testing: the dependence of measured error rates on the application classification, the need for a large test population which adequately models the target population, and the necessity for a time delay between enrollment and testing. This section will present a mathematical development of the five basic system parameters: false match rate, false non-match rate, binning error rate, penetration coefficient and transaction speed. We will also discuss Receiver Operating Characteristic curves [1-5] and confidence interval estimations.

Features

The features extracted by the signal processing sub-system of Figure 1 are generally vectors in a real or complex [6] metric space, with components generally taking on integer values over a bounded domain. In some systems [7], the domain of each component is restricted to the binary values of {0,1}. Fingerprint systems are the primary exception to this rule, using features not in a vector space. In this chapter, we will suppose that the components are any real number.

If each feature vector, X, has J components, x_j, we can write

$$X = (x_j)\ ,\ j=1,2...,J \tag{1}$$

where

$$x_j = \mu_j + \varepsilon \tag{2}$$

The components of the feature vector, X, consist of a fundamental biometric measure, μ_j, and an error term, ε,. both assumed to be time-invariant and independent over all J. The error term, ε, has some distribution, $\xi(0,\sigma^2)$, not presumed to be normal. To simplify the development, we will assume that the distribution of the errors is identical for all components of X. We can say, therefore, that the components of X come from a distribution

$$x_j \sim \xi(\mu_j,\sigma^2) \tag{3}$$

Errors arise from the data collection sub-system of Figure 1, perhaps owing to random variations in the biometric pattern, pattern presentation or the sensor. Errors owing to the transmission or compression processes of the transmission sub-system of Figure 1 may also be important. Assuming the errors to be uncorrelated, we can write

$$\sigma^2 = \sigma^2_{\text{biometric}} + \sigma^2_{\text{presentation}} + \sigma^2_{\text{sensor}} + \sigma^2_{\text{transmission}} \tag{4}$$

where subscripted terms on the right hand side are the error variances associated with changes in the biometric pattern, the presentation and the sensor, accordingly, along with the transmission error variance. In reality, compression errors, included in the transmission error term, may be a function of the sensor error, adding correlations to the error process.

We note the first major problem with error testing of biometric devices: the error variance of each of the terms in (4) is highly application dependent. There is currently no way to predict the error terms for all applications from measurements made in any one test environment. Consequently, test results are always dependent upon the test environment and will not reflect errors in dissimilar application environments of the "real

world". Testing of the individual error variances as noted in (.4) has not been done, so we will consider in this chapter only the composite variance σ^2.

Templates

At the time of enrollment, the user presents $M \geq 1$ samples of the biometric measure for the creation of a "template", $\bar{X}$, from the M feature vectors, X^i. The superscript, i=1,2…,M, has been added to the feature vector, X, to indicate multiple samples from the same user. The template, $\bar{X}$, may be computed as the average of the M feature vectors, X^i, in which case we can write,

$$\bar{X} = (\bar{x}_j), \tag{5}$$

where

$$\bar{x}_j = \frac{1}{M}\sum_{i=1}^{M} x_j^i . \tag{6}$$

When computing the weighted sum of uncorrelated random variables, the following relationships hold [8]:

$$\mu = \sum_{i=0}^{M} c_i \mu_i \tag{7}$$

and

$$\sigma^2 = \sum_{i=0}^{M} c_i^2 \sigma_i^2 , \tag{8}$$

where c_i i=1,2…M is the weighting vector.

Under some weighting vectors, such as uniform c_i, the distribution of the weighted sum of uncorrelated random variables will, by the Central Limit Theorem, tend toward normality as M increases

Applying (7) and (8) to (6), the components of the template $\bar{X}$ are seen to be distributed as

$$\bar{x}_j \sim \bar{\xi}(\mu_j, \frac{\sigma^2}{M}), \tag{9}$$

where $\bar{\xi}$ indicates a distribution tending toward normality. In many systems, however, M may be one or three, meaning that $\bar{\xi}(\mu_j, \frac{\sigma^2}{M})$ cannot generally be considered normal.

"Genuine" Distances

The feature vectors, X, vary across users, which we will express by adding a subscript, h=1,2..N, to our notation, where N is the number of enrolled users. Our original assumption, that the components of the feature vector, X, are independently distributed random variables, is now expanded to include independence over users, as well. The sample data across the entire population of users is "non-stationary", meaning that multiple measures from a single user cannot be used to approximate the distribution for the entire population. This adds a second major complication to biometric testing, the requirement for a large test population that adequately represents the target population of the application.

For every user, a template is created from M samples of the biometric measure. Then for each user, the biometric feature vector is re-sampled and a distance measure, $d_{h\,h}$, is computed between the additional sample and the user's template.

$$d_{h\,h} = \left\| \bar{X}_h - X_h^{M+1} \right\| = \left\| \Delta X_{h\,h} \right\|, \tag{10}$$

where the double brackets indicate a general distance measure and

$$\Delta X_{h\,h} = (\bar{x}_{h\,j} - x_{h\,j}^{M+1}) \text{ for j=1,2…J.} \tag{11}$$

Applying equation (6), the components of $\Delta X_{h\,h}$ become

$$\Delta x_{h\,h\,j} = \frac{1}{M}\sum_{i=1}^{M} x_j^i \; - \; x_j^{M+1}. \tag{12}$$

Referring to equation (2),

$$\Delta x_{h\,h\,j} = \mu_{h\,j} - \mu_{h\,j} + \varepsilon = \varepsilon\,, \tag{13}$$

where, by equation (8), the error term, ε, is distributed as

$$\varepsilon \sim \bar{\xi}(0, \frac{(M+1)\sigma_h^2}{M}) \tag{14}$$

Consequently, the distribution of the components, $\Delta x_{h\,h\,j}$, used to compute the genuine distance $d_{h\,h}$, do not tend toward normality with increasing M, but rather to $\xi(0, \sigma_h^2)$, the original distribution of the error terms for user h.

One of the tasks in testing will be to develop the probability distribution of these distance measures over the entire user population. We will call this density function $F'_{GEN}(d)$ where "GEN" indicates "genuine", indicating that samples are being compared to each user's own (genuine) template.

We have assumed, for simplicity, that both $\mu_{h\,j}$ and σ_h^2 are time-invariant. In reality, however, both may drift over time. The measurement means, $\mu_{h\,j}$, may move as a "random walk". In general, biometric system identification errors increase with the passage of time after enrollment. This phenomenon is generally attributed to changes in the underlying biometric measures, $\mu_{h\,j}$, and, consequently, is referred to as "template aging". Sensor and presentation changes over time may also be contributing factors. This represents a third major problem in the error testing of biometric devices: performance estimation may depend upon the time difference between enrollment and test samples.

Consider $\mu_{h\,j}=f(t)$, where t is time. Then the $\mu_{h\,j}$ of (13) are also functions of time, and can be given by

$$\Delta x_{h\,h\,j}(t) = \mu_{h\,j}(t_1) - \mu_{h\,j}(t_2) + \varepsilon\,, \tag{15}$$

where, again,

$$\varepsilon \sim \bar{\xi}\left(0, \frac{(M+1)\sigma_h^2}{M}\right) \tag{16}$$

and t_1 and t_2 are the times at enrollment and later sampling, respectively. To understand the effects of a time-varying mean, we compare the time invariant case of of (13) to the time varying case of (15). If our distance measure is Euclidean, then any variation over time in the $\mu_{h\,j}$ causes an increase in the expected distance values, $E(d_{h\,h})$, over the population, because

$$E\left(\left(\sum_{j=1}^{J}\left(\mu_{h\,j}(t_1) - \mu_{h\,j}(t_2) + \varepsilon\right)^2\right)^{\frac{1}{2}}\right) > E\left(\left(\sum_{j=1}^{J}\varepsilon^2\right)^{\frac{1}{2}}\right) \tag{17}$$

if $\mu_{h\,j}(t_1)$ and $\mu_{h\,j}(t_2)$ are not always equal and the ε and μ terms are uncorrelated, as originally assumed.

Ideally, the time interval between enrollment and sampling in any test should be similar to the interval expected in the application. This is usually not possible to estimate or attain so, as a "rule of thumb", we would like the time interval to be at least on the order of the healing time of the body part involved. This would allow any temporary variations in the biometric measures to be considered in the computation of the template-to-sample distances. This requirement, of course, greatly increases test time and expense.

It has been commonly noted in practice [9] that users can be roughly divided into two groups depending upon distance measurements, $d_{h\,h}$: a large group, $\{N_1\}$ with small distance measures, called "sheep", and a smaller group, $\{N_2\}$, with high distance

measures, called "goats" [10], where the total population N = N_1 + N_2. The preceding development leads us to believe that "sheep" and "goats" may be distinguished either by the value of σ^2_h, with users in $\{N_1\}$ having smaller error variance σ^2 then users in $\{N_2\}$, or by the time-variability of their fundamental measures, $\mu_{h\,j}$.

More precisely, the terms "goats" and "sheep" have generally been applied to indicate the chronic classification of individuals, "goats" being users who consistently return large distance measures when samples are compared to stored templates. Multiple test samples over time from the same user do not return additional independent data for population estimates and may result in the mixing of "habituated" and "non-habituated" user interaction with the system.. Consequently, previous tests have not followed users in time. From any single set of test distance samples, the large distances will represent both chronic "goats" and a few "sheep", who simply happen to return a large distance score at the tail of the "sheep" error distribution.

Histograms of "genuine" distances are noted in practice to be bi-modal, the distance measures from the "sheep" contributing to the primary mode, and the distance measures from the "goats" contributing to the secondary mode.

"Impostor" Distances

Using the same metric as used for establishing the "genuine" distances, a set of samples X_k^{M+1} could be compared to non-matching templates $\bar{X}_h$, $h \neq k$, to arrive at a non-matching distance, $d_{h\,k}$. We can rewrite Eqs. (10), (11) and (12) to get

$$\Delta x_{h\,k\,j} = \frac{1}{M}\sum_{i=1}^{M} x^i_{h\,j} \; - \; x_{k\,j} \tag{18}$$

By (2), these components of the difference vector $\Delta X_{h\,k}$ can be written

$$\Delta x_{h\,k\,j} = \mu_{h\,j} - \mu_{k\,j} + \varepsilon\,, \tag{19}$$

where, by (7) and (8)

$$\varepsilon \sim \bar{\xi}(0, \frac{\sigma_h^2}{M} + \sigma_k^2) \tag{20}$$

The distribution of the error term, ε, does not tend to normal with increasing M, but rather to the original, unspecified distribution $\xi(0, \sigma_k^2)$.

Over the entire population, these distance measures, $d_{h\,k}$, for $h \neq k$, have a density function $F'_{IMP}(d)$ where "IMP" means "impostor", so named because the density is of measures from an "impostor" sample to a non-matching template. Some researchers [11] have suggested the use of additional templates not matched by samples for the calculation of "impostor" distributions. This is sometimes called a "background" database. In our

notation, this would create two groups of templates, $\bar{X}_h, h \in \{H_1\}$, those matched by test samples, and $\bar{X}_h, h \in \{H_2\}$, those not matched. The genuine distance distribution, based on distances whose components are distributed as (14), considers only the σ_h^2 for the users, $h \in \{H_1\}$, with matching samples. By (20), however, the impostor distribution is impacted by the distribution of variances σ_h^2 for users in both matched and unmatched groups. Unless we are certain that the populations are the same, such the distribution of the terms $(\mu_{h\,j} - \mu_{k\,j})$ do not depend upon the group $\{H_1\}$ or $\{H_2\}$ from which the members come, and that the application environment is the same, such that σ_h^2 is also group independent, "background" databases only add uncertainty to the measurements.

"Inter-Template" Distances

Between each pair of templates, $\bar{X}_h$ and $\bar{X}_k$, $h \neq k$, a distance, $\delta_{h\,k}$, can be computed using the same metric as was used to compute the genuine and impostor distances:

$$\delta_{h\,k} = \left\| \bar{X}_h - \bar{X}_k \right\| = \left\| \Delta \bar{X}_{h\,k} \right\|. \tag{21}$$

We use the Greek symbol, δ, to differentiate this "inter-template" distance from the "impostor" distance of the preceding section.

Because we are working in a metric space, the distances are symmetric such that $\delta_{h\,k} = \delta_{k\,h}$, and the distance of any vector from itself is zero, so $\delta_{k\,k} = 0$. Therefore, N(N-1)/2 non-independent distances can be computed between all templates. From (6), we have for the components of $\Delta X_{h\,k}$,

$$\Delta \bar{x}_{h\,k\,j} = \frac{1}{M} \sum_{i=1}^{M} x_{h\,j}^{i} - \frac{1}{M} \sum_{i=1}^{M} x_{k\,j}^{i} \quad \text{for } h \neq k. \tag{22}$$

By (2),

$$\Delta \bar{x}_{h\,k\,j} = \mu_{h\,j} - \mu_{k\,j} + \varepsilon, \tag{23}$$

where, by (7) and (8),

$$\varepsilon \sim \bar{\xi}\left(0, \frac{\sigma_h^2 + \sigma_k^2}{M}\right) \tag{24}$$

For the inter-template distance, the error term is from a distribution tending toward normality as M increases. Further, in the limit, the variance of the error term goes to zero with increasing M. This indicates that the inter-template terms are not impacted by the measurement error for large M.

We denote the density function of $\delta_{h\ k}$ over the population as $F'_{IT}(\delta)$, where "IT" indicates "inter-template". Comparing 20 to 24, the distributions of the terms composing the "impostor" and "inter-template" distributions are equivalent only when M=1. For M>1,

$$\frac{\sigma_h^2}{M} + \sigma_k^2 > \frac{\sigma_h^2 + \sigma_k^2}{M} \tag{25}$$

indicating that the variance in the error terms of components comprising the impostor distance vector will be larger than that of the terms comprising the inter-template distance vector. For uncorrelated means and errors, as assumed, and Euclidean distances, the expected values of the impostor distances will be greater than for the inter-template distances. Consequently, the impostor and inter-template distributions will only be equivalent for M=1. The inter-template distribution makes an increasingly poor proxy for the impostor distribution as M increases.

The three distributions, "genuine", "impostor" and "inter-template", are shown as Figure 2. Both the impostor and inter-template distributions lie generally to the right of the genuine distribution, which shows the second mode noted in all experimental data.

Decreasing the difficulty of the application category (changing from non-habituated, non-attended to habituated, attended, for instance) will effect the genuine distribution by making it easier for users to give repeatable samples, decreasing the value of σ_h^2, and thus moving the genuine curve to the left. Decreasing the measurement errors, σ_h^2 and σ_k^2, also causes movement in the impostor distribution to the left, but causes movement in the "inter-template" distribution only for small M.

Operational systems store templates and transaction distance measures, but rarely store the samples acquired during operations. Consequently, under the assumption that all users are "genuine", the genuine distribution can be constructed directly from the transaction distance measures. The "inter-template" distribution can be constructed by "off-line" comparison of the distances between templates. The "impostor" distribution, however, cannot be reconstructed without operational samples. Methods for convolving $F'_{GEN}(d)$ and $F'_{IT}(\delta)$ to determine $F'_{IMP}(d)$, under some simplifying assumptions, have been discussed in [12] and [13].

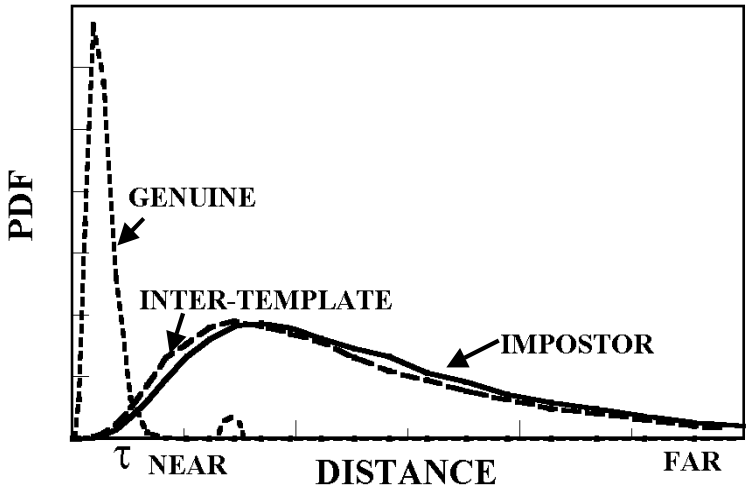


FIGURE 2 Distance distributions.

A decision policy commonly accepts as genuine any distance measure less than some threshold, τ. In non-cooperative applications, it is the goal of the deceptive user ("wolf") not to be identified. This can be accomplished by willful behavior to increase his/her personal σ_h^2, moving a personal genuine distribution to the right and increasing the probability of a score greater than the decision policy threshold, τ. We do not know for any non-cooperative system the extent to which "wolves" can willfully increase their error variances.

ROC Curves

Even though there is unit area under each of the three distributions, the curves themselves are not dimensionless, owing to their expression in terms of the dimensional distance. We will need a non-dimensional measure, if we are to compare two unrelated biometric systems using a common and basic technical performance measure.

The false non-match rate, FNMR, at any τ is the percentage of the distribution $F'_{GEN}(d)$ greater than d= τ and can be given by

$$FNMR(\tau) = \int_{\tau}^{\infty} F'_{GEN}(d)\, d\, d = F_{GEN}(d)\,|_{\tau}^{\infty} = 1 - F_{GEN}(d)\,|_{0}^{\tau}. \qquad (26)$$

The false match rate, FMR, at any τ is the percentage of the distribution $F'_{IMP}(d)$ smaller than d= τ and can be given by

$$FMR(\tau) = \int_{0}^{\tau} F'_{IMP}(d)\, d\, d = F_{IMP}(d)\,|_{0}^{\tau}. \qquad (27)$$

The "Receiver Operating Characteristic" (ROC) curve is the two-dimensional curve represented parametrically in τ by the points $\left[F_{IMP}(d)|_0^\tau, F_{GEN}(d)|_0^\tau \right]$. We find the ROC curve to be more intuitive when displayed as the points $\left[FMR(\tau), FNMR(\tau) \right]$.

As previously noted, the probability densities $F'_{IMP}(d)$ and $F'_{IT}(\delta)$ are equivalent only when M=1. When M>1 and the distances are Euclidean with component means and error uncorrelated, the expected values of the impostor distances will be larger than the expected values of the inter-template distances. Consequently, in this case, if we compare the integrals of the distributions between 0 and some threshold, τ,

$$F_{IMP}(d)|_0^\tau = \int_0^\tau F'_{IMP}(d)\, d\,d \tag{28}$$

and

$$F_{IT}(\delta)|_0^\tau = \int_0^\tau F'_{IT}(\delta)\, d\delta\,, \tag{29}$$

we find that

$$F_{IT}(\delta)|_0^\tau \geq F_{IMP}(d)|_0^\tau\,. \tag{30}$$

By this last equation, we can see that if the inter-template distribution, $F_{IT}(\delta)|_0^\tau$, is used to replace the impostor distribution under these conditions, the false match rate will be overestimated.

We note that the ROC curve is non-dimensional. Other non-dimensional measures have been suggested for use in biometric testing [14], such as "D-prime"[1,2] and "Kullback-Leibler" [15] values. These are single, scalar measures, however, and are not translatable to error rate prediction. The "equal error rate" (EER) is defined as the point on the ROC where the false match and false non-match rates are equivalent. The EER is non-dimensional, but not all biometric systems have meaningful EERs owing to the tendency of the genuine distribution to be bimodal. False match and false non-match error rates, as displayed in the ROC curve, are the only appropriate test measures for system error performance prediction.

Penetration Rate

In systems holding a large number, N, of templates in the database, search efficiencies can be achieved by partitioning them into smaller groups based both upon information contained within (endogenous to) the templates themselves and upon additional (exogenous) information, such as the customer's name, obtained at the time of enrollment. During operation, submitted samples are compared only to templates in appropriate partitions, limiting the required number of template-to-sample comparisons. Generally, a single template may be placed into multiple partitions if there is uncertainty regarding its classification. Some templates of extreme uncertainty are classified as

"unknown" and placed in all of the partitions. In operation, samples are classified according to the same system as the templates, then compared to only those templates from the database which are in communicating partitions. The percentage of the total database to be scanned, on average, for the each search is called the "penetration coefficient", P, which can be defined as

$$P = \frac{E(\text{number of comparisons})}{N}, \tag{31}$$

where E(number of comparisons) is the expected number of comparisons required for a single input sample. In estimating the penetration coefficient, it is assumed that the search does not stop when a "match" is encountered, but continues through the entire partition. Of course, the smaller the penetration coefficient, the more efficient the system. Calculation of the penetration coefficient from the partition probabilities is discussed in [16,17].

The general procedure in testing is to calculate the penetration coefficient empirically from the partition assignments of both samples and templates. Suppose there are K partitions, C_i, for i=1,2,…,K and there are L sets, S_l, l=1,2,…,L, indicating which partitions communicate. For instance, an "unknown" partition communicates with every other partition individually.

There are N_S samples, X_h , h=1,2…,N_S , and N_T templates, $\bar{X}_k$, k=1,2…,N_T. Each sample, X_h, can be given multiple partitions, $C_{h\,i}$, the precise number of which, I_h, will depend upon the sample. Similarly, each template, $\bar{X}_k$, can have multiple partitions $\bar{C}_{k\,j}$, j=1,2,…,I_k. For any sample-template pair, if any of the partitions are in a communicating set, the sample and template must be compared. However, they need to be compared at most only once, even if they each have been given multiple partitions in multiple communicating sets.

We define the "indicator" function,

$$1_l\left(C_{h\,i}, \bar{C}_{k\,j}\right) = \begin{cases} 1 \text{ if } C_{h\,i} \text{ AND } \bar{C}_{k\,j} \in \{S_l\} \\ 0 \text{ if } C_{h\,i} \text{ OR } \bar{C}_{k\,j} \notin \{S_l\} \end{cases}, \tag{32}$$

so that the function equals unity if the partitions C_h and $\bar{C}_k$ are both elements of the set S_l and zero otherwise. For each of the samples, h, and each of the templates, k, we must search all partitions, i=1,2,…,I_h, and j=1,2,…,J_k, against all L sets to determine if any $C_{h\,i}$ and $\bar{C}_{k\,j}$ communicate. However, a single sample and single template never need be compared more than once. The penetration coefficient will be

$$P = \frac{\sum_{h=1}^{N_S}\sum_{k=1}^{N_T} H\left(\sum_{i=1}^{I_h}\sum_{j=1}^{J_k}\sum_{l=1}^{L} 1_l(C_{h\,i}, \bar{C}_{k\,j})\right)}{N_S N_T}, \tag{33}$$

where H(.) is the Heavyside unity function, defined as unity if the argument is greater than zero and zero otherwise.

There may be multiple, say B, independent, filtering and binning methods, P_i, i=1,2,…,B, used in any system. If the methods are truly independent, the total penetration coefficient for the system, P_{SYS}, using all B methods, can be written

$$P_{SYS} = \prod_{i=1}^{B} P_i \,. \tag{34}$$

If correlations exist between any of the partitioning schemes, Eq. (34) will under-estimate the true penetration coefficient.

Bin Error Rate

The bin error rate reflects the percentage of samples falsely not matched against their templates because of inconsistencies in the partitioning process. This error rate is determined by the percentage of samples not placed in a partition which communicates with its matching template. For each partitioning method employed, a single test can be designed to determine the bin error rate, e. Consider N matched sample-template pairs, X_h and $\overline{X}_h$. The percentage of the pairs for which each member is placed in a communicating partition is an estimate of the complement of the bin error rate. This percentage can be computed by

$$1-e = \frac{\sum_{h=1}^{N} H\left(\sum_{i=1}^{I_h} \sum_{j=1}^{J_h} \sum_{l=1}^{L} 1_l(C_{hi}, \overline{C}_{hj}) \right)}{N}. \tag{35}$$

The bin error rate for the system, however, will increase as the number, B, of independent binning methods increase. If any one of the methods is inconsistent, a system binning error, ε_{SYS}, will result. Therefore, the probability of no system binning error over B binning methods is

$$1-e_{SYS} = \prod_{i=1}^{B} (1-e_i). \tag{36}$$

Transaction Speed

The time required for a single transaction, $T_{transaction}$, is the sum of the data collection time, $T_{collect}$, and the computational time, $T_{compute}$.

$$T_{transaction} = T_{collect} + T_{compute} \,. \tag{37}$$

For positive identification systems, only a very few comparisons between templates and submitted samples are required and generally $T_{collect} > T_{compute}$. The collection times are highly application dependent, varying from a very few seconds [18] to a couple of

minutes [19]. Transaction times are best estimated by direct measurement of the system throughput, S, as given by

$$S = \frac{1}{T_{transaction}}. \tag{38}$$

For large-scale, negative identification systems, the computational time can be much greater than the collection time. The challenge is to reduce the computational time so that the throughput is not limited by the computer hardware. The computational time can be estimated from the hardware processing rate, C, and the number of comparisons required for each transaction. If m is the number of biometric records collected and searched from each user during a transaction and N is the total number of records in the database, then

$$T_{compute} = \frac{m\, P_{SYS}\, N}{C}, \tag{39}$$

where P_{SYS} is again the system penetration coefficient. Methods for estimating hardware processing speeds are given in texts such as [20].

Confidence Intervals

The concept of "confidence intervals" refers to the inherent uncertainty in test results owing to small sample size. These intervals are *a posteriori* estimates on the uncertainty in the results on the test population in the test environment. They do not include the uncertainties caused by errors (mislabeled data, for example) in the test process. Future tests can be expected to fall within these intervals only to the extent that the distributions of $\mu_{h\,j}$ and σ_h^2, and the errors in the testing process, do not change. The confidence intervals do not represent *a priori* estimates of performance in different environments or with different populations. Because of the inherent differences between test and application populations and environments, confidence intervals have not been widely used in reported test results and are of limited value.

The method of establishing confidence intervals on the ROC is not well understood. Traditionally, as in [14], they have been found through a summation of the binomial distribution. The confidence, β, given probability p, of K distances, or fewer, out of N independent distances being on one side or the other of some threshold, τ, would be

$$1-\beta = \Pr\{i \le K\} = \sum_{i=0}^{K} \binom{N}{i} \frac{N!}{i!(N-i)!} p^i (1-p)^{N-i}. \tag{40}$$

When computing the confidence interval on the false non-match rate, for instance, K would be the number of the N independent, genuine distance measures greater than the threshold τ. The best "point estimate" of the false non-match rate would be K/N.

The probability, p, calculated by inversion (40), would be the upper bound on the confidence interval. The lower bound could be calculated from the related equation for $\Pr\{i \geq K\}$. In practice, values of N and K are too large to allow equation (40) to be computed directly and p may be too small to allow use of normal distribution approximations. The general procedure is to use the "incomplete Beta function" [21,22]

$$I_p(K+1, N-K) = \sum_{i=K+1}^{N} \binom{N}{i} \frac{N!}{i!(N-i)!} p^i (1-p)^{N-i} = \beta \tag{41}$$

and numerically invert to find p for a given N, K, and β.

One interesting question to ask is "What is the lowest error rate that can be statistically established with a given number of independent 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 gives us the 95% confidence level, β. We apply Eq. (40) using K=0,

$$0.05 > \Pr(K=0) = \sum_{i=0}^{0} \frac{N!}{i!(i-N)!} p^i (1-p)^{N-i} = (1-p)^N . \tag{42}$$

This reduces to

$$\ln(0.05) > N \ln(1-p) \tag{43}$$

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

$$N > 3/p . \tag{44}$$

Recent work indicates that while this approach is satisfactory for error bounds on the false non-match rate, where distance measures are generally calculated over N independent template-sample pairs, it cannot be applied for computing confidence intervals on false match results where cross-comparisons are used. Bickel [23] has given the confidence intervals for the false match rate when cross-comparisons are used and the templates are created from a single sample, such that M=1. For N samples, there are N(N-1) non-independent cross-comparisons. We will denote a cross-comparison distance less than or equal to the threshold, τ, by

$$r(h,k) = 1(d_{hk} \leq \tau), \tag{45}$$

where 1(.) is again the indicator function. So the best estimate of the probability, FMR(τ), of a cross-comparison being τ, or less, would be the number of such cross-comparisons divided by the total number available,

$$\hat{FMR}(\tau) = \frac{1}{N(N-1)} \sum_{h=1}^{N} \sum_{k=1}^{N} r(h,k) \text{ for } h \neq k \tag{46}$$

The (1- α)% confidence bounds are

$$\hat{\mathrm{FMR}}(\tau) \pm z_{\left(1-\frac{\alpha}{2}\right)} * \left(\frac{\hat{\sigma}(\tau)}{\sqrt{N}} \right), \tag{47}$$

where

$$\hat{\sigma}^2 = \frac{1}{N(N-1)^2} \sum_{h=1}^{N} \left(\sum_{k \neq h} r(h,k) + \sum_{k \neq h} r(k,h) \right)^2 - 4 * \left(\hat{\mathrm{FMR}} \right)^2 \tag{48}$$

and $Z_{\left(1-\frac{\alpha}{2}\right)}$ indicates the number of standard deviations from the origin required to encompass $\left(1-\frac{\alpha}{2}\right)\%$ of the area under the standard normal distribution. For α=5%, this value is 1.96. The explicit dependency on τ of all quantities in (48) has been dropped for notational simplicity.

In practice, time and financial budgets, not desired confidence intervals, always control the amount of data that is collected for the test. From the test results, we can calculate the upper bound on the confidence interval, "guess-timate" the potential effect of differences between test and operational populations and environments, then over-design our system decision policy to account for the uncertainty.

Testing Protocols

The general test protocol is to collect one template from each of N users in an environment that closely approximates that of the proposed application, ideally within the same application partitions as described in Section 2. The value of N should be as large as time and financial budget allow and the sample population should approximate the target population as closely as practicable. Some time later, on the order of weeks or months if possible, one sample from each of the same N users is collected. Then, in "off-line" processing, the N samples are compared to the N previously stored templates to establish N^2 non-independent distance measures. For all distance thresholds, τ, point estimates of the false match and false non-match error rates are given by

$$\hat{\mathrm{FMR}}(\tau) = \frac{\sum_{h=1}^{N} \sum_{k \neq h}^{N} 1\left(d_{h\,k} < \tau\right)}{N(N-1)} \tag{49}$$

and

$$\hat{\mathrm{FNMR}}(\tau) = 1 - \frac{\sum_{h=1}^{N} 1\left(d_{h\,h} < \tau\right)}{N}, \tag{50}$$

where 1(.) is the indicator function, equal to unity if the argument is true and zero otherwise, and the hat indicates the estimation. When testing from operational data, substitution of the inter-template distances, $\delta_{h\,k}$, for the impostor distances, $d_{h\,k}$, in (49) will generally result in overestimation of the false match rate.

For systems employing binning, estimates of penetration coefficient and binning error rate are estimated from Eqs. (33) and (35) by comparing partition assignments of the templates to those of the samples. Results from one test [24] on four Automatic Fingerprint Identification System (AFIS) vendors are given in Figures 3 and 4.

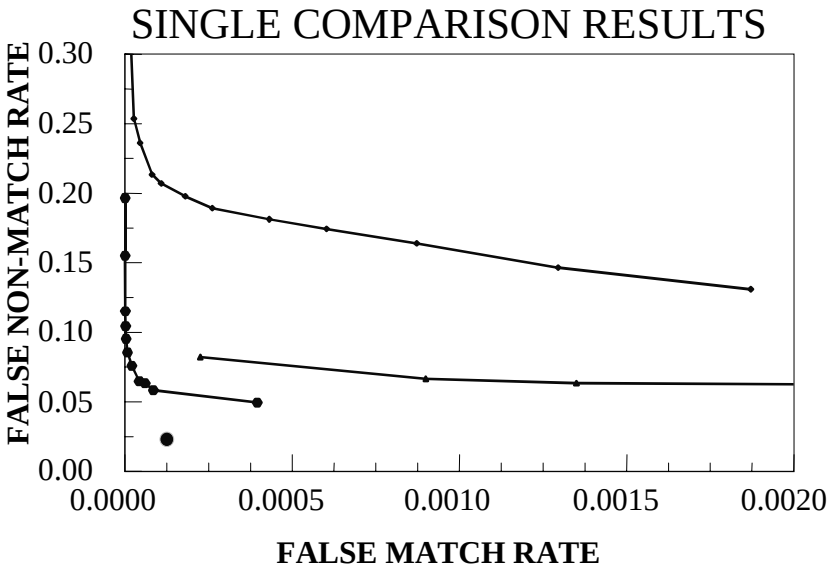


FIGURE 3 ROC curve.

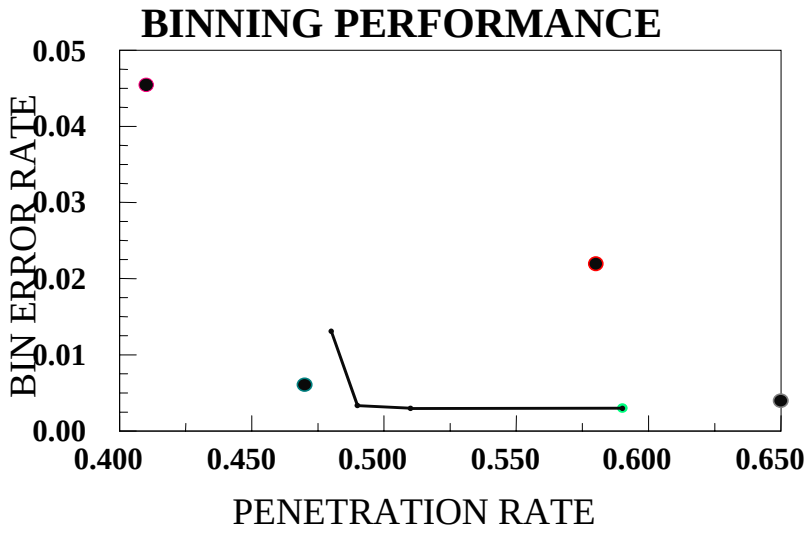


FIGURE 4 Binning error and penetration rates.

System Performance Prediction from Test Results

The five basic system performance parameters, false match rate, false non-match rate, penetration coefficient, bin error rate and transaction speed, can be used to predict system "false accept/false reject" rates and throughput under a wide variety of system decision policies [16]. Recall that the concern of every system manager is three-fold: the number of false rejections requiring human intervention, including the percentage of the population who are unable to enroll; the deterrence value of the false acceptance rate; and the ability of the system throughput rate to meet the input demand. In this section, we will consider the single example of a negative identification system using two independent biometric measures and a system policy that declares a "rejection" if both of the measures are found to match both measures of any previously enrolled individual. All calculations will assume statistical independence.

If the penetration coefficient is found to be 0.5 on each measure, then by Eq. (34), the system penetration coefficient will be P_{SYS} = 0.5*0.5=0.25. If the bin error rate is 0.01 for each measure, by Eq. (36) the system bin error rate will be e_{SYS} =1-(1-0.01)*(1-0.01)=0.02.

A false rejection occurs if both submitted samples from a single user are found to falsely match both templates of one of the previously enrolled N individuals. Assuming statistical independence of error rates, if the first sample pair is compared to the two stored templates from just one enrolled user, the chance of a false rejection, FRR, occurring is

$$\mathrm{FRR} = \mathrm{FMR}^2 . \tag{51}$$

For notational simplicity, we have not indicated the dependence of FMR on the threshold, τ. The probability of not getting a false rejection over of P_{SYS}*N searched template pairs is given by

$$1 - \mathrm{FRR} = (1 - \mathrm{FMR}^2)^{N*P_{SYS}} . \tag{52}$$

Suppose that our working estimate of the false match rate, based on testing, is 10^{-6} and that the system will be designed for N=4x10^6 users. Then, $\mathrm{FRR} \approx 10^{-6}$. The expected number of users falsely rejected during enrollment of the entire population will be $N * \mathrm{FRR} < 4$, thereby requiring limited human intervention for exception handling over the course of enrollment of the population.

Assuming that the database is "clean", meaning only one template set for any single user, a false acceptance will not occur if both samples are matched to the enrolled templates and no binning error occurs. Therefore, the complement of the false acceptance rate, FAR, can be given as

$$1 - \mathrm{FAR} = (1 - e_{SYS})(1 - \mathrm{FNMR})^2 . \tag{53}$$

If our working estimate of the false non-match rate is 7%, then FAR=15%. The number of fraudsters, F, in the system will be

$$F = N * FR * FAR, \tag{54}$$

where FR is the fraud rate, or percentage of the population that is attempting to defraud the system. The fraud rate depends not only on the inherent honesty of the population, but also on the perceived chance of getting caught. The true chance of getting caught, of course, is 1-FAR, or 85% in this example, but the perceived rate may be different. Consequently, estimation of the fraud rate is best left to social scientists. We hypothesize that a FAR of 15% is more than adequate for most real systems.

Usually, in large-scale systems, a throughput rate is specified as a system requirement and the throughput equations (38) and (39) are used to determine the necessary hardware processing speed. If our system is designed for $4x10^6$ users, we may want to enroll them over a four-year period, about 1000 days. We might design the system for a maximum capacity of 6,000 enrollments per day when the last of the users are being enrolled. We assume transaction time is controlled by the hardware processing rate. In our system, the number of samples, m, used for each individual is 2. Therefore, the processing rate, as calculated using (38) and (39), must be $1.2x10^{10}$ computations per day, if no daily backlog is acceptable. Assuming 20 hour per day availability of the processing system, the required rate will be about 170,000 comparisons per second.

Available Test Results

Results of some excellent tests are publicly available. The most sophisticated work has been done on speaker verification systems. Much of this work is extremely mature, focusing on both the repeatability of sounds from a single speaker and the variation between speakers [25-31]. The scientific community has adopted general standards for speech algorithm testing and reporting using pre-recorded data from a standardized "corpus" (set of recorded speech sounds), although no satisfactory corpus for speaker verification systems currently exists. Development of a standardized database is possible for speaker recognition because of the existence of general standards regarding speech sampling rates and dynamic range. The testing done on speech-based algorithms and devices has served as a prototype for scientific testing and reporting of biometric devices in general.

In 1991, the Sandia National Laboratories released an excellent and widely available comparative study on voice, signature, fingerprint, retinal and hand geometry systems [32]. This study was of 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 evaluation of the IriScan prototype [33] in an access-control environment.

A major study of both fingerprinting and retinal scanning, using people unacquainted with the devices and in a non-laboratory setting, was conducted by the California Department of Motor Vehicles and the Orkand Corporation in 1990 [19]. This report measured the percentage of acceptance and rejection errors against a database of fixed

size, using device-specific decision policies, data collection times, and system response times. Error results cannot be generalized beyond this test. The report includes a survey of user and management acceptance of the biometric methods and systems.

In 1996, an excellent comparative study on facial recognition systems was published by the U.S. Army Research Laboratory [34]. This study used as data facial images collected in a laboratory setting and compared the performance of four different algorithms using this data. Both test and enrollment images were collected in the same session and false match and false non-match rates are reported as a type of "rank order" statistic, meaning that the results are dependent on the size of the test database and cannot be used for general performance prediction. Earlier reports from this same project included a look at infrared imagery as well [35].

In 1998, San Jose State University released the final report to the Federal Highway Administration [24] on the development of biometric standards for the identification of commercial drivers. This report includes the results of an international automatic fingerprint identification benchmark test.

The existence of a dozen annual industry conferences, including the U.S. Biometric Consortium and the European Association for Biometrics meetings and the CardTech/SecurTech conferences, in addition to other factors such as the general growth of the industry, has encouraged increased informal reporting of test results

Conclusions

The science of biometric device analysis and testing 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.

Acknowledgements

The framework for a scientific approach to biometric testing was established several years ago in a series of questions posed to the biometric identification community by Joseph P. Campbell, Jr. This paper was created as a response to those questions. The mathematical notation and approach was suggested by Peter Bickel. Thoughtful input by Kang and Barry James of the University of Minnesota, Deluth, and assistance from Jim Maar and MAJ. John Colombi, USAF, was most helpful and appreciated. The author, however, claims sole credit all errors and omissions.

References

[1] W. W. Peterson, and T. G. Birdsall, "The Theory of Signal Detectability," Electronic Defense Group, U.niversity of Michigan, Technical Report 13 (1954)
[2] W. P. Tanner, and J. A. Swets, "A Decision-Making Theory of Visual Detection," *Psychological Review*, Vol. 61, (1954), pg. 401-409
[3] D. M. Green, and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, 1966),
[4] J. A. Swets, (editor), *Signal Detection and Recognition by Human Observers* (Wiley, 1964)
[5] J. P. Egan, *Signal Detection Theory and ROC Analysis*, (Academic Press,1975)
[6] J.L. Wayman, "A Dual Channel Approach to Speaker Verification," *Proceedings Speech Research Symposium XII*, June 1992, Rutgers University
[7] J. G. Daugman, "Biometric personal identification system based on iris analysis," *U.S. Patent 5291560*, 1994
[8] S.S. Wilks, *Mathematical Statistics*, (Wiley and Sons, New York, 1962) pg. 83
[9] J.P. Campbell, address to the 9th Biometric Consortium meeting, Crystal City, VA, April 1997
[10] J. Williams, *Glossary of Biometric Terms*, Security Industry Association, 1995
[11] W. Shen, et. al, "Evaluation of Automated Biometrics-Based Identification and Verification Systems," *Proceedings IEEE*, Vol. 85, Sept. 1997, pp. 1464-1479.
[12] P. Bickel, response to NSA-MSP Problem #97-21, University of California, Berkeley, Department of Statistics, 1998.
[13] C. Franzen, "Convolution Methods for Mathematical Problems in Biometrics," U.S. Naval Postgraduate School Technical Report , 1998
[14] J. Williams, "Proposed Standard for Biometric Decidability," *Proceedings CTST'96*, pp. 223-234
[15] S. Kullback and R. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol.22, (1951), pp. 79-86
[16] J.L. Wayman, "Error Rate Equations for the General Biometric System," to appear in *Automation and Robotics Magazine*, Special Issue on Biometric Identification, January 1999
[17] K. James and B. James, *NSA SAG Problem 97-25*, UC Berkeley, Dept. of Statistics. Monograph, 1998
[18] Presentation by Dan Welsh and Ken Sweitzer, of Ride and Show Engineering, Walt Disney World, to CardTech/SecurTech '97 , May 21, 1997.
[19] Orkand Corporation, "Personal Identifier Project: Final Report," April 1990, State of California Department of Motor Vehicles report DMV88-89, reprinted by the U.S. National Biometric Test Center.
[20] J.L. Hennessy and D.A. Patterson, *Computer Architecture: A Quantitative Approach*, 2nd ed., (Morgan Kaufman, San Francisco, 1996)
[21] M. Abromowitz and I. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables," (John Wiley and Sons, New York, 1972)
[22] W.H. Press, *et al*, *Numerical Recipes*, 2nd ed., (Cambridge University Press, Cambridge, 1988)
[23] P. Bickel, personal correspondence in response to NSA-SAG Problem #97-23, March 1998

[24] J.L. Wayman, "Biometric Identifier Standards Research Final Report," College of Engineering, San Jose State University, October, 1997, sponsored by the Federal Highway Adminstration.
[25] B. Atal, "Automatic Recognition of Speakers from Their Voices," *Proceedings IEEE*, Vol. 64, (1976), pp. 460-475
[26] A. Rosenberg, "Automatic Speaker Verification," *Proceedings IEEE*, 64, (1976), pg. 475-487
[27] N. Dixon and T. Martin, *Automatic Speech and Speaker Recognition* (IEEE Press, NY, 1979)
[28] G. Doddington, "Speaker Recognition: Identifying People by Their Voices", *Proceedings IEEE*, Vol. 73, pp. 1651-1664, 1985.
[29] A. Rosenberg and F. Soong, "Recent Research in Automatic Speaker Recognition" in S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing* (Marcel Dekker, 1991)
[30] J. Naik, "Speaker Verification: A Tutorial," *IEEE Communications Magazine*, (1990), pp. 42-48
[31] J.P.Campbell, Jr., "Speaker Recognition: A Tutorial," *Proceedings IEEE*, Vol. 85, September 1997, pp. 1437-1463
[32] J.P. Holmes, et al, "A Performance Evaluation of Biometric Identification Devices," Sandia National Laboratories, SAND91-0276, June 1991.
[33] F. Bouchier, J. Ahrens, and G. Wells, "Laboratory Evaluation of the IriScan Prototype Biometric Identifier," Sandia National Laboratories, SAND96-1033, April 1996
[34] P.J. Phillips, et al., "FERET (Face-Recognition Technology) Recognition Algorithm Development and Test Results," Army Research Laboratory, ARL-TR-995, October 1996
[35] P.J. Rauss, et al., "FERET (Face-Recognition Technology) Recognition Algorithms," *Proceedings of ATRWG Science and Technology Conference*, July 1996

Confidence Interval and Test Size Estimation for Biometric Data

James L.Wayman, Director
National Biometric Test Center

I. The Uncertainty in ROC Confidence Estimation

The standard method for expressing the technical performance of a biometric device for a specific population in a specific application is the "Receiver Operating Characteristic" (ROC) curve. Methods for establishing "confidence intervals" (uncertainty bounds) on the ROC are not well understood [1,2]. Each point on the ROC curve is calculated by integrating "genuine" and "impostor" score distributions between zero and some threshold, t. Traditionally [2,3], confidence intervals for the ROC at each threshold, t, have been found through a summation of the binomial distribution under the assumption that each comparison represents a "Bernoulli trial". The confidence, β, given a **non-varying** probability p, of K sample/template comparison scores, or fewer, out of N **independent** comparison scores being in the region of integration would be

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

In practice, (1) is calculated using the "incomplete Beta function" representation [4,5] to avoid the numerical problems with factorial computation for high values of N.

Equation (1) might be inverted to determine the required size, N, of a biometric test for a given level of confidence, β, if the error probability, p, 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 Law" 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 [6], provided that (1) is applicable.

The comparison of biometric measures will not be Bernoulli trials and equation (1) will not be applicable if: 1) trials are not independent; 2) the error probability varies across the population. Trials will not be independent if users stop after successful use and continue after non-successful use. Further, if cross-comparisons (all samples compared to all templates except the matching one) [7] are used to establish the "impostor distribution", the comparisons will not be independent. In either case, the biometric comparisons are not Bernoulli trials and the cumulative binomial distribution will not apply The varying error probability across the population [8,9] ("goats" with high false non-match errors and "lambs" with high false match errors) similarly invalidates the cumulative binomial equation as appropriate for developing uncertainty bounds. An equation for confidence intervals in the more general case of cross comparisons and population-varying error probability has been given by Bickel in [10].

Assume that each user is represented by one sample and one template. Let $d_{h,k}$ be the score or distance measurement between the sample of user h and the template of user k. If $h \neq k$, then the scores will represent "impostor" comparisons. If $h = k$, the scores represent "genuine" comparisons. Then we'll define

$$r(h,k)=\begin{cases}0, \text{if } d_{h,k} > \tau \\ 1, \text{if } d_{h,k} \le \tau\end{cases} \tag{2}$$

When cross comparisons , comparing N samples to the N-1 non-self templates, are used to establish the false match error rate, the "nominal" value of the error probability, $p(\tau)$, at any threshold, τ, is given using this notation as

$$p(\tau)=\frac{1}{N(N-1)}\sum_{h=1}^{N}\sum_{k=1}^{N} r(h,k) \text{ for } h \neq k \tag{3}$$

Bickel establishes the uncertainty bounds on this value of $p(\tau)$, as

$$\pm z_{\left(1-\frac{\alpha}{2}\right)} * \left(\frac{\hat{\sigma}(\tau)}{\sqrt{N}}\right) \tag{4}$$

where

$$\hat{\sigma}^2=\frac{1}{N(N-1)^2}\sum_{h=1}^{N}\left(\sum_{k\neq h} r(h,k)+\sum_{k\neq h} r(k,h)\right)^2 - 4p^2 \tag{5}$$

σ is bounded by $2p(1-p)/N$. $Z_{\left(1-\frac{\alpha}{2}\right)}$ indicates the number of standard deviations from the origin required to encompass $\left(1-\frac{\alpha}{2}\right)\%$ of the area under the standard normal distribution. For α=5%, this value is 1.96. The explicit dependency on τ of all variables (except N) in (5) has been dropped for notational simplicity.

Equation (5) requires empirical data, and consequently, will not allow the inversion of (4) to establish required test size N, even if error rates could be accurately estimated in advance of the test.

II. Experimental Test

To test the variation of the true confidence bounds from those predicted by (1) and (4), we repeatedly sampled from a large data set obtained from an existing biometric application. The data set consisted of user identification numbers and vector samples and templates for each user interaction with the biometric application. This data was arbitrarily edited to remove: 1) outliers indicative of a hardware failure or a "failure to acquire" condition (for instance, null vectors); 2) subsequent uses by a single identification number. We are assuming that single individuals have only one identification number in the system. This is not required by the system, but is a reasonable assumption because there is no general motivation for multiple enrollments. We are also assuming that there are no impostors among the users. Although this assumption is less reasonable and there are no estimates of the rate of occurrence of impostor transactions, we feel that the incidence is probably low. Nonetheless, our results may be affected by violations of this assumption.. With these cautionary notes, we treat the 48,478 remaining records as sample-template vector pairs from genuine transactions of distinct individuals.

This editing of the data makes each genuine transaction "independent". Elimination of "subsequent" user attempts, removes the impact of the "use-until-successful" stopping rule. It also eliminates the impact of "goats", as each user has only one trial, and there is a "one-to-one" correspondence between trials and users. Each genuine transaction can be taken as an event with uniform error probability

$$\bar{p}(\tau) = \frac{1}{N}\sum_{h=1}^{N} p_h(\tau) \qquad (6)$$

where $p_h(\tau)$ is the error probability for each user h=1,2…N. If multiple transactions from each user were allowed, each transaction would have an error probability chosen randomly from the set of user error probabilities { $p_1(\tau)$, $p_2(\tau)$….$p_N(\tau)$}, with selection weighted by the frequency of each user's transactions. Consequently, our test, as constructed, can say nothing about the impact of "goats" on the confidence intervals.

A further complication with possible impact on our results is that the user templates are updated after each successful use of the system by averaging the sample vectors into the templates with some unknown weighting. We have no information for any user as to how many successful transactions have been averaged into the template.

The "true" ROC curve for the system was established from the 48,478 sample-template records using each as a "genuine" transaction and randomly combining each of the 48,478 samples with a non-matching template as the "impostor" transactions.

Using four values of N (50, 100, 200, 400), 600 trials were conducted. Each trial consisted of a random selection of N sample-template pairs from the database without replacement. Replacement would have allowed the possibility that a single individual could be compared to himself as an impostor. "Genuine" distributions were established from the N sample-template pairs. "Impostor" distributions were established in three distinct ways: 1) Using all N(N-1) cross comparisons; 2) Using ½ N(N-1) cross comparisons, so that each pair of individuals would be compared only once[1]; 3) Using N random assignments of samples and templates from different individuals. For each trial, three ROC curves were developed, each using the same "genuine" distribution, but with different "impostor" distributions. So for each N, 600 ROC's were developed with each of the three methods. The number 600 was taken as our arbitrary trade-off between increasing accuracy and computational time.

For each N and each method, the 600 ROC's were sorted at each threshold to empirically establish the 0.025% upper and lower limits on their values. The region between these values corresponds to the 95% confidence bound. This approach to interval estimation is not very robust and may lead to substantial variation in estimates depending upon the particular 600 trials used.

III. Results

Figure 1 shows that the mean ROC over the 600 trials closely approximates the "true" ROC for each N. Figures 2-5 shows good agreement between the sampled

---

[1] We compare the sample of user 1 to the template of user 2, but do not allow the sample of user 2 to be compared to the template of user 1.

confidence intervals on the false non-match rate with those calculated from (1) over the N independent comparisons. This verifies that our data editing produced the equivalent of independent transactions at a fixed error rate. This seems to support the claim that cross-comparisons produce unbiased estimates of the threshold-dependent false match error rates.

Figures 6-9 show good agreement between the binomial confidence intervals on the false match rate and the sampling tests when the N random-selection technique is used for impostor comparisons. This shows that the impostor comparisons using this method were independent.

Figures 10-11 shows very poor agreement between the binomial confidence interval on the false match rate and sampling tests when the ½ N(N-1) technique is used. For brevity, only the N=50 and N=100 cases are graphed, but the N=200 and N=400 differences between binomial and sampling uncertainty bounds are even more pronounced. Use of '(1), with N taken as the total number of comparisons [½ N(N-1)] , underestimates the expected uncertainty. Comparison with Figures 6 and 7 show that the ½ N(N-1) confidence interval is overestimated by (1) using as N the number of samples.

Figures 12-13 show very poor agreement between the binomial confidence bounds on the false match rate and the sampling tests when N(N-1) cross comparisons are used. Use of (1), with N taken as the total number of comparisons, grossly underestimates the expected uncertainty. Only N=50 and N=400 cases are shown for brevity, and Figure 13 is rescaled for clarity. Comparing with Figures 6 and 7, we see that the sampling confidence interval decrease significantly when cross-comparisons are used. Therefore, the binomial confidence interval calculated with N taken as the number of samples, overestimates the uncertainty in the false match rate when cross comparisons are used.

Figures 14-17 shows the "true" ROC falling within the confidence interval predicted from the Bickel equations (3)-(5) for a single, randomly chosen trial at each value of N. Further, the widths of the confidence intervals show good agreement for all values of N tested.

IV. Conclusions

We can conclude the following:

1. When each of N users has one trial, the cumulative binomial distribution adequately models the false non-match rate ROC confidence interval.
2. When each of N samples (one from each user) is randomly paired with a non-matching template for the impostor comparisons, the cumulative binomial distribution adequately models the false match rate ROC confidence intervals.
3. The use of cross comparisons seems to produce unbiased estimates of the false match error rate.
4. Use of N(N-1) cross comparisons in estimating the impostor distribution decreases the width of the confidence interval, and consequently is a more efficient estimator of the false match rate than N random impostor pairings.
5. When N(N-1) cross comparisons are used, the false match rate confidence interval is grossly overestimated by the cumulative binomial distribution calculated using as N the number of data samples.

6. When N(N-1) cross-comparisons are used, the false match rate confidence interval is grossly underestimated by the cumulative binomial distribution calculated using as N the total number of comparisons.
7. When ½ N(N-1) cross-comparisons are used, the false match rate confidence interval is grossly underestimated by the cumulative binomial distribution calculated using as N the total number of comparisons.
8. When ½ N(N-1) cross-comparisons are used, the false match rate confidence interval is grossly overestimated by the cumulative binomial distribution calculated using as N the number of data samples.
9. The Bickel equations appear to accurately bound the "true" ROC curve when N(N-1) cross comparisons are used.
10. These tests have not considered the false non-match rate confidence interval when multiple attempts from each user are allowed.

V. References

[1] J.L. Wayman, "Fundamentals of Biometric Technologies", *Proc. CTST'99*, Chicago, May, 1999, pg. 390-410. Also available on-line at www.engr.sjsu.edu/biometrics/

[2] K.V. Diegert, "Estimating Performance Characteristics of Biometric Identifiers", *Proc. Biometric Consortium 8*, San Jose, CA, June, 1996

[3] W. Shen, etal, "Evaluation of Automated Biometrics-Based Identification and Verification Systems", Proc. IEEE, vol.85, Sept. 1997, pg. 1464-1479.

[4] M. Abromowitz and I. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", (John Wiley and Sons, New York, 1972)

[5] W.H. Press, *et al*, Numerical Recipes, 2nd ed., (Cambridge University Press, Cambridge, 1988)

[6] J. E. Porter, "On the '30 error' criterion", ITT Industries Defense and Electronics Group, April 1997, available from the National Biometric Test Center

[7] L. O'Gorman, "Fingerprint Verification", in A. Jain, etal (eds), Biometrics: Personal Identification in a Networked Society, (Kluwer Academic Press, 1998)

[8] G. Doddington, etal " Sheep, Goats, Lambs and Wolves: An Analysis of Individual Differences in Speaker Recognition Performance", ICSLP'98, Sidney, Australia, November 1998

[9] J.L. Wayman, "Multi-Finger Penetration Rate and ROC Variability For Automatic Fingerprint Identification Systems", unpublished NBTC report, August1999.

[10] J.L. Wayman, "Technical Testing and Evaluation of Biometric Identification Devices", in A. Jain, etal , op.cite.

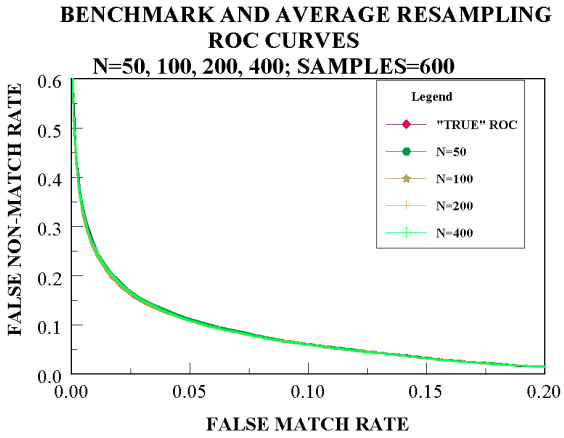


Figure 1

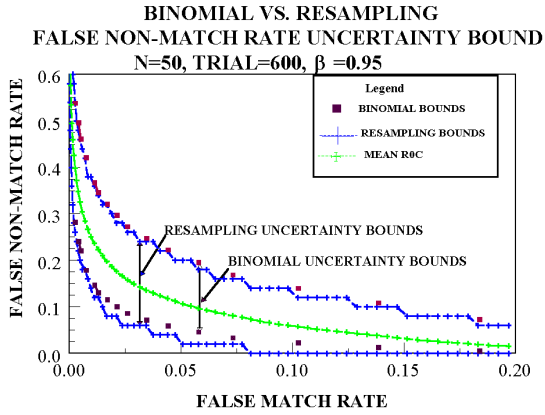


Figure 2

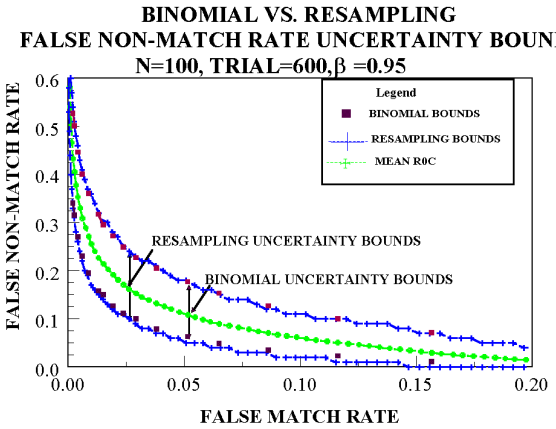


Figure 3

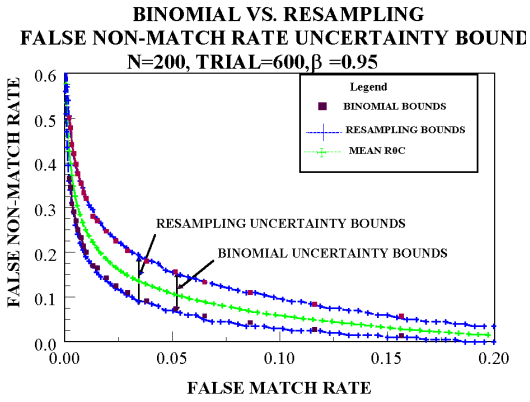


Figure 4

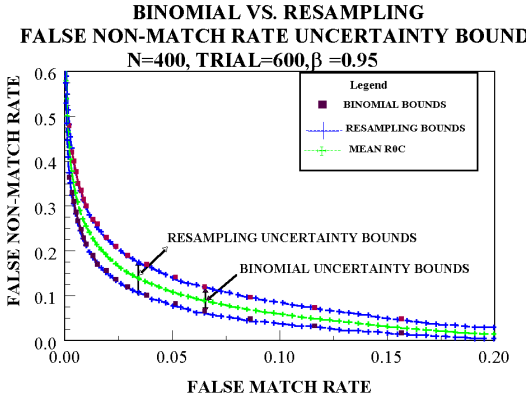


Figure 5

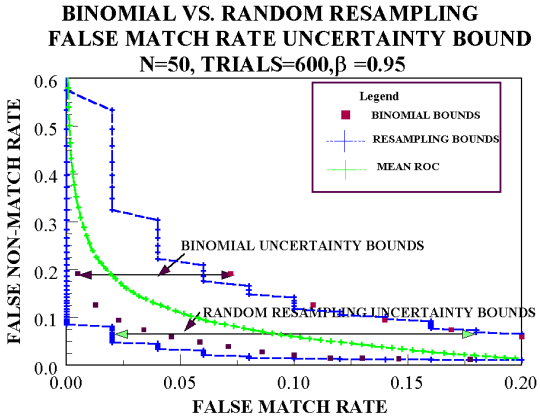


Figure 6

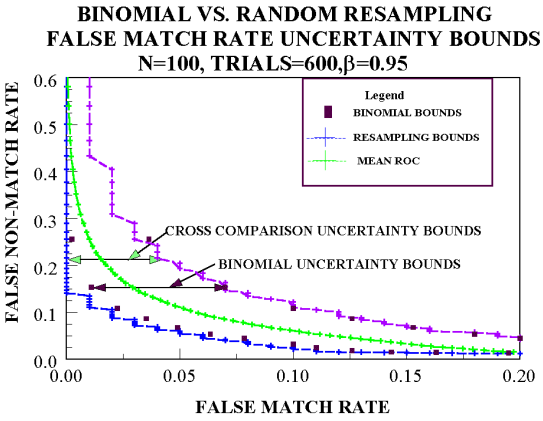


Figure 7

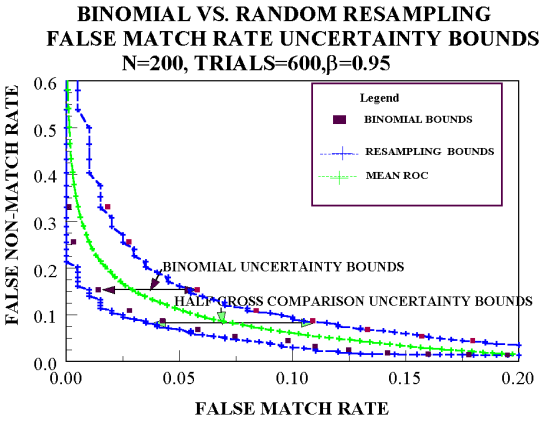


Figure 8

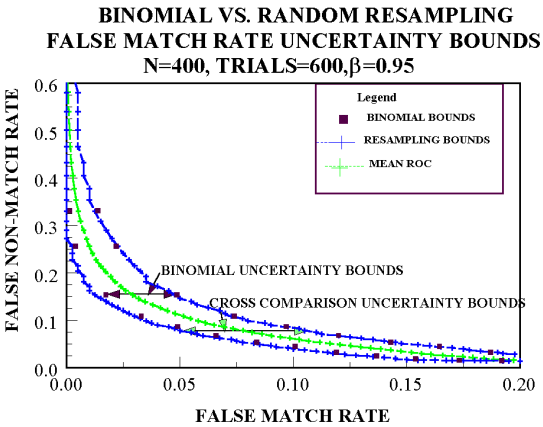


Figure. 9

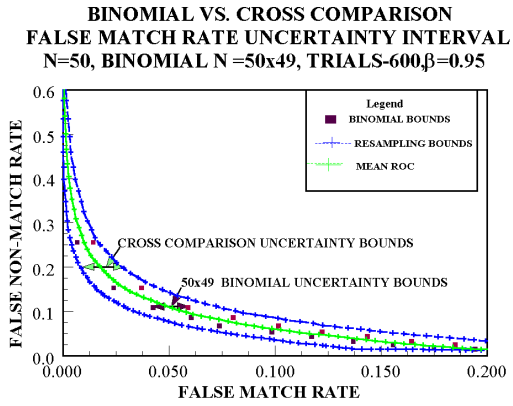


Figure 10

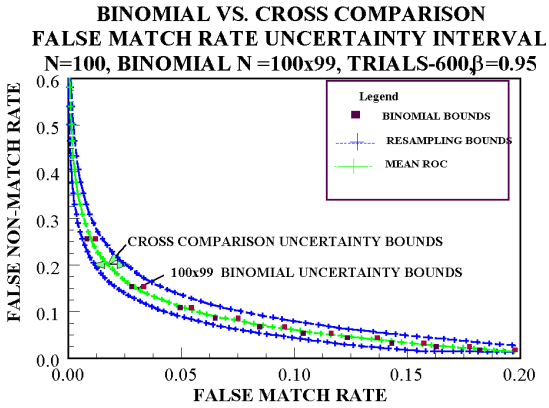


Figure 11

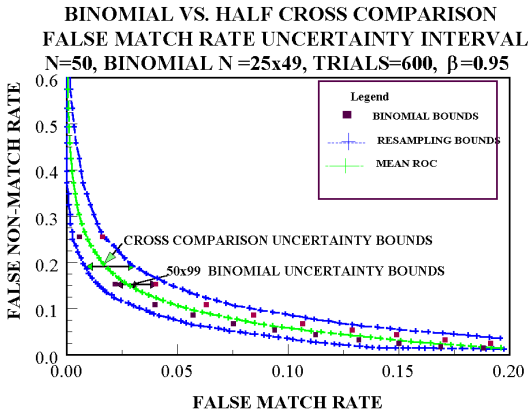


Figure 12

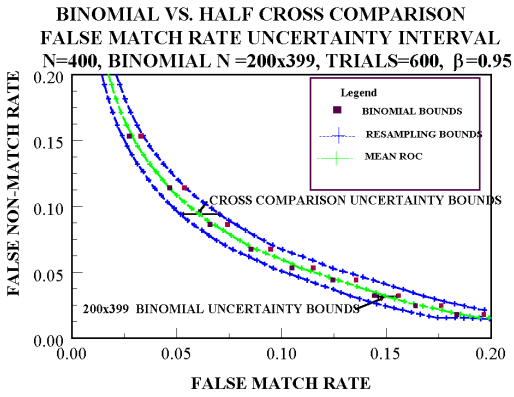


Figure 13

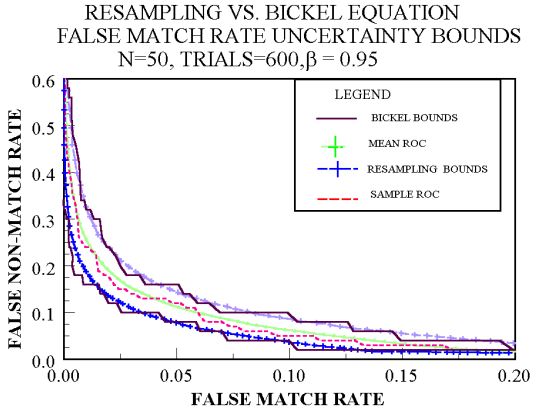


Figure 14

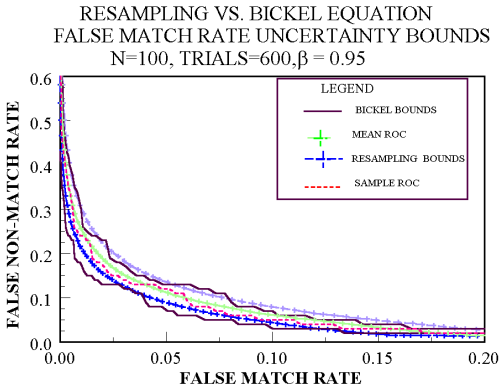


Figure 15

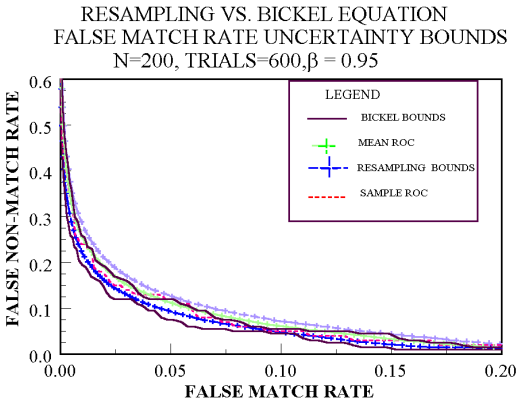


Figure 16

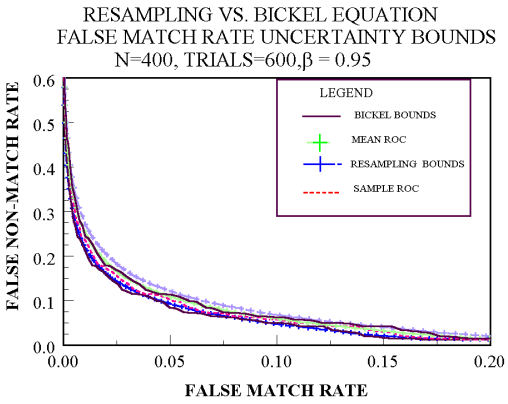


Figure 17

Error Rate Equations for the General Biometric System

James L. Wayman, Director
U.S. National Biometric Test Center

ABSTRACT

This paper will derive two equations for error rate prediction in the general M-to-N biometric identification system: one for system false match rate, and one for system false non-match rate. Under the simplifying, but approximate, assumption of statistical independence of all errors, independent variables are bin error rate, penetration rate, sample-template ("genuine") and "impostor" distance distributions, number of active templates or user models in the database, N, and the number of samples submitted for each transaction, M

Depending upon the system, each of the users might enroll one or many biometric measures or models. These measures might be different presentations of the same biometric pattern, or representations of independent biometric patterns, perhaps acquired using different biometric technologies. During use, each user might present multiple patterns for comparison with the database.

We use the word "binning" to indicate database partitioning based on information contained within (endogenous to) the biometric patterns. We reserve the word "filtering" for partitioning through the use of exogenous information about the user not discernable from the biometric patterns. The penetration rate is the total search efficiency gained by both binning and filtering. For systems with a large number of users, U, the system throughput rate is seen to be dependent upon both the hardware processing speed and the penetration rate. Error rates are dependent upon the penetration rate and are thus linked to the hardware processing speed through the throughput equation.

The general error rate equations will be shown to degenerate to those previously published [1,2] for: the one-to-one (so-called "verification") case, when M=N=1; the one-to-N (so-called "identification") case, when M=1; and the three-to-one ("three strikes and you're out" verification) case, when M=3 and N=1. Examples, using parameters established in previous studies, are given.

0.0 NOTATION

N	number of active stored templates or models in the database
M	number of samples submitted during each transaction
m	number of samples used in an initial search
U	number of active, enrolled users
T	number of independent templates or models stored as an ensemble for each enrolled user
K	number of partitions in a filtering or binning method
B	number of binning and filtering methods
p_i	probability that a sample will be in the i^{th} partition
P_i	penetration rate owing to the i^{th} filter or binning method
P_{sys}	system penetration rate

ε_j	bin error rate of the j^{th} bin
ε_{sys}	system bin error rate
D	similarity or distance measure
Ψ_G (D)	"genuine" distance distribution function
Ψ_I (D)	"impostor" distance distribution function
Ψ_T (D)	inter-template distance distribution function
τ	similarity or distance score threshold
FMR(τ)	false match rate: the probability that a sample will be mistakenly matched with a non-self template.
FNMR(τ)	false non-match rate: the probability that a sample will be mistakenly judged not to match a self template when compared.
FNM_i	the probability that the i^{th} sample will be falsely not matched because of binning or matching errors.
Q	number of matches required by decision policy to declare an identification
C	hardware comparison rate
S	system throughput rate

1.0 INTRODUCTION

The function of a biometric identification system is verify claims of "customers" (users) that they are who they say they are, or are not who they say they are not. More specifically, the biometric system seeks to verify a customer's claim that her/his physiological or behavioral characteristics do or do not match those of some number of previously enrolled individuals. In the literature of biometric identification, a distinction is made between "verification" and "identification", "one-to-one" and "one-to-many" matching, based on whether the size of the searched database is one or more than one. Past testing [3-18] of biometric devices has focused on measuring "false acceptance" and "false rejection" rates, or developing "candidate lists" [14,15], in either "one-to-one" or "one-to-many" tests, often using unreported system decision policies. Device performance is often convolved with test design and system decision policy, making results difficult, or impossible, to compare between tests. Needed is a consistent approach which clearly de-couples device performance from the size of the search, the test design and the decision policy.

The general biometric system allows a single user to enroll multiple measures or multiple presentations of the same measure, and, during operation, to input multiple samples for matching. Consequently, the general system may perform multiple comparisons, even when the customer is claiming to match a single identity. A single mathematical system model of throughput and error rates, using common, testable measures, can be constructed for the general "M-to-N" biometric system in which both the "one-to-one" and "one-to-many" models are seen as degenerate cases. Here, M refers to the number of samples submitted for each transaction, and N refers to the number of active templates or user models in the database. The M samples will be of one or more physiological or behavioral characteristics. Multiple characteristics might be acquired using different biometric technologies. We call this collection of samples a "sample ensemble".

There are U active, enrolled users and T stored templates or models for each. Like the submitted samples, the template ensemble will consist of one or more biometric characteristics. We call the set of T templates a "template ensemble. In this paper, to limit complexity, we will consider the T templates in an ensemble to be independent. This is rarely, if ever, the case in reality, where interdependencies between the models may be subtle. The effect of the interdependence of the T models on the comparison error rate is very difficult to estimate from current data and, at the cost of inaccuracies in the equations developed, will largely be ignored in this paper. We will leave for future studies the most general case of template ensembles containing multiple representations of several multiple, dependent characteristics[1].

Although some systems allow the number of stored templates to vary over the enrolled individuals, in this paper we will assume that T is fixed by s $N = T * U$ stem policy so that

$$N = T * U \tag{1}$$

Depending upon the system enrollment policy, each of the templates might be created from a single enrollment sample, from multiple samples given in a single enrollment session, or as a weighted moving average of samples submitted during use over time.

In previous papers [1,2,19,20], we developed a general system description and governing equations for the "one-to-one" and "one-to-many" systems. The goal of this paper is to derive general error rate and throughput equations for the more general "M-to-N" system under a variety of decision policies. It will be seen that error rates are strongly impacted by the system throughput requirements; that is, that system speed and error rates are closely related.

2.0 BASIC MEASURES

There are five important, interrelated measures that govern the performance of the general biometric system. These are: 1) the "penetration coefficient", reflecting the expected proportion of the enrolled ensembles to be compared to a single input sample; 2) the "bin error rate", or probability that a search for a matching template in the database will be unsuccessful because the sample and template were erroneously placed different "bins"; 3) the single comparison false match rate, or probability that a non-self ("impostor") template will be incorrectly matched to a sample; 4) the single comparison false non-match rate, or probability that a truly matching template will be missed; 5) the comparison rate (sample-template comparisons per unit time) of the hardware, perhaps averaged over a time period long enough to include system availability considerations.

2.1 System Penetration Rate

Search efficiencies can be achieved by partitioning the N templates into smaller groups based both upon information contained within (endogenous to) the templates themselves and upon additional (exogenous) information, such as the customer's name,

---

[1] We will also not consider "cohort" modeling techniques, often used in speaker recognition systems, wherein single input samples are compared to multiple models in a closely related subset of users.

obtained at the time of enrollment. During operation, submitted samples are compared only to templates in appropriate partitions, limiting the required number of sample-to-template comparisons. We refer to partitioning based on exogenous information as "filtering" and reserve the word "binning" for the use of endogenous information

Generally, a single template may be placed into multiple partitions if there is uncertainty regarding its classification. Some templates of extreme uncertainty as to classification are labeled as "unknown" and placed in all of the partitions. In operation, samples are classified according to the same system as the database, then matched against only those templates from the database which are in the same classification or classifications. The portion of the total database to be scanned, on average, for the each search is called the "penetration coefficient", P, which can be defined as

$$P = \frac{E[\text{number of comparisons}]}{N} \tag{2}$$

where E{number of comparisons} is the expected number of comparisons required for a single input sample.

In estimating the penetration rate, it is assumed that the search does not stop when a "match" is encountered, but continues through the entire partition. Of course, the smaller the penetration rate, the more efficient the system.

Methods used to partition the database will depend upon the operational purpose of the system. In "verification" systems, where the goal is to verify the customer's claim to a specified identity, template ensembles might be stored on a card in the customer's possession. For each transaction, the database is simply the T templates on the card. In other systems of similar purpose, the templates for all enrolled users are stored centrally. In such a system, it is possible to partition the database of N templates into U partitions, based on the claimed identity of the enrollee. In such a case, where templates are placed exclusively in one partition and each partition contains the same number of templates, the penetration rate, P, owing to the filter is

$$P = \frac{1}{K} \tag{3}$$

where K is the number of partitions. In the case where K = U, combining equations (1), (2) and (3) shows that the expected number of comparisons required of an input sample is T, the size of the user's enrolled sample ensemble. The equations developed in this paper, therefore, are independent of the architecture chosen for storage.

In systems where the goal is to verify the customer's claim to an unspecified enrolled identity, or the negative claim of no enrolled identity, we might consider the total database as T partitions of U templates each. Each template in each group is linked to templates in each of the T-1 other groups through the identity of the enrolled user. For example, consider a system for verifying customer's negative claims of no enrolled identity in which fingerprint templates from left and right index fingers of each of U persons are stored. In this case, T= K = 2. Data in each partition will be linked by the identity of the enrollee. Separation of left and right prints is based on information not found in the prints themselves, so partitioning is a "filtering" operation performed at the time of enrollment. As in the previous example, the bins are exclusive and there is

equality in the partition assignments so equation (3) applies. By equations (1), (2) and (3), the expected number of searches per input sample is seen to be U.

For more general filtering and binning, however, such as the partitioning of the database by gender[2], equality in partition size does generally not apply and equation (3) cannot be used. A more general approach must be taken. If there are K partitions and p_i is the positive probability that a template is placed in the i^{th} partition, then the i^{th} partition will hold $N*p_i$ templates. If the samples and templates are from the same population, p_i is also the probability that the sample is in the i^{th} partition. If a sample or template can only be placed in a single partition, then

$$\sum_{i=1}^{K} p_i = 1 \tag{4}$$

In cases where (4) holds and the partitions are exclusive (no "unknown" partition), the expected number of comparisons is

$$E\{\text{number of comparisons}\} = \sum_{i=1}^{K} p_i N p_i = N \sum_{i=1}^{K} p_i^2 \tag{5}$$

and the penetration rate, P, can be seen to be

$$P = \sum_{i=1}^{K} p_i^2 \tag{6}$$

We will now consider the case where the K^{th} bin represents an "unknown" classification. The unknown bin must always be searched and samples classified as "unknown" must be searched against all templates regardless of bin. Nonetheless, equation (4) continues so hold. So, the expected number of comparisons becomes

$$E\{\text{number of comparisons}\} = N * p_K + \sum_{i=1}^{K-1} p_i N(p_i + p_K) = N\left[p_K + \sum_{i=1}^{K-1} (p_i + p_K) p_i \right] \tag{7}$$

The term in brackets on the right-hand side is the penetration rate under these conditions.

In the case of samples and templates of ambiguous, but not completely unknown, the general procedure is to place them into multiple bins, such that equation (4) does not hold. Rather,

$$\sum_{i=1}^{K} p_i > 1 \tag{8}$$

---

[2] Currently, only speaker verification can perform gender-based partitioning on the basis of information within the biometric pattern. For other technologies, gender-based filtering must be done on the basis of information given by the customer or by assessment of the system supervisory personnel.

and equations (5) through (7) do not hold. Calculation of the penetration rate as a function of bin probabilities, p_i , under the more general condition expressed by (8) has been given in [21]. Penetration coefficient can be calculated empirically from the binning assignments of both samples and templates by

$$P_{AVE} = \frac{\sum^{M} \text{samples} \sum^{N} \text{templates with bin(s) in common with sample}}{M\ N} \tag{9}$$

where P_{AVE} is the average, or expected value, taken over all users.

There may be multiple, say B_i, independent, filtering and binning methods used with each biometric measure in the ensemble. We will therefore add two subscripts to the penetration rate, $P_{i,i}$, to indicate the i^{th} meaure and the j^{th} binning or filtering method. If the methods are truly independent, the total penetration rate, P_i, for the i^{th} measure, using B_i methods, can be written

$$P_i = \prod_{j=1}^{B_i} P_{i,j} \tag{10}$$

If positive correlations exist between any of the partitioning schemes, equation (10) will under-estimate the true penetration rate, meaning that the real penetration rate will be higher (worse).

The above equation applies to systems that use any ensemble size, T. In those systems where T>1 and M=T, partitioning of the can be done on the basis of the classification of the entire ensemble of the independent samples and templates. That is, a sample ensemble can be compared to only those template ensembles which are partitioned similarly on all measures. We call this "ensemble binning". The system penetration rate for the ensemble becomes

$$P_{ensemble} = \prod_{i=1}^{T} P_i \tag{11}$$

2.2 Bin Error Rate

The bin error rate reflects the percentage of samples falsely not matched against the database because of inconsistencies in the binning process. This error rate can be easily measured by comparing binning partitions assigned for samples and matching templates. In general, the more bins that are used, the greater the probability that the bins will be inconsistently applied to truly matching measures. Errors are a function of the action of the bin classification algorithms on the input sample, and consequently, methodologies for inducing such errors are difficult to predict without a thorough knowledge of the algorithm.

Filtering errors, such as in the classification of an individual as "male" or "female", are due to mistakes in the externally collected data, generally made by human operators during the customer interview process. Like binning errors, filtering errors also cause samples to be falsely not matched to templates in the database. With filtering, however, system vendors can "externalize" these errors, blaming them on the data collection process of the system administrator, not on the computational algorithms.

Filtering allows the beneficial decrease in the penetration rate without the responsibility for the associated increase in false non-match error rate. Individuals wishing not to be matched to previously enrolled templates can increase the probability of filtering errors through deliberate actions. Thus, the use of filtering can create system vulnerabilities to fraud that generally do not occur with binning. Because filtering errors are the result of inconsistencies in human judgement or deliberate fraud, they cannot be easily measured by purely technical tests and will not be considered in this paper.

Binning errors can be measured by determining the percentage of truly matching biometric patterns that are placed by the system in non-communicating bins. For each binning method employed, a single test can be designed to determine the binning penetration rate, by equation (9), and the bin error rate, ε, calculated by,

$$\varepsilon = \frac{\text{number of inconsisently binned sample} \rightarrow \text{template pairs}}{\text{number of sample} \rightarrow \text{template pairs tested}} \tag{12}$$

In a system using multiple binning methods on a single measure, not to make a bin error requires that none of the individual binning methods produce an error. This awkward English actually best describes the underlying probabilistic relationship

$$1-\varepsilon_i = \prod_{j=1}^{B_i}(1-\varepsilon_j) \tag{13}$$

where ε_i is the bin error rate on the i^{th} measure, ε_j is the bin error rate for the j^{th} of the B_i binning method used on that measure. Equation (13) assumes that bin errors are independent. If the B_i methods have the same bin error rate, ε, equation (13) can be rewritten as

$$\varepsilon_i = 1-(1-\varepsilon)^{B_i} = B_i * \varepsilon - O(\varepsilon^2) \tag{14}$$

where $O(\varepsilon^2)$ indicates terms of order ε^2 and smaller. For small ε, as is the general case, equation (14) reduces to

$$\varepsilon_i \approx B_i * \varepsilon \tag{15}$$

For systems using ensemble binning, the ensemble penetration rate is calculated using (11) and the ensemble bin error rate is calculated as

$$1-\varepsilon_{ensemble} = \prod_{i=1}^{T}(1-\varepsilon_i) \tag{16}$$

where the ε_i are the bin error rates for the binning on each measure.

2.3 "Geninue", "Impostor", and "Inter-template" Distance Distributions

The function of the pattern matching module in Figure 1 is to send to the decision subsystem a positive, scalar measure, D, for every comparison of a sample to a template. We can presume, without loss of generality, that D increases with increasing difference between sample and template. We will loosely call this measure a "distance", recognizing that it will technically be such only if resulting from a vector comparison in a

metric space. The general biometric system does not require that sample and template features compose such a space[3].

Regardless of the mathematical basis for the comparison, from a series of comparisons of samples to truly matching templates, we can construct a histogram which approximates the "genuine" distance probability distribution function [22]. We will call this distribution, $\Psi_G(D)$, as shown in Figure 2. It is both device and measure dependent. This "genuine" distribution is a measure of the repeatability of the biometric pattern. Repeatability is negatively impacted by any factor causing changes in the measurement. Such factors generally accumulate over time, so the "genuine" distribution appears to drift in the direction of increasing distance with the passage of time. This phenomenon is called "template aging", although changes in the sample, not the stored template, are responsible for this decrease in repeatability.

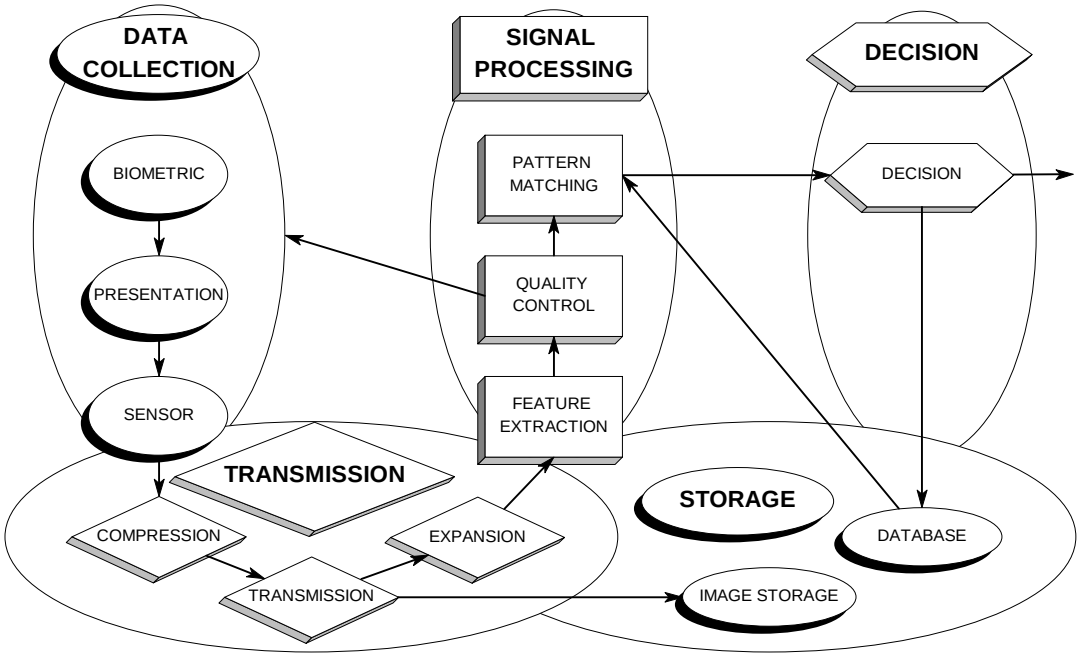


FIGURE 1: THE GENERAL BIOMETRIC SYSTEM

[3] Minutiae-based fingerprint systems are an example of biometric system with sample and template features not composing a metric space. In general, fingerprint samples and templates will have unequal numbers of features, distances are not symmetric, and the triangular inequality does not hold.

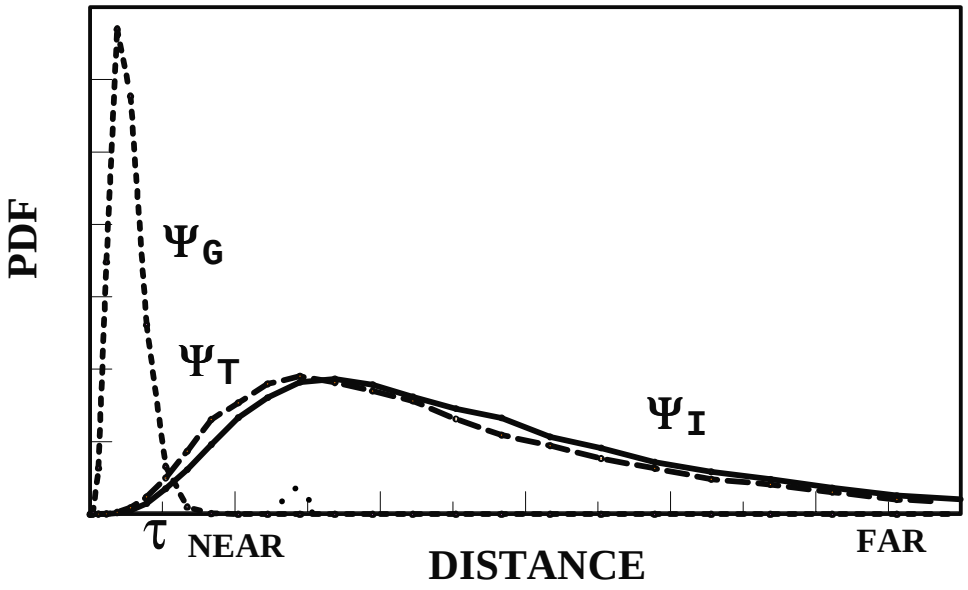


FIGURE 2: DISTANCE DISTRIBUTION FUNCTIONS

Similarly, from a series of comparisons of samples to different user's, or non-self, templates, we can construct a histogram which approximates the "impostor" distance probability distribution function, $\Psi_I(D)$. There are several ways of doing this. The "impostor" histogram can be constructed by comparing each sample to a single non-self template [12], by comparing every sample to every non-like template [5,13-17], or through "re-sampling"[18,23], drawing samples and templates from a pool at random with replacement. Some researchers [17] have suggested the use of a "background" database of templates for which there is no matching sample[4]. This allows the "impostor" comparisons to be from a greater pool of samples than the "genuine" comparisons.

Consequently, for all approaches except that of comparing each sample to a single non-self template [12], the resulting histogram for the "impostor" distribution is much smoother than the histogram used to construct the "genuine" distribution, even if the same number of independent data points were used to construct both. However, the number of independent comparisons ("degrees of freedom") resulting from each method, needed for the development of confidence intervals, cannot exceed the number of independent samples.

Ideally, the genuine and impostor distributions will be disjoint (non-overlapping), allowing us to discriminate completely between "genuine" and "impostor" comparisons using a distance threshold. Of course, this is never the case in practice; one side of the problem being large distances between samples and truly matching templates caused by changes in the underlying biometric measure, in its presentation to the sensor, or in the sensor itself. We have noticed, in practice, that $\Psi_G(D)$ is usually bimodal, with the second mode coincident with the primary mode of $\Psi_I(D)$. This means that changes in the biometric measure, or its presentation, have caused an individual to appear clearly as an impostor. We hypothesize that in the general system, the "genuine" distance does not increase smoothly with changes in the biometric pattern, but undergoes rapid increase with changes past a particular threshold. In any case, the general biometric system shows significant overlap in the tails of the "genuine" and "impostor" distributions.

There is actually a third distribution, the "inter-template" distribution, $\Psi_T(D)$, which expresses the distinctiveness between the templates. In practice, only the templates and the (presumed) genuine comparison distances may be available to the researcher. The actual samples may be discarded by the system. In the case of a system that creates templates from but a single sample, templates are samples. In this case, the "inter-template" distribution is identical to the "impostor" distribution.

The general biometric system might use multiple samples taken at a single "enrollment" session to create the template, or may update the template from a moving,

---

[4] In "large-scale" biometric systems with N exceeding ten million, estimation the false match rate from $\Psi_I(D)$ becomes much more critical than estimation of the false non-match rate from $\Psi_G(D)$. Further, single comparison false non-match rates of even 10% might give satisfactory system performance, while single comparison false match rates of 10^{-6} might be required. In this context, it is not possible to dismiss the need for a "background database" and the "independent degrees of statistical freedom" which ensue, provided that all templates and samples are collected from a similar population in a similar environment.

weighted average of samples presented over time. Simulation models have shown us that, in these cases, the "inter-template" distribution is closer to the origin than the "impostor" distribution, and consequently, does not make a good proxy in calculating the relationship between "false match" and "false non-match" rates as a function of decision threshold.

With the proper assumptions, we can construct the "impostor" distribution from the higher-dimensional convolution of the "genuine" and "inter-template" distributions. If we can assume that the "genuine" scalar distance measures result from an isotropic distribution of samples around the true templates, and that such distributions are "stable", meaning that the distribution resulting from the set of single sample-to-template distances is the same as the distribution of each sample about its own template, then we can reconstruct the sample-to-template distance distribution from the genuine and inter-template distributions [24]. The reconstruction algorithm must account for the template creation policy. This is an area of current emphasis in our research and classifies as a "hard" problem.

2.4 The Single Comparison False Match Rate

A single comparison false match occurs when a sample is incorrectly matched to a template in the database by the decision subsystem because the distance measure between the two is less than a threshold, τ, established by the decision policy. The single comparison false match rate, FMR(τ)can be computed from the integral of the "impostor" distribution function, $\Psi_I(D)$, between zero and the threshold, as

$$\mathrm{FMR}(\tau) = \int_0^{\tau} \Psi_I(D)\mathrm{d}D \tag{17}$$

which increases with increasing decision threshold. Although in practice τ might be user dependent, our analysis will consider τ to be at a single, fixed value for all users. The single comparison false match rate can be seen in Figure 2 as the area under Ψ_I between the origin and τ.

2.5. The Single Comparison False Non-Match Rate

A single comparison false non-match occurs when a sample is incorrectly not matched to a truly matching template by the decision subsystem because the distance between the two is greater than the fixed threshold. The single comparison false non-match rate, FNMR(τ), can be given as

$$\mathrm{FNMR}(\tau) = \int_{\tau}^{\infty} \Psi_G(D)\mathrm{d}D = 1 - \int_0^{\tau} \Psi_G(D)\mathrm{d}D \tag{18}$$

where $\Psi_G(D)$ is the genuine probability distribution function. FNMR(τ) decreases with increasing decision threshold. The single comparison false non-match rate can be seen in Figure 2 as the area under Ψ_G to the right of τ. It is clear from equations (17) and (18) that false match and false non-match rate are competing factors based on the threshold and can be set based on comparative risks false match and false non-match system errors.

2.6 Hardware Comparison Rate

The "one-to-one", or "cold match", comparison rate, C, is the number of comparisons per second, of a single sample to a single database template, that can be made by the hardware. It is a function of the hardware processing speed, the template size, and the efficiency of the matching algorithm. System availability must be considered when predicting the number of comparisons that can be made over longer time periods, such as a day or a month.

The architecture for large-scale (large N) biometric systems is modular in the sense that processing speed can be designed to meet seemingly any requirement, although there are no doubt limits of scale as speed requirements get too great. In general, a single comparison may take as many as a few million operations. Measurement and prediction of system processing speed from component architecture or from direct measurement will not be considered further in this paper

3.0 SYSTEM PERFORMANCE EQUATIONS

We are now in a position to write some first-principal equations reflecting the dependence of system performance on the parameters explained in the preceding section. By "system performance", we mean the timely and correct matching and non-matching of customers to identities in a database of N template ensembles, based on a system decision policy utilizing M samples from each customer. In the development of the equations, we will assume that one sample of each independent measure is submitted, such that M=T. Departures from this assumption will be handled in the applications section of this paper. The possibilities for system decision policies are limited only by the imaginations of the system developers. We will develop equations capable of modeling the most common approaches. Owing to both complication and lack of data, we will ignore any and all correlations between errors, expressing where we can the impact these assumptions have. Our goal will be to give system performance estimates and bounds, based on these simplifying, but admittedly inexact, assumptions.

In a multi-measure system with large N, search speed becomes an important issue. The usual approach is to conduct an initial search a subset, *m*, of the collected samples, M, where $m \leq M$ and M=T. This limited initial search will rule out most of the U enrolled users as potential matches with *m*, not M, comparisons for each, thus greatly increasing search efficiency.

Let's assume a system decision policy that requires, for a system "match" decision, Q matches of the M samples to a single enrolled ensemble of T templates. To do this, we will conduct an initial search of the relevant partitions of the database against *m* of the M samples. These *m* samples are searched sequentially through the entire database. In other words, each of the initial *m* searches will result in P_i*N comparisons and a total of $\sum_{i=1}^{m} P_i * N$ comparisons will be made over the *m* searches. Any matches found can be verified by comparisons of the remaining of the M samples against the remaining of the T templates from the same template ensemble. In other words, if the first of the *m* initial searches against the database produce no match, yet the second results in a match identifying a candidate template ensemble, the remaining M-2 input samples will be compared to the T templates in the identified ensemble. If this results in Q-1 or more matches, a system match is declared. Accordingly, if all *m* samples in the

initial search falsely non-match, or more than T-Q false non-matches occur against a correctly matched enrolled ensemble, a system "non-match" is falsely declared.

We will allow each of the *m* samples in the initial search to have independent error rates, ε_i, $FMR_i(\tau)$ and $FNMR_i(\tau)$. This reflects the differences in the underlying "genuine" and "impostor" distributions for each sample and allows for sample-dependent, but not user-dependent, thresholds. For purposes of mathematical tractability, however, the samples in the remaining comparisons (which may include samples from among the *m*) will be assumed to have uniform error rates, ε_U, $FMR_U(\tau)$ and $FNMR_U(\tau)$.

3.1 System False Non-Match Rate

When comparing a single input sample to a single stored template, for false non-match not to occur, there must be: 1) no binning error; 2) no single comparison false non-match. Assuming these errors to be independent, the probability of a correct match of a single sample with a truly matching template can be written

$$\Pr\{\text{correct match for i}^{\text{th}}\text{ sample}\} = 1 - FNM_i = (1-\varepsilon_i)(1-FNMR_i) \qquad (19)$$

where FNM_i is the probability that the i^{th} sample will not be properly matched for any reason and the explicit dependence of $FNMR_i$ on threshold ,τ, has been dropped for notational simplicity. Rewriting equation (19),

$$FNM_i = \varepsilon_i + FNMR_i - \varepsilon_i * FNMR_i \qquad (20)$$

Some simple systems make a series of sample-to-template comparisons without using any ensemble concepts. The decision policy for such systems may only require a single match on one of these comparisons for a system match to be declared. A system false non-match occurs only when all *m* comparisons result in a false non-match. Assuming independence of false non-matches,

$$FNM_{sys} = \prod_{i=1}^{m} FNM_i \qquad (21)$$

The development of (31) assumes the comparisons to be independent. From elementary probability theory, we know that

$$\Pr\{A \cap B \cap C \cdots\} = \Pr\{A\} * \Pr\{B \mid A\} * \Pr\{C \mid AB\} \cdots \qquad (22)$$

where Pr{B|A} indicates the conditional probability of event B occurring given that A has occurred. If Pr{A|B}=Pr{A}, Pr{C|AB}=Pr{C}, we say that events A, B, and C are independent. In practice, we find this not to be the case. In operational data, we have observed that a single comparison false non-match increases slightly the probability that a subsequent biometric sample of the same characteristic from the same customer will also be falsely non-matched. We can reasonably expect the probability of a false non-match to approach to one as the number of previous false non-matches from the same session by the same customer increases. Further, we expect this hypothesis to hold for any reasonable threshold. Consequently, we expect that equation (21) will underestimate by some unknown amount the true system false non-match rate.

For systems using ensembles of multiple measures, the i^{th} of *m* searches against an entire ensemble to not result in a false non-match requires that: 1) the initial

comparison of sample to template not result in a false non-match; and 2) Q-1 or more of the remaining patterns in the ensemble be correctly matched. Therefore, the probability of a correct identification being declared on the i^{th} of the *m* sample comparisons is

$$\Pr\{\text{correct identification declared on } i^{th} \text{ sample}\} =$$

$$(1-\text{FNM}_i)\sum_{j=Q-1}^{T-i}\binom{T-i}{j}(1-\text{FNM}_U)^j(\text{FNM}_U)^{T-i-j} \tag{23}$$

The complement, that the correct identification is not declared on the i^{th} sample, can be given as

$$\Pr\{\text{correct identification not declared on } i^{th} \text{ sample}\} =$$

$$1-(1-\text{FNM}_i)\sum_{j=Q-1}^{T-i}\binom{T-i}{j}(1-\text{FNM}_U)^j(\text{FNM}_U)^{T-i-j} \tag{24}$$

The concept of expressed by equation (21) still applies, but with the more complicated definition of FNM_i given by (24). Assuming that the *m* searches are independent, the probability that a system false non-match occurs is, therefore

$$\text{FNM}_{sys}=\prod_{i=1}^{m}\left[1-(1-\text{FNM}_i)\sum_{j=Q-1}^{T-i}\binom{T-i}{j}(1-\text{FNM}_U)^j(\text{FNM}_U)^{T-i-j}\right] \tag{25}$$

where FNMR_{SYS} is the system false non-match rate.

For systems that use ensemble binning, the probability of a bin error is the same for all samples, namely $\varepsilon_{ensemble}$. If $\varepsilon_{ensemble}$ replaces ε_i in (19), then probabilities of correct match calculated by (19) will not be independent for each sample and cannot be used in developing the system false non-match rate equation. When using ensemble binning, the bin error is not independent over the M comparisons, as each comparison looks in the same database partition. We can modify the above development by removing consideration of the binning error from (19), writing

$$\Pr\{\text{correct match for } i^{th} \text{ sample}\} = 1-\text{FNM}_i = 1-\text{FNMR}_i \tag{26}$$

so

$$\text{FNM}_i = \text{FNMR}_i \tag{27}$$

Equations (21), (23), and (24) continue to hold using (27).

For the system to return a proper identification, we require: 1) no binning error for the entire ensemble; 2) no failure of all initial *m* searches to identify the ensemble.

For a correct identification to be made, we

$$1-\text{FNM}_{sys}=\left[1-\varepsilon_{ensemble}\right]\left[1-\prod_{i=1}^{m}\left[1-(1-\text{FNM}_i)\sum_{j=Q-1}^{T-i}\binom{T-i}{j}(1-\text{FNM}_U)^j(\text{FNM}_U)^{T-i-j}\right]\right] \tag{28}$$

Equation (28) can be rewritten as

$$\text{FNM}_{sys}=\varepsilon_{ensemble}+\left[1-\varepsilon_{ensemble}\right]\prod_{i=1}^{m}\left[1-(1-\text{FNM}_i)\sum_{j=Q-1}^{T-i}\binom{T-i}{j}(1-\text{FNM}_U)^j(\text{FNM}_U)^{T-i-j}\right] \tag{29}$$

3.3 System False Match Rate

In simple systems not using multiple independent measures arranged as ensembles, a match will be declared if any of the *m* sample-to-template comparisons over the entire database result in a match. Consequently, no system false match requires no single comparison false match over the entire database.

$$1-\mathrm{FMR}_{\mathrm{sys}}=\prod_{i=1}^{m}\left[1-\mathrm{FMR}_{i}\right]^{N*P_i} \tag{30}$$

Rewriting,

$$\mathrm{FMR}_{\mathrm{sys}}=1-\prod_{i=1}^{m}\left[1-\mathrm{FMR}_{i}\right]^{N*P_i} \tag{31}$$

For large *m**N*P, the system false match rate approaches 1 even for very small single comparison false match rates, FMR_i. Consequently, such a system design cannot be used for large-scale "identification" systems.

In systems using ensembles of multiple measures, a system false match occurs if Q or more samples are falsely matched against the enrolled ensemble of a single individual. One general approach is to search $m \le M = T$ samples against a partition of the database. If any matches are found, the remaining samples are compared to the associated templates in the matched enrolled ensembles. If Q-1 additional false matches are found in any single enrolled ensemble, a match will be falsely declared by the system. The probability that the i^{th} of the initial *m* searches will result in a false match against a single non-matching ensemble can be given by

$$\Pr\{\text{false match on the } i^{th} \text{ sample}\}=\mathrm{FMR}_i * \sum_{j=Q-1}^{T-i}\binom{T-i}{j}\mathrm{FMR}_U{}^{j}(1-\mathrm{FMR}_U)^{T-i-j} \tag{32}$$

Again the explicit dependence of the false match rate on threshold, τ, has been dropped for notational simplicity and the false match rates, FMR_i, within the summation sign are considered uniform.

Numerical computation of (32) from the single comparison false match rates for each sample, FMR_i, is straight forward, as the ensemble size, M, will always be small, perhaps reaching ten in the case of a ten-print fingerprint identification system.

The search of the i^{th} of the initial *m* patterns against the entire database will not result in a false match only if none of the $N*P_i$ searches end in a false match. Therefore, the probability that the i^{th} of the *m* initial searches against the relevant partition of the database will not end in a false match is

$$\Pr\left\{\text{incorrect identification not made on } i^{th} \text{ sample}\right\}=$$

$$\left[1-\mathrm{FMR}_i * \sum_{j=Q-1}^{T-i}\binom{T-i}{j}\mathrm{FMR}_U{}^{j}(1-\mathrm{FMR}_U)^{T-i-j}\right]^{N*P_i} \tag{33}$$

For a search of a sample ensemble against the database to not end in a false match requires that none of the *m* initial comparisons falsely match. Therefore, the system false match rate can be given as

$$1-\mathrm{FMR}_{sys}=\prod_{i=1}^{m}\left[1-\mathrm{FMR}_i*\sum_{j=Q-1}^{T-i}\binom{T-i}{j}\mathrm{FMR}_U{}^{j}(1-\mathrm{FMR}_U)^{T-i-j}\right]^{N*P_i} \tag{34}$$

which can be rewritten as

$$\mathrm{FMR}_{sys}=1-\prod_{i=1}^{m}\left[1-\mathrm{FMR}_i*\sum_{j=Q-1}^{T-i}\binom{T-i}{j}\mathrm{FMR}_U{}^{j}(1-\mathrm{FMR}_U)^{T-i-j}\right]^{N*P_i} \tag{35}$$

Equation (35) holds regardless of the type of binning chosen. Note that the system false match rate decreases with decreasing system penetration rate. If ensemble binning is used, the penetration rates, P_i, are replaced with the single penetration rate, $P_{ensemble}$. Unlike the simple design used for development of equation (31), this ensemble-based design allows for reasonable system false match rates even for systems with large N.

3.4 System Throughput

The final set of system equations is an approximation for the system throughput rate, S, which depends upon:1) the hardware "one-to-one" comparison rate, C; 2) the number of input samples compared to the database, *m*; 3) the number of samples in the database, N; and 4) the penetration rate, either computed on each sample, P_i; or over the ensemble, $P_{ensemble}$.

We are assuming that the system throughput rate is entirely limited by computational speed, not data collection time. This will be a fair assumption only for systems with large N. For systems with small N, throughput times will be limited by data collection speed, and other human factors.

Under the assumption that no matches will be found, the computational throughput rate, S, in customers per unit time, can be written as

$$S=\frac{C}{\sum_{i=1}^{m}P_i*N} \tag{36}$$

where C is the hardware "one-to-one" computational rate. In the case where ensemble binning is used, (36) becomes

$$S=\frac{C}{m*P_{ensemble}*N} \tag{37}$$

Violation of our assumption regarding binning independence increases penetration rate and decreases throughput. Any matches found (false or correct) require additional comparisons over the remaining portion of the ensemble, further decreasing throughput, so equations (36) and (37) are an optimistic upper bound.

This throughput rate must match the customer input on a time scale driven by operational requirements. Because of the various time units used, care must be taken in dimensional balancing when applying (36) or (37).

It is generally true that hardware system costs increase with processing speed, C. Minimizing costs against a fixed customer throughput requirement pushes the system designer to decrease P_{sys}, through additional binning or filtering, thereby increasing false non-match rate by equation (7) , (25) or (29) , and decreasing false match rate by equation . We are presented with the somewhat surprising result that, through the penetration rate, system error rates depend upon system processing speed.

4.0 EXAMPLES

In this section we will apply the above equations to several types of biometric systems, specifically "one-to-one" "verification" systems, with and without a "three-strikes you're out" policy, and "one-to-many" and "M-to-many" "identification" systems .

4.1 "One-to-One" Systems

Consider a system in which a single sample is given and compared to a single enrolled template, perhaps contained on an identification card or associated with an enrolled user in a centralized database. The number of stored templates for each user is T=1. If the templates are stored on a card, N=T=1 and P_{sys}=1. If the templates are in a centralized database, then N=UT and P_{sys}=1/U. In either case, $N*P_{sys}=T=1$. There is no binning, so the bin error rate is zero and the penetration rate is unity.

By equation (35), FNM=FNMR. Using equation (21) with *m*=1, we get FNM_{SYS} = FNMR, which is as expected.

Equation (25) could also be used, with N=M=T=Q=1. We get

$$FNM_{SYS} = 1-(1-FNMR)\sum_{j=0}^{0}\binom{0}{j}(1-FNMR)^{0}(FNMR)^{0} = FNMR$$

Application of equation (29) also produces the same result.

The system false match rate is most easily calcluated with equation (31). We get FNM_{SYS} = FNMR, which is as expected.

We could also calculate the system false match rate using (35). We have

$$FMR_{sys} = 1-\prod_{i=1}^{1}\left[1-FMR_i * \binom{0}{0} FMR_U{}^{0}(1-FMR_U)^{0}\right]^{1} = FMR_i$$

as expected.

Applying equation (37) for throughput rate,

$$S = \frac{C}{m * P_{ensemble} * N} = C$$

and we see that the thoughput rate is exactly the single-sample-to-single-template hardware comparison rate. This can be assumed to be so much faster than the time required for sample input that the throughput rate will be limited by human factors, not by hardware considerations.

4.2 "One-to-One" Systems with a "Three-Strikes" Decision Policy

Consider a system in which a single sample is given and compared to a single enrolled template, but the customer is given three tries to be identified. Any single match over the three tries results in a system "match" decision The single comparison error rates are assumed invariant over the match attempts. There is no binning, so the bin error rate is zero and $N*P_{sys}=T=1$, as in the previous example. By equation (20), FNM=FNMR. Using equation (21) with m=3, we get

$$FNM_{sys} = \prod_{i=1}^{m} FNMR_i = FNMR^3$$

Equation (25) could also be applied, taking Q=T=1 and M=m=3. Because Q= T=1, the requirement for Q-1 matches against remaining templates has a probability of 1. Under these conditions, equation (25) yields

$$FNM_{sys} = \prod_{i=1}^{m} \left[1-(1-FNM_i)(1)\right] = FNMR^3$$

As previously noted, our computation above will underestimate the true false non-match rate.

The false match rate is computed using equation (31). Again, the probability of Q-1 matches against remaining templates is 1.

$$FMR_{sys} = 1-\prod_{i=1}^{3}\left[1-FMR_i * 1\right]^1 = 1-\left[1-FMR\right]^3 = 3*FMR-O\left(FMR^2\right)$$

where $O(FMR^2)$ indicates terms on the order of the square of the false match rate. The violation of the assumption of independence will cause the product in the above computation to be too large. Accordingly, the this calculation will overestimate the true false match rate.

These results for system false match and false non-match rate are identical to previously published results for the "three strikes" case [1].

By equation (37),

$$S = \frac{C}{m * P_{ensemble} * N} = \frac{C}{3}$$

The throughput rate is one-third the hardware comparison rate. Again, this is insignificant compared to the data collection time. A trick that is usually employed to decrease the collection time is to collect and test the samples one at a time, collecting further samples only if a match is not determined.

4.3 "One-to-Several" Verification Systems

Now we will consider a system using only one biometric measure, but allowing several input samples and stored templates of varying presentations of that measure for each enrolled customer. If any input sample matches any of the enrolled templates, a system "match" results. As the the previous examples, $N*P_{sys}=T$, but here, $T>1$. Similar to the development of equation (21), a false non-match occurs only if all comparisons of

the *m* input samples to the T stored templates result in a false match. If the sample-to-template comparisons were independent, we could write

$$FNM_{sys} = \prod_{i=1}^{m} FNM_i^T = FNM^{m*T}$$

Similarly, equation (31) could be rewritten as

$$FMR_{sys} = 1 - \prod_{i=1}^{m} [1 - FMR_i]^{N*P_i} = 1 - \prod_{i=1}^{m} [1 - FMR_i]^T = m*T*FMR - O(FMR^2) \approx m*T*FMR$$

where $O(FMR^2)$ indicates terms on the order of the square of the single comparison false match rate and the approximation is valid to the extent that this rate is small.

These equations indicate that the system false non-match rate would go to 0 and the system false match rate to 1, as the number of total comparisons, *m*T*, increases. As previously noted, we have observed that a single comparison false non-match increases the probability of subsequent non-matches by the same customer within the same session. Additionally, we have operationally observed that the absence of a false match by a customer against a template decreases the probability of subsequent false match by that customer against the same template. Consequently, the development in this section over-estimates the probability of a system false match and under-estimates the probability for a system false non-match.

4.4 "One-to-Many" Single Comparison Systems

Now we will consider a system in which a single sample is given and compared to a partitioned database of N individuals, enrolled with one template each. The system "match/non-match" decision is made on the basis of the single sample. In this case, N=large, T=M=m=Q=1. This time there is individual sample binning, so the bin error rate is non-zero and the penetration rate is less than one.

Using data from the recent international automatic fingerprint identification system (AFIS) benchmark test [2], we will take values of performance equation parameters that are consistent with large-scale fingerprint systems. Let's allow the penetration rate from endogenous binning be P = 0.5 and apply gender-based filtering. Further, we will take the values $\varepsilon_{BIN} = 0.01$, $FMR=10^{-5}$, and FNMR=0.05 as obtainable by a general large-scale system.

We will first calculate the gender-based filter factor by applying equation (7). We will assume the population to be evenly divided between male and female and guess that no more than 2% of the population will be of unknown gender and that these unknowns will be equally male and female. The gender filter factor becomes by (7),

$$P_{gender} = p_K + \sum_{i=1}^{K-1} (p_i + p_K) p_i = 0.02 + 0.51*0.49 + 0.51*0.49 \approx 0.51$$

Using equation (10), the total system penetration rate becomes

$$P_1 = \prod_{j=1}^{2} P_{1,j} = 0.5*0.51 = 0.26$$

By equation ,

$$FNM_1 = FNMR_1 + \varepsilon - FNMR_1 * \varepsilon \approx 0.06$$

Using equation (29) to calculate the false non-match rate for the system as

$$FNM_{sys} = \prod_{i=1}^{1}\left[1-(1-FNM_i)\binom{0}{0}(1-FNM_U)^0(FNM_U)^0\right] = FNM_1 = 0.06$$

We can see that the false non-match rate is not directly dependent on N. However, as N increases, the system designer will be under increasing pressure to keep down the required computational rate by trading decreases in P_{SYS} for increases in ε_{BIN}, and thereby increasing the false non-match rate.

Now we consider the false match rate as given by equation (31), which gives the probability that a sample will have one **or more** false matches over the P_{sys}* N comparisons made. The expected number of system false matches, $E\{FM_{sys}\}$, for a single sample over P_{sys} N comparisons is

$$E\{FM_{sys}\} = P_{sys} * N * FMR_{sys} \tag{38}$$

Equation (38) comes directly from the expected value of a binomial distribution with parameters n= P_{sys}* N and p= $Pr\{FM_{sys}\}$. Equation (38) is interesting in its predictive value in estimating the formation of "candidate lists", the return of several false matches with each correct match as $E\{FM_{sys}\}$ approaches and exceeds 1. Equation (38) tells us that candidate lists can be avoided only if $P_{sys} * N << 1/Pr\{FM_{sys}\}$. This sets a natural limit for database size for the general biometric system as a function of the system penetration, single comparison false match rate, and system decision policy, if human intervention in the adjudication of candidate lists is to be avoided.

In our current example, with P_{sys}=0.26 and FMR=10^{-5}, this implies that N<< 400,000, if false matches are to be avoided. Let's take N=20,000 as the size of our system.

Computing the false match rate using (35) with N=20,000,

$$FMR_{sys} = 1-\prod_{i=1}^{1}\left[1-FMR_i * \binom{0}{0} FMR_U{}^0(1-FMR_U)^0\right]^{N*P_i} = 1-[1-FMR]^{N*P_{sys}} \approx P_{sys} * N * 10^{-5} = 0.05$$

the approximation arising from the binomial expansion of the expression and being valid to the extent that P_{sys}**N***FMR<<1. This indicates that in such a system, approximately 5% of the customers would be falsely matched to one or more templates in the enrolled database.

Now, let's assume a customer input rate of N per year, or about 160 customers per working day, based on 250 working days per year. By equation (37), if the system throughput over an 8 hour period is to equal the customer input rate per day, then

$$S = \frac{C}{P_{ensemble} * N} = \frac{C}{0.26 * 2\times 10^4} = \frac{160}{\text{day}}$$

The required hardware comparison rate can be calculated as

$$C = \frac{160 * 0.26 * 2\times 10^4}{8 \text{ hour day}} \approx \frac{9\times 10^5}{3\times 10^4 \text{ sec}} = 30 \text{ comparisons / sec}$$

4.5 "M-to-N" System Example

Now we will consider a specific instantiation of the M-to-N system in which four independent measures are used[5]. Both input samples and stored templates consist of a single ensemble of 4 measures. So, for this system M=T=4 and N=4U. We will consider a large-scale system where $U=10^8$ so that $N= 4x10^8$. An initial search will be made sequentially over two samples, so m=2 and a total of $m*P_{sys}*N$ comparisons will be made. A "match" decision is made only if at least three input samples match the ensemble of a single enrolled individual, so Q=3. We will use the same values for single comparison error rates and penetration rate as in the above example. Gender-based filtering will again be used. It will be assumed that uniform bin error rates, single sample penetration rates and single comparison error rates apply to all measures in the ensemble.

Because we are using an ensemble of multiple, independent measures, we will have the choice of either binning on individual samples or the entire ensemble. It is interesting to compare the performance differences in the two approaches.

Regardless of approach used, we will partition the database by the four independent measures, placing all measures of each type into different, non-communicating bins. For example, in a multiple fingerprint system, right thumb prints will be placed in one partition, left thumb prints in another. This is a filtering operation performed by the system operator at the time of data collection. Consequently, the inevitable errors in this procedure will not be considered in this analysis. Equation (3) applies and the penetration rate, P_f, owing to this filtering method is

$$P_f = \frac{1}{K} = \frac{1}{4} = 0.25$$

Assume that we will use gender-based filtering and bin individually on each measure, using the penetration rates of the previous example. Then, by equation (10), the individual sample penetration rate would be

$$P_i = \sum_{j=1}^{3} P_{i,j} = 0.25 * 0.51 * 0.5 = 0.06$$

If ensemble binning were used and if all the measures can be considered independent, by equation (11), the ensemble penetration rate would be

$$P_{ensemble} = 0.25 * 0.51 * 0.5^4 \approx 0.008$$

If the partitionings of any of the measures are correlated, this value will be higher. We can see that ensemble binning decreases the penetration rate by nearly an order of magnitude over binning only on the individual samples. This translates into nearly an order of magnitude decrease in hardware computation rate for a fixed throughput requirement.

---

[5] The Republic of the Philippines Social Security System identification project will use an ensemble of four fingerprints (both thumbs and both forefingers), with an initial search on the forefingers only. Confirmation of any matches will be against the remainder of the ensemble.

A less obvious difference between systems using ensemble binning and systems using individual sample binning is in the system error rates.

Binning on individual samples, we can use (20) to write

$$\mathrm{FNM} = \mathrm{FNMR} + \varepsilon - \mathrm{FNMR} * \varepsilon \approx 0.06$$

Using equation (29),

$$\mathrm{FNM}_{\mathrm{sys}} = \prod_{i=1}^{2}\left[1-(1-\mathrm{FNM})\sum_{j=2}^{4-i}\binom{4-i}{j}(1-\mathrm{FNM})^{j}(\mathrm{FNM})^{4-i-j}\right] =$$

$$\left[1-(1-\mathrm{FNM})\left(3(1-\mathrm{FNM})^{2}\mathrm{FNM}+1(1-\mathrm{FNM})^{3}\right)\right]\left[1-(1-\mathrm{FNM})^{3}\right] \approx 0.01$$

With ensemble binning, we calculate the ensemble binning error from equation (16) as

$$\varepsilon_{\mathrm{ensemble}} = 1-\prod_{i=1}^{4}(1-\varepsilon) \approx 0.04$$

We use (27) to write

$$\mathrm{FNM} = \mathrm{FNMR} = 0.05$$

Using equation (28) to calculate the system false non-match rate, we find

$$\mathrm{FNM}_{\mathrm{sys}} = \varepsilon_{\mathrm{ensemble}} + \left[1-\varepsilon_{\mathrm{ensemble}}\right]\prod_{i=1}^{2}\left[1-(1-\mathrm{FNM})\sum_{j=2}^{4-i}\binom{4-i}{j}(1-\mathrm{FNM})^{j}(\mathrm{FNM})^{4-i-j}\right] =$$

$$\varepsilon_{\mathrm{ensemble}} - \left(1-\varepsilon_{\mathrm{ensemble}}\right)\left[1-(1-\mathrm{FNM})\left(3(1-\mathrm{FNM})^{2}\mathrm{FNM}+1(1-\mathrm{FNM})^{3}\right)\right]\left[1-(1-\mathrm{FNM})^{3}\right] \approx 0.05$$

Thus, in this example, the system false non-match rate is five times higher using ensemble binning than using binning on individual samples.

Considering the false match rate, equation (35) applies for both binning methods. The difference in its application is in the penetration rate used. For either method, (35) becomes

$$\mathrm{FMR}_{\mathrm{sys}} = 1-\prod_{i=1}^{3}\left[1-\mathrm{FMR}*\sum_{j=2}^{4-i}\binom{4-i}{j}\mathrm{FMR}^{j}(1-\mathrm{FMR})^{2-i-j}\right]^{N*P_i}$$

In this example we assume the penetration rate to be the same for all samples, even if individual sample binning is used. Therefore, the above equation can be rewritten as

$$\mathrm{FMR}_{\mathrm{sys}} = 1-\left[\prod_{i=1}^{2}1-\mathrm{FMR}*\sum_{j=2}^{4-i}\binom{4-i}{j}\mathrm{FMR}^{j}(1-\mathrm{FMR})^{2-i-j}\right]^{N*P} \approx 1-\left[1-4*\mathrm{FNM}^{3}\right]^{N*P}$$

If individual sample binning is used, P=0.06 and the system false match rate is approximately $1\text{x}10^{-7}$ or one false match in every 10^{7} customer transactions. If ensemble

binning is used, P=0.008 and the system false match rate is approximately $1x10^{-8}$. With the ensemble binning method, the system has a smaller penetration rate, allowing fewer comparisons and fewer opportunities for a false match. We emphasize again that the above development assumed, without proof, statistical independence of all errors.

After the system is fully operational, with 10^8 users enrolled, we will assume that renewals and re-issuances occur at a rate of about 1/5 U per year. In a five year period, we would expect about 10 false match errors to occur with a system using individual sample binning and about 1 to occur with ensemble binning.

Based on the above, the input rate will be about $4x10^5$ customers per week, based on 50 working weeks per year. By equations (36) and (37), the required hardware comparison rate for both individual sample and ensemble binning can be calculated as

$$S = \frac{C}{m * P * N} = \frac{4 \times 10^5}{\text{week}}$$

Assuming a system availability of 20 hours per day, 7 days a week for the same 50 weeks per year, the required hardware computational rate becomes

$$C = \frac{4 \times 10^5 * 2 * 4 \times 10^8 * P}{5 \times 10^5 \,\text{sec}} \approx 6 \times 10^8 * P \text{ computations/sec}$$

For individual sample binning with P=0.06, $C=4x10^6$ computations per second. With ensemble binning, P=0.008, and $C=5x10^5$ computations per second. Large-scale AFIS vendors are currently designing hardware systems with target processing rates on the order of a few hundred thousand comparisons per second.

5.0 CONCLUSIONS

In this paper, we derived equations for false match and false non-match error rate prediction for the general M-to-N biometric identification system, under the simplifying, but limiting, assumption of statistical independence of all errors. For systems with large N, error rates were shown to be linked to the hardware processing speed through the system penetration rate and the throughput equation. These equations are somewhat limited in their ability to handle sample-dependent decision policies, and were shown to be consistent with previously published cases for "verification" and "identification" [1,2]. Applying parameters consistent with the Philippine Social Security System benchmark test results for automatic fingerprint identification system (AFIS) vendors [2], we established that biometric identification systems can be used in populations of 100 million people. Development of more generalized equations, accounting for error correlation and general sample-dependent thresholds, establishing confidence bounds, and substituting the inter-template for the impostor distribution under the template generating policy, remain for future study.

6.0 ACKNOWLEDGEMENTS

The author is greatly indebted to Drs. John M. Colombi and Joseph P. Campbell for their careful reading of the text and their many helpful suggestions, and to Dr. Larry O'Gorman for providing the inspiration to get these ideas written down.

7.0 REFERENCES

[1] J.L. Wayman, "A Scientific Approach to Evaluating Biometric Systems Using a Mathematical Methodology", Proc. CTST'97, pg. 477-492
[2] J.L. Wayman, "Benchmarking Large-Scale Biometric System: Issues and Feasibility", Proc. CTST Government'97, Sept. 1997
[3] D.C. Bright, "Examining the Reliability of a Hand Geometry Identity Verification Device for Use in Access Control", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1987
[4] M. Fuller, "Technological Enhancements for Personal Computers", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1992
[5] S.C. Geshan, "Signature Verification for Access Control", Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1991
[6] D. Helle, "Examination of Retinal Pattern Threshold Levels and Their Possible Effect on Computer Access Control Mechanisms", Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1985
[7] G. Poock, "Fingerprint Verification for Access Control", Naval Postgraduate School Report NPSOR-91-12, Monterey, CA, April 1991
[8] G. Poock, "Voice Verification for Access Control", Naval Postgraduate School Report NPSOR-91-01, Monterey, CA, October 1990
[9] H. Kuan, "Evaluation of a Biometric Keystroke Typing Dynamics Computer Security System", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1992
[10] L. Tirado, "Evaluation of Fingerprint Biometric Equipment", Master's Thesis, Naval Postgraduate School, Monterey, CA, March 1991
[11] Orkand Corporation, "Personal Identifier Report", California Department of Motor Vehicles, DMV 88-89, May 1990
[12] J.P. Holmes, et al, "A Performance Evaluation of Biometric Identification Devices", Sandia National Laboratories, SAND91-0276, June 1991.
[13] F. Bouchier, et al., "Laboratory Evaluation of the IriScan Prototype Biometric Identifier", Sandia National Laboratories, SAND96-1033, April 1996.
[14] P.J. Phillips, et al, "FERET (Face-Recognition Technology) Recognition Algorithm Development and Test Results", Army Research Laboratory, ARL-TR-995, October 1996
[15] P.J. Rauss, et al, "FERET (Face-Recognition Technology) Recognition Algorithms", *Proceedings of ATRWG Science and Technology Conference*, July 1996
[16] A.K. Jain, etal, "An Identity-Authentication System Using Fingerprints", *Proc. of the IEEE,* Vol. 85, No.9, Sept. 1997, pg 1365-1388.
[17] W. Shen, etal., "Evaluation of Automated Biometrics-Based Identification and Verification Systems", *Proc. of the IEEE,* Vol. 85, No.9, Sept. 1997, pg 1464-1478.
[18] J.P. Campbell, "Speaker Recognition : A Tutorial", *Proc. of the IEEE*, Vol. 85, No.9, Sept. 1997, pg 1437-1462.
[19] J.L. Wayman, "The Science of Biometric Technologies: Testing, Classifying, Evaluating", Proc. CTST'97, pg. 385-394
[20] J.L. Wayman, "A Generalized Biometric Identification System Model", Proc. of the IEEE Asilomar Conference on Signals, Systems, and Computers, Nov.,1997

[21] K. James and B. James, "NSA SAG Problem 97-25", UC Berkeley Dept. of Stats. Monograph, to be published
[22] D.W. Scott, Multivariate Density Estimation, (New York, Wiley, 1992)
[23] B. Efron and R.J.Tibshirani, An Introduction to the Bootstrap (Chapman and Hall, New York, 1993)
[24] P. Bickel, "SAG Problem 97-2-1", UC Berkeley Department of Statistics monograph, to be published

Memo on Non-Identically Distributed Bernoulli Model Problems for System Performance Prediction

Hani Doss
Department of Statistics
Ohio State University

(Editors Note: This and the following paper were in response to the following question to the Statistical Advisory Group:

Statement: Suppose N not necessarily fair, but independent coins are flipped. If the probability of heads were the same, then of course the probability of seeing exactly x heads, P{X=x}, is given by the binomial.

1. *If each of the N coins (say N is small, like 3 or 200 has a different, but known probability p_I of a head, can a simple or approximate equation be written for P{X}?*

2. *If we approximate P{X} as a binomial distribution with the common probability of heads p equal to the average of the p_I , is there a way to characterize the error in P{X} as a function of some measure of variability of the p over the N coins?*

3. *Under what conditions can the problem above be inverted, that is from numerous experiments involving N flips and observing the number of heads in each experiment, estimate the best fitting p to the model in 2), including the variability of the estimator?*

Comments:

1) Of course we know that 1) starts from the expression

$$P(X) = \sum_{(x_1+\ldots+x_N=X)} \prod_{i=1}^{N} p_i^{x_i} (1-p_i)^{1-x_i}$$

where x_i is 1 for a "head" and 0 for a "tail" on the i^{th} coin's toss.

2) We are aware of the general forms of the Central Limit Theorem for non-i.d. r.v.s, such as in S.S. Wilks (1962) Mathematical Statistics , John Wiley & Sons, Sec. 9.2, pp.256-259)

Background and General Statement of the Problem

From our discussions at the SAG meeting of July, 1998, we understand the problem to be as follows. We wish to determine whether an individual is who he claims to be or is an intruder. We have K tests. The probability that the k^{th} test will determine that the person is an intruder when in fact the person is not is p_k, k=1,2…,K. We form the total score

$$X = \sum_{k=1}^{K} I(\text{test } k \text{ determines individual is intruder})$$

and we claim that the individual is an intruder if X exceeds a certain threshold. We wish to

1. Determine the distribution of X, assuming that the p_k's are known.
2. Estimate the p_k's from a sample X_1, ….X_n. The estimation procedure should include both point estimates and confidence intervals.

The Distribution of X

We will find a simple, closed form expression for the distribution of X. Let $X = \sum_{k=1}^{K} Y_k$, where the Y_k's are independent Bernoulli random variables with success probability p_k. Consider the probability generating function

$$f(s) = E(s^X) = E\left(\prod_{k=1}^{K} s^{Y_k}\right) = \prod (1 - p_k + p_k s), \qquad (1)$$

where the third equality follows from the independence of the Y_k's.This is a polynomial in s, of degree K. The coefficient of s^j is P(X=j).

Fortunately, the calculation of the polynomial (1) can be done in about K^2 steps. (We form the product iteratively. At the j^{th} step, we need to do 2j multiplications.)

Estimation of the p_k's from a Sample X_1, ….X_n

Let

$$\hat{f}_n(s) = \frac{1}{n} \sum_{i=1}^{n} s^{X_i}$$

be the empirical probability generating function. By the law of large numbers, for each s, $\hat{f}_n(s) \to f(s)$, and it is easy to see that convergence is actually uniform in s over compact sets. Consider the generating function f(s). From (1), we see that this is a polynomial of degree K. All its roots are real. In fact, the roots are equal to

$$r_k = 1 - 1/p_k, \quad k = 1, \ldots, K. \qquad (2)$$

Equation (2) shows that if we can estimate the roots of f(s), then we can estimate the probabilities of p_k, k=1,…,K.

Suppose now that the roots of f(s) are distinct. In this case, the derivative of the polynomial at each root is positive, meaning that the polynomial actually crosses the s axis (i.e. it changes sign at the root). Since $\hat{f}(s)$ converges uniformly to f in small neighborhoods of the roots, for large n, $\hat{f}(s)$ has K real roots, and these roots converge to those of f.

Let $r^{(k)}$ and $r_n^{(k)}$ denote the k^{th} root of f(s) and $\hat{f}_n$, respectively. We will show that

$$n^{1/2}(r_n^{(k)} - r^{(k)}) \xrightarrow{d} N\left(0, \frac{f(2r^{(k)}) - f^2(r^{(k)})}{\left(f'(r^{(k)})\right)^2}\right) \qquad (3)$$

where f' is the derivative of f. Of course, since the probabilities are related to the roots via (2), we can get an asymptotic nomality result for the vector of probabilities through a very simple application of the delta method. Note that for each s,

$$n^{1/2}(\hat{f}_n(s) - f(s)) \xrightarrow{d} N(0, f(2s) - f^2(s)) \qquad (4)$$

by the central limit theorem. (It is even possible to obtain a functional convergence result, i.e. a result that gives convergence of $\hat{f}_n$ as a function of s.) We are estimating the solution of the equation f(s)=0. Now, near the root, f is nearly linear, with slope f'. Intuitively, the flatter f is near the root, the worse the estimation is going to be; hence the division by the derivative of f in (3). (This is very similar to the argument that gives an asymptotic variance of $[F(q)(1-F(q))]/(F'(q))^2$ for the estimate of the q^{th} quantile of the distribution F from a sample of size n from F.) A formal proof can be obtained from the main result in Doss and Gill (1992).

If f has a root r of multiplicity greater than 1 (to be more precise, of even multiplicity), then at r, f does not cross the s axis. Therefore, it is possible that f_n will not have a root near r. (To make things concrete, consider the case K=2, and suppose that the two probabilities p_1 and p_2 are equal. In this case, f is a parabola that has a minimum at the root r = 1 - 1/p, and for arbitrarily large n, the parabola f_n may lie entirely above the s axis, without contradicting the fact that f_n is uniformly close to f near the root.) Now, it is not difficult to establish that the roots of f_n converge to those of f, although some of the roots of f_n may be complex. In particular, the real parts of the roots of f_n converge to the roots of f, which are real. Thus, one can obtain the roots of f_n numerically (a number of math packages do this), and take the real part.

An alternate method is maximum likelihood. From (1) we have the distribution of $X_1, \ldots, X_n$, as a function of $p_1, \ldots, p_K$. Thus, we can write down a likelihood function, and in principle, we can maximize it. However, the maximization may have to be done numerically. If K is small, this is not difficult. Difficulties do arise if $K \geq 6$ or so, which is the attraction of the method based on estimating the roots of the probability generating function.

Once we have found the MLE, getting the estimates of variability is not difficult. We can calculate the observed Fisher information matrix by taking the second derivative of the log likelihood at the MLE. This does not involve a search. Although I have not checked this, I believe that the usual regularity conditions for asymptotic normality of MLE's are satisfied in this case. I see nothing pathological in this situation (except of course for the fact that the MLE is not obtainable in closed form).

References

An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *The Journal of the American Statistical Association* 87, 869-877, 1992 (with Richard Gill).

Non-identically distributed Bernoulli sums

Satish Iyengar
Department of Statistics
University of Pittsburg

(Editor's note: See the previous paper for a statement of the original problem)

Preliminaries

Let $\{X_j : j=1,\ldots, N\}$ be independent Bernoulli random variables with X_j having success probability p_j. We need the following notation:

$P = (p_1,\ldots p_N)$ is the vector of probabilities,

1 = (1,….,1) be an N-vector of 1s,

$S = \sum_{j=1}^{N} X_j$ is the total number of successes,

$q_j. = 1- p_j$ is the failure probability for X_j,

$\bar{p} = \frac{1}{N}\sum_{j=1}^{N} p_j$ is the average success probability, and $\bar{q} = 1-\bar{p}$,

$\sigma_p^2 = \frac{1}{N}\sum_{j=1}^{N}(p_j-\bar{p})^2$ is the variance of success probabilities, and

$\kappa_p = \frac{1}{N}\sum_{j=1}^{N}(p_j-\bar{p})^3$ is the third central moment of the success probabilities.

The mean of S is $E(S) = N\bar{p}$; its variance

$$\text{var}(S) = \sum_{j=1}^{N} p_j q_j = N\bar{p}\,\bar{q} - N\sigma_p^2$$

is smaller than that for a binomial with parameters $(N,\bar{p})$; call the distribution of the latter $B(N,\bar{p})$, and let T be a $B(N,\bar{p})$ random variable. The distribution of S is unimodal (by Theorem 4.8 of [2]); thus, S and T both have the same mean, with S more concentrated around it than T. We might therefore expect the tail probabilities of S to be smaller than those of T. Indeed, we have (by Jensen's inequality)

$$P(S=0) = \prod_{j=1}^{N}(1-p_j) \le (1-\bar{p})^N = P(T=0)$$

and

$$P(S=N) = \prod_{j=1}^{N} p_j \le \bar{p}^N = P(T=N).$$

There are several results that extend these inequalities. They are usually stated in the language of majorization: see Marshall and Olkin [4] for background. Here is one result. Let $h_t(p) = P(S \le t|p)$; then

$$h_t(p) \le h_t(\bar{p}\,\mathbf{1}) \text{ if } 0 \le t \le n\bar{p}-1$$

$$h_t(p) \ge h_t(\bar{p}\,\mathbf{1}) \text{ if } n\bar{p} \le t \le n$$

Other extensions that compare probabilities (and expectations) corresponding to p to those corresponding to $\bar{p}\,\mathbf{1}$ are given in [4, pages 359-375). They may be of interest where a bound or other qualitative information about such quantities is needed.

When more precise probability calculations are needed, the method depends on N. For small N, a simple enumerations is easy. In his memo, Hani Doss suggests evaluating the probability generating function, $g(t)=E(t^S)$, of S (using MAPLE or other package that does symbolic calculations), and then reading off the probabilities from the coefficients of that polynomial of degree N. For larger N, a central limit argument suggests a normal approximation. However, for finite N the distribution of S may well be skewed, and a correction for that is needed; an Edgeworth expansion with the skewness term provides such a correction. The third central moment of S, which is an indication of its skewness is

$$E(S-N\bar{p})^3=\sum_{j=1}^{N}p_jq_j(q_j-p_j)=N\bar{p}\bar{q}(\bar{q}-\bar{p})-3N\sigma_p^2(\bar{q}-\bar{p})+2N\kappa_p\,.$$

The first term on the right is the third central moment for the $B(N,\bar{p})$ variate T, and the next two terms are due to the variability among the components of p and their own third central moment, respectively. The comparison of the third central moment of S and T is more complicated than that of their variances. For instance, even if $\bar{p}=1/2$, so that T is symmetric, S can be skewed if $\kappa_p\neq 0$ (for instance N = 3 and p = (0.2,0.4,0.9)); and if $\kappa_p=0$, the third central moments of S and T can have opposite signs (for instance N = 2 and p = (0.01,0.97)).

An Approximation to the Distribution of S

Next, while the normal approximation works well near the center of the distribution (in this case near $N\bar{p}$), it can break down in the tails; exponential tilting provides a solution by modifying the distribution of S so that it is centered near the tail of interest. I will now describe an approximation using what is called the tilted Edgeworth expansion. This approach is well described by Barndorff-Nielsen and Cox [1, Chapter 4]. Maxumdar and Gaver [5] used this technique in another reliability context, that of estimating the loss of load probability for power generating systems. They considered also a correction for kurtosis, but found it to be not as satisfactory as the one including only the skewness term. Recently, Mazumdar and Iyengar [3] developed the details of the bivariate version; this paper studied only the skewness correction. Both [3] and [5] involved weighted sums of non-identically distributed Bernoulli random variables; the problem at hand is the special case with equal weights. Below, I give a sketch of the approximation; see [1] for further discussion. The details appear a bit involved, but the approximation is easy to program.

We need a few preliminaries. Let $\phi(x)$ denote the standard normal density, and $H_k(x)$ the k^{th} order Hermite polynomial with respect to ϕ; we will need $H_3(x)=x^3-3x$. Also, let

$$H_k(\alpha,\beta_1,\beta_2)=\int_{\beta_1}^{\beta_2}e^{\alpha x}H_k(x)\phi(x)dx\,.$$

To express $H_k(\alpha,\beta_1,\beta_2)$ conveniently, we need the following notation: for any function f(x), let $\Delta f(\beta_1,\beta_2) = f(\beta_1) - f(\beta_2)$. A direct integration gives

$$H_0(\alpha,\beta_1,\beta_2) = e^{\alpha^2/2}[\Phi(\beta_2-\alpha) - \Phi(\beta_1-\alpha)] = e^{\alpha^2/2}\Delta\Phi(\beta_1-\alpha,\beta_2-\alpha).$$

The repeated differentiation with respect to α of this expression, with the recursion $H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x)$ yields

$$H_k(\alpha,\beta_1,\beta_2) = e^{\alpha^2/2}\sum_{i=0}^{k}(-1)^i\binom{k}{i}\alpha^{k-i}\Delta\Phi^{(i)}(\beta_1-\alpha,\beta_2-\alpha)$$

for all integer k>0; below, we will need $H_0(\alpha,\beta_1,\beta_2)$ and $H_3(\alpha,\beta_1,\beta_2)$, both of which are easy to compute.

Returning to the problem at hand, let F denote the distribution of $S = \sum_{j=1}^{N} X_j$, and let

$$K(\theta) = \log E(e^{\theta S}) = \sum_{j=1}^{N}\log(p_j e^{\theta} + q_j)$$

be its cumulant generating function. Let the tilted distribution F_θ be given by

$$dF_\theta(x) = e^{\theta x - K(\theta)}\, dF(x),$$

and let S_θ be a random variable with distribution F_θ. The cumulant generating function of S_θ is

$$K_\theta(\eta) = \log E(e^{\eta S_\theta}) = K(\theta+\eta) - K(\theta).$$

Hence, the first three cumulants (which are the mean, variance, and third central moment) of S_θ are, respectively,

$$\kappa_1(\theta) = \sum_{j=1}^{N}\frac{p_j e^{\theta}}{q_j + p_j e^{\theta}},$$

$$\kappa_2(\theta) = \sum_{j=1}^{N}\frac{p_j q_j e^{\theta}}{(q_j + p_j e^{\theta})^2},$$

$$\kappa_3(\theta) = \sum_{j=1}^{N}\frac{p_j q_j (q_j - p_j e^{\theta})e^{\theta}}{(q_j + p_j e^{\theta})^3}.$$

By varying θ, the mean of the associated random variable S_θ can be anywhere from 0 to N; we will choose an appropriate value of θ later. The standardized variable

$$Z_\theta = \frac{S_\theta - \kappa_1(\theta)}{\sqrt{\kappa_2(\theta)}}$$

has first three cumulants 0, 1, and $\kappa_3(\theta)/\kappa_2(\theta)^{3/2}$. We will approximate the distribution of Z_θ by a standard normal, with an Edgeworth correction for skewness. Now, for any $0 \le a_1 < a_2 \le N$,

$$P(a_1 \le S \le a_2) = \int_{a_1}^{a_2} dF(x) = \int_{a_1}^{a_2} e^{-\theta x + K(\theta)} dF_\theta(x)$$

$$= E\left[e^{-\theta S + K(\theta)} I(a_1 \le S_\theta \le a_2) \right]$$

$$= E\left[e^{K(\theta) - \theta\kappa_1(\theta) - \alpha Z_\theta} I(\beta_1 \le Z_\theta \le \beta_2) \right] \qquad (1)$$

$$\simeq e^{K(\theta) - \theta\kappa_1(\theta)} \int_{\beta_1}^{\beta_2} e^{\alpha x} \phi(x) \left[1 + \frac{\kappa_3(\theta)}{6\kappa_2(\theta)^{3/2}} H_3(x) \right] dx$$

$$= e^{K(\theta) - \theta\kappa_1(\theta)} \left[H_0(\alpha, \beta_1, \beta_2) + \frac{\kappa_3(\theta)}{6\kappa_2(\theta)^{3/2}} H_3(\alpha, \beta_1, \beta_2) \right],$$

where $\alpha = \alpha(\theta) = -\theta\sqrt{\kappa_2(\theta)}$, $\beta_1 = \beta_1(\theta) = (a_1 - \kappa_1(\theta))/\sqrt{\kappa_2(\theta)}$, $\beta_2 = \beta_2(\theta) = .$ $(a_2 - \kappa_1(\theta))/\sqrt{\kappa_2(\theta)}$ Another approximation is the tilted normal approximation

$$= e^{K(\theta) - \theta\kappa_1(\theta)} H_0(\alpha, \beta_1, \beta_2),$$

which omits the adjustment for skewness.

Next comes the choice of θ. My approach is based on suggestions in [1]. If $N\bar{p}$ is contained in $[a_1, a_2]$, that interval is a central region, so I use $\theta = 0$; if $a_1 > N\bar{p}$, then I used θ satisfying $\kappa_1(\theta) = a_1$; and if $a_2 < N\bar{p}$, then I use θ satisfying $\kappa_1(\theta) = a_2$. That is, I tilt in such a way that the point in the interval nearest to $N\bar{p}$ is the mean of the tilted

distribution. Since $K(\theta)$ is a convex function of θ, $\kappa_1(\theta)$ is a monotone increasing function of θ, so the solution to each equation is unique. Some elementary analysis shows that if $r = a_1 / N > \bar{p}$, the solution is in the interval

$$\log\left[\frac{r/(1-r)}{\bar{p}/(1-\bar{p})}\right] \le \theta \le \log\left[\frac{r/(1-r)}{p_{min}/(1-p_{min})}\right],$$

where p_{min} is the smallest success probability; and that if $s = \alpha_2 / N < \bar{p}$, the solution is in the interval

$$\log\left[\frac{s/(1-s)}{p_{max}/(1-p_{max})}\right] \le \theta \le \log\left[\frac{s/(1-s)}{\bar{p}/(1-\bar{p})}\right],$$

where p_{max} is the largest success probability. I get the solution by simple bisection search.

I have written a Fortran program that evaluates the last expression in (1) using the tilted distribution when needed. It takes as input N, (a_1, a_2) and p. It outputs two probability estimates, a "tilted normal approximation" that does not use the skewness term, and a "tilted with skew" approximation that does use it.

I have checked the program for a few cases. The table below summarizes the results. I computed the exact probabilities by enumeration; TN refers to the tilted normal approximation, and TS refers to the tilted approximation with correction for skewness. I have used a continuity correction, so the boundary points a_i are adjusted by 0.5.

a. N = 5, p = (0.1, 0.1, 0.2, 0.3, 0.8)

(a_1,a_2)	Exact	TN	TS
(0.0,0.5)	**0.091**	**0.085**	**0.095**
(3.5,5.0)	**0.012**	**0.015**	**0.016**

b. N = 10, p = (0.1, 0.2, 0.2, 0.3, 0.4, 0.5, 0.6, 0.6, 0.7, 0.9)

(a_1,a_2)	Exact	TN	TS
(0.5,2.5)	**0.091**	**0.085**	**0.095**
(5.5,10.0)	**0.012**	**0.015**	**0.016**

Both the tilted normal and skew-corrected approximations seem fairly good for these small values of N. They both capture the first significant digit, and their accuracy is similar to that reported in [5].

Estimation of p

Hani Doss has outlined some approaches to the estimation of the probability vector p given data. *(ed. note: See preceding paper in this collection, H. Doss, "Memo on Non-Identically Distributed Bernoulli Model Problems for System Performance Prediction".)* Given the computational difficulties encountered there, I thought it might

be worthwhile to address a slightly different problem: the estimation of summaries of p, in particular, $\bar{p}$, σ_p^2 and κ_p. One solution to this problem is elementary, and it may provide useful information.

Let the data be $S_1,\ldots,S_n$, which are independent and identically distributed with the same distribution as that of S. Let

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i \quad \text{and} \quad \hat{\sigma}_S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(S_i - \bar{S})^2$$

be the usual unbiased estimates of the mean and variance of S. Then a simple estimate of $\bar{p}$ is $\bar{S}/N$. Since $\text{var}(S) = N\bar{p}\bar{q} - N\sigma_p^2$, a simple plug-in estimate of σ_p^2 is

$$\tilde{\sigma}_p^2 = \frac{\bar{S}}{N}\left(1 - \frac{\bar{S}}{N}\right) - \frac{1}{N}\hat{\sigma}_S^2,$$

which is slightly biased. An unbiased estimate of σ_p^2 is

$$\hat{\sigma}_p^2 = \frac{\bar{S}}{N}\left(1 - \frac{\bar{S}}{N}\right) - \frac{1}{N}\left(1 - \frac{1}{Nn}\right)\hat{\sigma}_S^2.$$

These estimates can be negative, so they must be truncated at zero; I have not checked to see how likely that is. Of course, unless n is small the two estimates $\tilde{\sigma}_p^2$ and $\hat{\sigma}_p^2$ will be quite close. Finally, since

$$\kappa_p = \frac{1}{2N}E(S - N\bar{p})^3 - \frac{1}{2}\bar{p}\bar{q}(\bar{q} - \bar{p}) + \frac{3}{2}\sigma_p^2(\bar{q} - \bar{p})$$

a plug-in estimate of κ_p is

$$\tilde{\kappa}_p = \frac{1}{2N}\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}(S_i - \bar{S})^3 - \frac{\bar{S}}{N}\left(1 - \frac{\bar{S}}{N}\right)\left(1 - \frac{2\bar{S}}{N}\right) + \frac{3}{2}\hat{\sigma}_p^2\left(1 - \frac{2\bar{S}}{N}\right)$$

References

Barndorff-Nielsen, O., Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics.* Chapman and Hall. London.

Dharmadhikari, S, Joag-dev, K. (1988) *Unimodality, Convexity and Applications.* Academic Press, New York.

Iyengar, S. and Mazumdar, M. (1998) "A Saddle Point Approximation for Certain Multivariate Tail Probabilities" *SIAM Journal of Scientific Computing.* In press.

Marshall, A., Olkin, I. (1979) *Inequalities: Theory of Majorization and its Applications.* Academic Press, New York

Maxumdar, M., Gaver, D.P. (1984) On the computation of power-generating system reliability indexes. *Technometrics* 26:173-185

Large-Scale Civilian Biometric Systems—Issues and Feasibility

James L. Wayman, Director
U.S. National Biometric Test Center

I. INTRODUCTION

The Requirement for Large-Scale Civilian Identification Systems

In the United States, two recent pieces of federal legislation are pushing limited development of large-scale civilian biometric identification systems: The Personal Responsibility and Work Opportunity (Welfare Reform) Act of 1995 and the Immigration Control and Financial Responsibility (Immigration Reform) Act of 1996. Although neither Act uses the word "biometric", both call for use "technology" for identification purposes.

The Welfare Reform Act[1] requires the implementation by the States of an electronic benefits transfer program, "using the most recent technology available that the State agency considers appropriate and cost effective and which may include personal identification numbers, photographic identification on electron benefit transfer cards, and other measures to protect against fraud and abuse". Further the Act requires penalties for anyone making "a fraudulent statement or representation with respect to the identity or place of residence of the individual in order to receive multiple benefits simultaneously under the food stamp program"[2].

The Immigration Reform Act calls for the "Development of a New Verification System"[3] and requires the President to "develop and recommend to the Congress a plan for the establishment of a data system or alternative system …to verify eligibility for employment in the United States, and immigration status in the United States for purposes of eligibility for benefits under public assistance programs…or government benefits."[4] This system "must be capable of reliably determining with respect to an individual whether… the individual is claiming the identity of another person." The President must "submit to Congress a report …including estimates of…the accuracy rate of the system; and the overall costs and benefits that would result from implementation...".

These two pieces of federal legislation are driving State efforts to construct large-scale biometric identification systems. Internationally, a number of identification efforts for voter registration, driver's licensing, social security enrollment, immigration control and national identification are under way in countries such as the Philippines, Malaysia, Honduras, Costa Rica, Panama, El Salvador, South Africa, Egypt and Guatamala.

---

1 U.S. Public Law 104-327, Section 1034

2 U.S.P.L. 104-327, Section 1029

3 U.S. House Resolution 2202, title of Part 2, subpart A.

4 U.S. H.R.2202, Section 111.

Putting aside issues of privacy and legality for discussion in another forum[5], the technical questions of, "Is a large-scale civilian biometric identification system technically feasible?", and "If so, how well will it perform?", have not been carefully addressed. The goal of this paper is to look at exactly these issues from a scientific point of view, creating mathematical performance models and developing model parameters from a set of carefully controlled experiments.

Large-Scale Use Of Biometrics: Fingerprinting

The only large-scale (involving over 30,000 individuals) application to date of biometric technology has employed fingerprinting[6]. Therefore, we have chosen to focus this study on the performance of large-scale Automatic Fingerprint Identification Systems (AFIS).

The function of an AFIS is to correctly match input "sample" fingerprints to prints already "enrolled" in a database, or to determine that no enrolled prints matching the samples exist. Depending upon the application, the database comparison may be over millions of enrolled prints, or may be to one print selected from a larger database. There are three primary goals in the design of such a system: 1) The number of incorrect matches must be minimized; 2) The number of prints incorrectly found to not match prints actually in the enrolled database must be minimized; 3) The throughput rate of these comparisons must match the input rate of the prints (as averaged over a time scale appropriate for the system operational requirements), so as to prevent backlogs from developing. These three goals are actually competing design parameters, the interactions of which are not generally understood outside the AFIS scientific community.

This paper will: 1) explain, in mathematically-based, scientific terms, the first-order interaction of these parameters in a "civilian" AFIS system; 2) report results of a closely-supervised international benchmark test of systems from four of the world's best known AFIS developers; and 3) predict large-scale civilian system performance based on the measured values. We will first offer two tutorial sections to explain the difference between "civilian" and "forensic" systems and to outline the operation of a "civilian" system.

II. TUTORIAL: "CIVILIAN" AND "FORENSIC" AFIS

There are two types of AFIS systems: 1) "Forensic" systems, used by law enforcement for the purpose of matching "latent" prints left at a crime scene, or prints collected from known or suspected criminals or persons applying for especially sensitive positions, against prints in a database of externally identifiable individuals; 2) "Civilian"

---

5 K. Nuger and J.L. Wayman, "Reconciling Government Use of Biometric Technologies with Due Process and Individual Privacy", U.S. National Biometric Test Center, to be published.

6 A study of retinal scanning using 26,000 individuals was performed by the California Department of Motor Vehicles in 1989 under contract with the Federal Highway Administration. See "Personal Identifier Project: Final Report", The Orkand Corporation, (DMV 88-89), April 26, 1990, reprinted by the U.S. National Biometric Test Center.

systems, used by non-law enforcement government and non-government agencies for the purpose of matching persons applying for or receiving benefits, privileges or services to records already in the database.

Forensic Systems

Forensic systems must be able to identify poor quality, partial prints left at a crime scene as well as to match prints submitted "live" (using electronic scanners) and on inked cards. The first function is the most difficult and drives the design of the system. Therefore, the enrollment record generally consists of the prints of all ten fingers from each enrolled individual. These prints are generally "rolled", meaning captured through a rolling motion of the fingers, so as to image the sides of the finger area as well as the flat "core" of the fingerprint. Historically, enrollment has used ink on cards, so these systems must be able to accept rolled fingerprint images from such cards, as well as images taken directly from the fingers of enrollees using electronic fingerprint scanners. Implicit in this process is the suspicion that the enrollee may have a criminal record, even when enrollment is for the purpose of qualifying for employment in a "sensitive" profession (jobs involving children, military service, or national security). Enrollment time requirements are secondary to obtaining both complete prints and external information (public identity, address, etc) identifying the individual from whom the prints are taken. Entering the enrolled prints into the system may require human intervention, including the classification by type (arch, loops, whorl, and associated subtypes) by trained examiners. The enrolled print images must be accessible for use by human examiners or for use in court during criminal proceedings and consequently must be of high image quality, as judged by the human eye. Maintenance of a "clean" database, meaning one set of prints per individual, may be desirable, but is not mandatory. In fact, the maintenance of several sets of prints on a single individual may actually serve to assure that future samples from that individual are matched to at least one identity in the database.

For the searching of sample prints against the enrolled database, the forensic system must accept as input photographs of partial, latent prints taken from crime scenes. Extensive human intervention during the input of such prints is assumed. During the search, the implicit goal is the matching of the input print against all possible "candidate" prints in the database. A high implicit penalty is assessed to the failure to match a sample print to one actually in the database. Consequently, searches may result in "candidate lists" of potentially matching prints, requiring final matching by expert human examiners.

Forensic systems must support interoperability standards, being able to exchange images with other jurisdictions.

Civilian Systems

Civilian systems are different from forensic systems in many ways: some subtle, some obvious. Even some AFIS vendors do not, to their peril, understand these distinctions. Civilian systems accept enrollment prints from only a few (one to four)

primary fingers, either thumbs or forefingers[7]. Accept to accommodate persons (considered as "customers") who cannot appear in person, prints are always input directly from electronic fingerprint scanners. These scanners image only the "flat", core area of the finger, with no requirement for the finger to be rolled. These prints must be input electronically into the AFIS system with no human intervention after the collection stage. Any classification done on the prints must be performed automatically. Externally identifying information, such as name, address, etc., may or may not be required, so the primary purpose of the system is to identify people to previously enrolled records, not identities. Enrollment time must be minimized for both the convenience of the customer and for keeping down operational costs. Implicit in this process is the assumption that enrollees are not suspected of any crime and that their identity is being protected within the system from exploitation by others who may wish to falsely claim it. Any indication that there is a presumption of criminality or that obtained prints will be used outside of the system is strongly avoided. Some state and local laws make illegal the general release of prints taken by government operated civilian systems to law enforcement[8]. Protection of the enrollees identity and the identification of the enrollee is not necessary, in fact may even be undesirable, outside the system. There is no implicit assumption of criminality of the customers. Future recall of the fingerprint images may or may not be a system requirement.

Depending upon the transaction, sample prints may be searched against the entire database or only against a single, claimed record. Unlike the forensic system, implicit in the search strategy is the assumption that the prints will not be found in the database except as claimed. Also in contradiction with the forensic system, a high implicit penalty is assessed against the return of any false matches. Consequently, civilian system design and the selection of operational parameters may be far different than that for forensic systems. Civilian systems must not develop candidate lists and must operate with but perhaps a single trained examiner used only for system fraud investigations. Interoperability with other systems is often prohibited by law. Some administrators have even sought assurances that their system will be unable to inter-operate with forensic systems.

---

[7] We know of no existing or proposed civilian systems accepting more than four fingers, nor systems using fingers other than the forefingers and thumbs. Social service systems generally use forefingers because of the perceived association of thumb print collection with criminal investigations. Driver's licensing systems within the U.S. use either thumbs or forefingers, depending upon State.

[8] H.R. 2202, Section 111.c.1.C. forbids law enforcement use of the "new verification system" except for direct enforcement of the provisions of the Act. State laws limiting law enforcement access to driver's licensing fingerprint records exist in California, Texas, and Georgia, as well.

III. TUTORIAL: HOW CIVILIAN AFIS WORK

Figure 1 shows a generic biometric identification system. In previous papers[9,10], we discussed in detail this diagram. In this section, we will focus on the storage, signal processing, matching, and decision policy subsystems in civilian AFIS.

Both sample and enrollment fingerprints may or may not be stored as images in the storage subsystem for future recall. Prior to the development of the National Institute of Standards and Technology (NIST) and Federal Bureau of Investigation (FBI) "Wavelet Scalar Quantization"[11] (WSQ) compression standard in 1992, images were generally stored using JPEG (Joint Photographic Experts Group) compression. Both WSQ and JPEG are "lossy" compression algorithms, but the effect of JPEG on image recognition can be severe. "Legacy" systems using JPEG storage still exist, but all recently constructed systems use the superior WSQ.

The signal processing subsystem extracts "features" from the input images. Theoretically, these features may be smaller portions of the image, mathematical transforms of the image, or "minutiae" points extracted from the image. For historical reasons[12] owing to the original development of AFIS for forensic applications, all large-scale civilian AFIS vendors currently use the "minutiae" extraction technique of human forensic experts, although other approaches are used by smaller vendors[13] and in university and government sponsored[14] projects. There are no standards for feature extraction and all such techniques are proprietary to each AFIS vendor. These extracted features may be 100 to 1000 bytes in length and are always stored in the storage subsystem. There is no way to recover the fingerprint image from the extracted features, so systems storing only features are non-interoperable with other systems. A change in vendor for such systems requires re-enrollment of the entire database.

Besides extracting features, the signal processing subsystem has three other functions: pattern classification, quality control and feature matching. The purpose of pattern classification is to allow the "binning" of fingerprints during the matching

---

9 J.L. Wayman, "A Scientific Approach to Evaluating Biometric Systems Using a Mathematical Methodology", Proc. CardTech/SecurTech'97, pg. 477-492

10 J.L. Wayman, "The Science of Biometric Technologies ", Proc. CardTech/SecurTech'97, pg.

11 NIST/FBI, "Minimum Image Quality Requirements for Live Scan, Electronically Produced Fingerprint Cards, Appendix F/G", IAFIS-IC-0010(V2), April 1993

12 Francis J. Galton, "Personal Identification and Description", *Nature*, June 21 and 28, 1888, pg. 173-177, 201-202.

13 The fingercprint matching software used by Comparitor Systems, for instance, was based on one-dimensional Fourier transforms. Two-dimensional correlation matching , Hough and Fourier transform techniques are being used at San Jose State University by students and faculty members in projects not associated with the National Biometric Test Center.

14 Jay Stosz and Lisa Alyea ,"Fingerprint Authentication", Proc. CardTech/SecurTech'95, pg. 201-219.

process. In large-scale systems, it is computationally inefficient to match features of each input print against the stored features of all enrolled prints. Some stored prints can be eliminated from comparison on the basis of differences in pattern classification. For instance, many civilian AFIS use a classification system very close to that originally proposed by Galton[15], based on arches, left loops, right loops, and whorls, but there are other approaches as well. Some vendors simultaneously use multiple binning techniques.

Many prints will be difficult to clearly classify and will be given multiple classifications or designated as "unknown". Input prints will be matched only against stored prints that have been given at least one of the same classifications or designated as "unknown". Consequently, input prints will require matching against only a portion of the stored database. This portion, expressed as a percentage, is known as the "penetration rate". The lower this "penetration rate", the fewer comparisons will be expected and the more efficient the system.

If the system uses multiple fingers, it is possible to bin each finger according to the classifications of the fingerprint ensemble. In other words, if two prints are taken from a customer, the first being of classification A and the second of classification B, each print can be binned according to the classification of the two print combination, AB. The sample first print will be compared only to first prints in the database from people whose ensemble classification was also AB, although unknowns must also be considered. This allows for a multiplicative decrease in penetration rate for multiple print systems.

There is a cost in terms of error to be paid for this increase in matching efficiency, however. If a sample print is placed in a different bin or bins than a truly matching print in the database, the two prints will never be compared and a false non-match will result. The probability that a print will be inconsistently binned is known as the "bin error rate".

At this point we should mention fingerprint "filtering", often confused with "binning" because its goals are the same. "Filtering" involves additional partitioning of the database based on information, such as gender or age of the customer, which is not contained in the fingerprint image itself. Identification of the finger ("right thumb", for instance) cannot be made based on the fingerprint pattern, so the partitioning of the database by finger, as done in all AFIS, is a filtering process. Because filtering is based on exogenous information, it is not part of the signal processing process, but rather, is part of the data collection process accompanying the sampling of the customers. Flow of this information is not shown in Figure 1.

Filtering also results in errors, but these errors are those made by the human operators of the system, perhaps encouraged by deceptive activities of customers attempting fraud, and cannot be estimated by engineering tests. In other words, filtering leads to search efficiencies while externalizing the associated errors, so consequently is greatly appreciated by AFIS vendors.

Civilian AFIS generally transmit the images to the signal processing subsystem and perform the signal processing in "real time", meaning while the customer is still at the image scanning device. After the extraction of the features, the sample is checked for "quality", generally related to the number of minutiae extractable from areas of the print where the ridge structure is coherent. No current vendor, to our knowledge, has quality

---

[15] Francis Galton, Fingerprints, (London, McMillan, 1892)

control software capable of detecting the presence of the fingerprint "core". The lack of this capability appears to adversely effect the performance of the matching module when comparing prints. If the quality is deemed insufficient, the system operator is instructed to collect a replacement sample from the customer. When the quality of the received image is deemed sufficient, or when overridden by the system operator in the case of difficult prints, the extracted minutiae features are sent to the matching module.

The function of the matching module is to compare features of the input print to features of candidate prints in the database one print at a time. In the case of verifying a claimed identity[16], there is only one candidate print matched from the database. In the case of a general search of the database to determine if the person is previously known to the system (used in social service applications, for instance, to prevent benefit fraud by multiple enrollment), the matcher will compare, one at a time, features from all stored prints which have one or more classifications in common with the sample print. Sample prints designated as being of "unknown" classification will require comparison to all prints in the database. Similarly, any prints in the database classified as "unknown" will be compared to all input sample prints. For each comparison, the matcher passes to the decision subsystem a numeric measure of the similarity of the sample and database features. This measure differs from that used in other types of biometric systems in that higher values indicate a closer match between the sample and database prints.

It is the function of the decision subsystem to determine if a "match" has been found based on some decision policy established by the system user (not the vendor). The policy may be to declare a "match" if the similarity measure is above some threshold, or if the sum of two consecutive measurements is above some threshold, or under any of a set of conditions limited only by the imagination of the designers. The system policy may be to "accept" a customer for enrollment if no match is found over some number of presented fingers and to "reject" a customer for enrollment if matches are found on some number of fingers. The system might also "accept" a customer for renewal if a match is found against some number of claimed prints and "reject" a customer for renewal if no match is found. In any case, the precise decision policy is a question to be worked out during implementation by the system user and will vary according to the functions the system is required to perform. System testing and performance issues must be considered as independently as possible from system decision policy.

IV. AFIS PERFORMANCE PARAMETERS

There are five important, non-independent parameters that govern the performance of Automatic Fingerprint Identification Systems. These are: 1) the "penetration rate", reflecting the expected percentage of the fingerprint database to be compared to a sample print; 2) the "bin error rate", or probability that a search for a print

---

16 It is important to remember that an AFIS can only match a person to a record inside the system. It cannot externally validate that person with a public identity. In other words, the AFIS can only match me with a record of someone previously claiming to be "James L. Wayman". It cannot assure that I really am that person. That information can only come from external documentation (birth certificate, passport, etc.) presented at the time of enrollment and is no more valid than that documentation.

in the database will be unsuccessful because the sample and template prints were placed different "bins"; 3) the single comparison false match rate, or probability that two non-matching prints will be incorrectly matched; 4) the single comparison false non-match rate, or probability that two matching prints will be incorrectly not matched when compared; 5) the "one-to-one", or "cold match" comparison rate of the hardware.

System Penetration Rate

The system penetration rate reflects the matching efficiencies achieved by placing the database fingerprints into "bins" based on classification type, ridge count or some other measure endogenous to the fingerprint itself, and gains based on exogenous "filtering" techniques, including identification of the finger in multiple finger systems and identification of the customer's gender. Generally, a single print can be placed into multiple bins or filter partitions if there is uncertainty regarding its classification. Some prints of extreme uncertainty as to classification are labeled as "unknown" and placed in all of the partitions. In operation, a sample print is classified according to the same system as the database, then matched against only those prints from the database which are in the same classification or classifications. The average, or expected, percentage of prints to be matched for each input sample is the "penetration rate". Of course, the smaller the penetration rate, the more efficient the system.

Binning and filtering are generally independent operations and can be considered separately. In multiple fingerprint systems, system penetration rate is a function of the binning of the single fingers. This binning is not statistically independent, meaning that if the left thumb is a loop, for instance, the right thumb is likely to be a loop. Correlations between finger binnings are currently not widely known, so in developing our equations, we make the incorrect assumptions that the binnings are statistically independent between fingers and the same for all fingers. As possible, we will point out the direction, although not the magnitude, of the error this causes. Under these incorrect assumptions in an M-finger system, the multiple finger penetration rate, P_{mf}, can be written as

$$P_{mf} = P_{sf}^{M} \tag{1}$$

where P_{sf} is the single finger penetration rate. We note that penetration rate decreases geometrically with M. Correlations between finger binning actually causes this penetration rate to be higher (worse) than calculated using this equation.

Providing that filters are independent, as is generally the case, filtering factors are similarly multiplicatively combined, where the system filter factor, F, can be calculated from each individual filter factor, Fi, as

$$F = \prod_{\text{all filters}} F_i \tag{2}$$

We can calculate the filter factors, Fi, from the partition distributions only in the case where a person can be placed in only one partition. In the case of gender-based filtering, if a person can be estimated as only male or female, with probabilities p_{male} and p_{female} , the size of the male bin will be $N*p_{male}$ and the size of the female bin will be $N*p_{female}$. Assuming that the database and sample prints are from populations of identical gender distribution, the expected number of searches is

$$E(\text{searches}) = p_{male} * N * p_{male} + p_{female} * N * p_{female} = N\sum p_j^2 \quad (3)$$

and the filter factor can be seen to be

$$F_i = \sum_{\text{non-intersecting bins}} p_j^2 \quad (4)$$

In the case of an "unknown" partition with probability, $p_{unknown}$, the unknown partition must always be searched and the expected number of searches becomes

$$E(\text{searches}) = N * p_{unknown} + (p_{male} + p_{unknown}) * N * p_{male} + (p_{female} + p_{unknown}) * N * p_{female} =$$

$$N\left[p_{unknown} + \sum (p_j + p_{unknown}) p_j\right] \quad (5)$$

The filter factor is

$$F_i = \left[p_{unknown} + \sum (p_j + p_{unknown}) p_j\right] \quad (6)$$

where the summation is over all bins j=1,2…K.

We have not yet been able to develop an expression for the filter factor based on partition distributions when prints can be placed in multiple partitions. Such an expression must include the correlations between partitions. Equations (4), (5), and (6) for filtering apply also to binning, but again only in the case where prints cannot be placed into multiple bins. Given that

$$p_{unknown} + \sum_{j=1}^{K} p_j = 1 \quad (7)$$

we can see the general principal that F decreases with increasing number of bins K of non-zero probability. This principal also applies to binning penetration rate.

The system penetration rate, P_{sys}, is the product of the bin penetration rate and the filter factor

$$P_{sys} = P_{mf} * F \quad (8)$$

Bin Error Rate

The bin error rate reflects the percentage of prints falsely not matched because of inconsistencies in the binning process. Filtering errors, made by human operators during the customer interview process, occur outside of the "automatic" boundaries of the Automatic Fingerprint Identification System, so are not considered within the AFIS. In a multiple print system using ensemble binning, to not make a bin error requires that no bin error be made for any print. This awkward English actually best describes the underlying probabilistic relationship

$$1 - \varepsilon_{\text{bin ensemble}} = (1 - \varepsilon_{sfbin})^M \quad (9)$$

where $\varepsilon_{\text{bin ensemble}}$ is the system bin error rate, $\varepsilon_{\text{sfbin}}$ is the single finger bin error rate, and M is the number of fingers in the ensemble. Equation (9) assumes that bin errors are independent between fingers. This might not be true if the extent of finger damage is related between fingers. If damage is related, the true system error rate would be lower.

Equation (9) can be rewritten as

$$\varepsilon_{\text{bin ensemble}} = \text{M} * \varepsilon_{\text{sfbin}} - \text{O}(\varepsilon_{\text{sfbin}}^2) \tag{10}$$

where $\text{O}(\varepsilon^2_{\text{sfbin}})$ indicates terms of order $\varepsilon^2_{\text{sfbin}}$. For small $\varepsilon_{\text{sfbin}}$, as is the general case, (10) reduces to

$$\varepsilon_{\text{bin ensemble}} \approx \text{M} * \varepsilon_{\text{sfbin}} \tag{11}$$

We saw in the above section that penetration rate decreased geometrically with M. Here we see that system bin error rate increases arithmetically with M. This indicates the general operational tradeoff between decreasing penetration rate and increasing bin error rate. This same relationship holds with increasing number of bins, K, for which $p_j \neq 0$: Penetration rate decreases while bin error rate increases.

The Single Comparison False Match Rate

A single comparison false match occurs when a sample print is incorrectly matched to a print in the database by the decision subsystem because the similarity score between the two exceeded a fixed threshold. The "impostor" probability distribution function, $\Psi_I(s)$, is a function of the positive similarity measure s, which increases with increasing similarity between compared prints. Unlike other biometric systems, the impostor distribution function is closer to the origin (s=0) on the abscissa than the "genuine" distribution of similarity scores between truly matching prints.

The single comparison false match rate can be expressed as a function of decision threshold, τ, as

$$\text{FMR}(\tau) = \int_{\tau}^{\infty} \Psi_{\text{I}}(\text{s})\text{ds} = 1 - \int_{0}^{\tau} \Psi_{\text{I}}(\text{s})\text{ds} \tag{12}$$

which decreases with increasing decision threshold.

The Single Comparison False Non-Match Rate

A single comparison false non-match occurs when a sample print is incorrectly not matched to a print from the same finger by the decision subsystem because the similarity score between the two is less than a fixed threshold. The single comparison false non-match rate, FNMR, can be given as a function of decision threshold, τ, as

$$\text{FNMR}(\tau) = \int_{0}^{\tau} \Psi_{\text{G}}(\text{s})\text{ds} \tag{13}$$

where $\Psi_G(s)$ is the genuine probability distribution function. FNMR increases with increasing decision threshold. It is clear from equations (12) and (13) that false match and false non-match rate are competing factors based on the threshold.

Hardware Comparison Rate

The "cold match" comparison rate is the number of "one-to-one" comparisons per second that can be made by the hardware of a single "sample" print to "template" prints retrieved from the database. It is a function of the hardware processing speed, the template size, and the efficiency of the matching algorithm. AFIS system architecture is modular in the sense that processing speed can be designed to meet seemingly any requirement, although there are no doubt limits of scale as speed requirements get too great. In general, a single comparison may take one or two million operations, and as a rule of thumb, hardware costs run several US$ per match per second. Measurement and prediction of system processing speed from component architecture or from direct measurement is beyond the scope of our current capabilities and will not be considered further in this paper.

V. AFIS PERFORMANCE EQUATIONS

We are now in a position to write some first-order equations reflecting the interaction between these parameters and system performance. By "system performance", we mean that we are concerned with the acceptance and rejection of customers, as they are represented by an ensemble of prints. As precise performance prediction will depend upon system decision policy, which will be determined only in full-scale implementation, our goal here is to bound performance. Owing to both complication and lack of data, we will ignore any and all correlations between errors, expressing where we can the impact this has on the computed performance bound.

System Throughput

The first equation is an approximation for the system throughput rate, T, which depends upon: the comparison rate, C; the system penetration rate, P_{sys}; the number of records in the database, N; and the number of fingers to be matched, m. We have given the lower case symbol, m, to the fingers to be matched to differentiate it from the previously used, M, the number of fingers collected, upon which ensemble binning is performed. In all cases, $m \leq M$

In multifinger system, initial search can be done on a subset, m, of the collected fingers, M. Any matches determined for any finger can be verified against the remaining M - m fingers. Under the assumption that no matches will be found, the throughput rate can be written as

$$T = \frac{C}{P_{sys} * N * m} \tag{14}$$

Violations of our assumptions regarding finger binning independence increase penetration rate and decrease throughput. Any matches found (false or correct) require an additional M - m comparisons, further decreasing throughput, so this equation is an optimistic upper bound.

This throughput rate must match the customer input rate, I, as averaged on a time scale driven by operational requirements. Because of the various time units used, care must be taken in dimensional balancing. Therefore, if customer throughput must match customer input on a daily basis, we can write

$$\frac{C \ \text{comparisons/sec} * O \ \text{operational secs/day}}{P_{sys} * N \ \text{comparisons/finger} * m \ \text{fingers/customer}} = I \ \text{customers/day} \quad (15)$$

We note that the penetration rate, P_{sys}, is a percentage, and therefore non-dimensional..

System False Non-Match Rate

A system false non-match occurs when some minimum number of prints from the customer's ensemble are falsely not matched to existing records. Let's assume a bounding policy that a customer is not matched if all m searched prints are not matched. A searched print can be falsely not matched because of a binning error or because of a single comparison false non-match. The probability that a single searched print has neither can be expressed as

$$1 - FNM_{sf} = (1 - \varepsilon_{bin\ ensemble})(1 - FNMR) \quad (16)$$

where FNM_{sf} is the probability that the system falsely does not match a single searched finger, $\varepsilon_{bin\ ensemble}$ is the probability of a binning error over the ensemble, and FNMR is the single comparison false non-match rate. The explicit dependency of FNMR and FNM_{sf} on threshold, τ, has been dropped for notational simplicity.

Providing that both FNMR and $\varepsilon_{bin\ ensemble}$ are small, equation (15) can be rewritten as

$$FNM_{sf} = \varepsilon_{bin\ ensemble} + FNMR \quad (17)$$

Therefore, a system false non-match occurs if all m searched prints are falsely non-matched, so

$$FNM_{system} = (\varepsilon_{bin\ ensemble} + FNMR)^m \quad (18)$$

Positive correlation on a single finger or over a single customer between binning errors and false non-match errors will cause FNM_{system} to be lower than predicted.

System False Match Errors

A system false match occurs when some number of prints from the customer's ensemble are falsely matched to prints of a single customer in the database. Let's assume a bounding policy that a customer is matched if all M collected prints are matched to another customer's. For this to happen requires that one of the searched prints be falsely matched to an enrolled print in the searched portion of the database and that the other M-1 prints be falsely matched to that same ensemble. The probability that a single searched print is not matched against any print in the searched database is

$$1 - FM_{sf} = (1 - FMR)^{P_{sys} * N} \quad (19)$$

where FM_{sf} is the system single comparison false match rate, FMR is the single comparison false match rate, P_{sys} is the penetration rate, and N is the size of the database. Again, explicit dependency of FM_{sf} and FMR on threshold, τ, has been dropped for notational simplicity.

Equation (19) can be rewritten as

$$FM_{sf} = P_{sys} * N * FMR - O(FMR^2) \tag{20}$$

where the last expression on the right-hand side indicates terms of order FMR^2 or higher. Assuming FMR to be small, the above equation reduces to

$$FM_{sf} \approx P_{sys} * N * FMR \tag{21}$$

If the bounding system match policy requires a match on all M fingers for a system match to be declared, the probability of a system false match, FM_{system} ,occurring is

$$FM_{sytem} \approx P_{sys} * N * FMR^M \tag{22}$$

Due to the apparent confusion over this equation[17], it might be instructive to insert a few values to demonstrate the feasibility of large-scale AFIS systems. Allowing P_{sys}=0.0125, N=8x10^7 , M=4 and FMR=10^{-3} , the system false match rate becomes 10^{-6} . This means that a system using an ensemble of four prints, containing records of 20 million persons, can operate with less than one false match per one million input customers, providing that the single comparison false match rate can be kept below one in one-thousand.

VI. THE PHILIPPINE SOCIAL SECURITY SYSTEM BENCHMARK TEST

The Republic of the Philippines Social Security System (SSS) Identification Card Project AFIS benchmark test was conducted in May, 1997, with four international AFIS vendors, a fifth vendor withdrawing immediately prior to the test. The goals of the test were to measure penetration rate, bin error rate, and single comparison false match and false non-match rates for each vendor. We did not attempt to measure hardware comparison rate, as the conditions of the contract require minimum throughput rate performance. The benchmark was undertaken in support of a national social security card project, with an anticipated eventual enrollment of 20,000,000 cardholders and an enrollment rate of 20,000 people per day when the full enrollment is reached. The system will use gender-based filtering and four fingers (left and right forefingers and thumbs) for enrollment.

Fingerprint Images

To facilitate the test, we collected three sets of images: “training”, “practice” and “test”. All images were taken from SSS adult employee volunteers, each giving eight prints at each session, thumb through ring finger of each hand. The volunteers were primarily managerial and clerical workers, although some volunteer laborers were

---

17 Richard Hopkins, “Benchmarking Very Large-Scale Identity Systems”, Proc. CTST’97, pg. 314-332

solicited, as well. Each volunteer signed a consent form authorizing release of the collected data. The collection was supervised by three SSS employees. No personal data was collected with the prints other than gender; 55% of the volunteers were women. Database bookkeeping was accomplished by assigning each volunteer a collection number. Handwritten data sheets connecting volunteers with collection numbers was maintained by the supervising employees and have since been destroyed.

Prints were imaged with an Identicator DF-90 "flat" scanner, believed to be "Appendix F"[18] compliant and an "MRT" frame grabber in a lap-top computer. Front-end quality control software from Identicator was employed. Database software was custom supplied by Identicator for this project. The prints were stored, using loss-less compression, as "TIFF" images. Some image quality loss, attributable to external electromagnetic noise at the time of collection, was noticed in the upper right hand quadrant of each image.

The "training" data consisted of 4080 prints taken from 510 volunteer employees of the Social Security System over a three-week period. It was our original intent that the "training" data set be "clean", meaning free from duplicate images. Subsequent analysis by the vendors indicated that there were, in fact, 4 repeated prints in the set made by inadvertently inaccurate finger presentations by volunteers which were not corrected by the supervisors.

In collecting the 4080 "training" images, we were allowed to physically touch the volunteers, manipulating their fingers on the scanner and applying slight pressure with the intent of obtaining the highest print quality possible as judged by the quality control software. Moisturizing compound was applied as needed. In general, three or more prints were collected from each finger of each volunteer, although the training data set consisted of only one image from each collected finger.

A second, "test", data set of 4128 images was collected from 506 volunteers, 409 of whom were included in the training set Although image quality was checked and moisturizing compound applied as needed, somewhat less care was generally taken to provide high quality images. Collection of the "test" set commenced one week after completion of the "training" set collection and was finished within three weeks. Consequently, individual volunteers were imaged at an interval of one to six weeks for the test and training sets. The test set contained 10 duplicate ensembles (80 prints) imaged from 10 volunteers in a separate session several weeks after the completion of the collection of the other images.

The order of the files and the file names were scrambled to prevent a determination of correlation between "test" prints and "training" prints or correlations within the "test" set, and a highly-secret key was created linking the "test" and "training" prints.

The third, "practice" set of 80 images was taken from 10 volunteers whose prints were in the "training" data set. The file names given these images were adjusted to identify them with their matching image files in the "training" set.

---

[18] "Minimum Image Quality Requirements for Live Scan, Electronically Produced, Fingerprint Cards, Appendix F -IAFIS Image Quality Specifications", NIST/FBI document IAFIS-IC-0010 (V2), April 22, 1993.

Both the "training" and "practice" data sets were mailed to the vendors several weeks prior to the tests. The "test" data set was hand delivered to the AFIS vendors on the day of the benchmark test. In all cases, testing was completed within the day.

Some Comments On Test Design

Our test is of a "symmetric" design, in that the training and test sets are about equal in size and about equal to the number of matching pairs. Another common design is to include with the training prints a "background" database of non-matching prints. This latter practice has the sole function of increasing the degrees of freedom in the false match testing, allowing for smaller uncertainty intervals in the evaluation of single comparison false match rate. The interpretation of resulting false match versus false non-match (ROC) curves of greatly varying degrees of freedom along abscissa and ordinate is not currently understood by us, although this problem is under investigation.

The choice of about 4000 matching pairs gave us the possibility of testing both false non-match and false match rates down to about one part in one thousand[19], although the issue of "degrees of freedom" for statistics arising from the approximately 4080x4128 "cross-matches" using only about 5000 independent fingerprints has not yet been completely resolved. Further, the practical problems of cost and record keeping associated with collecting and handling more than 8,000 prints and 16,000,000 cross comparisons are indeed daunting. For these reasons, we settled on the use of 4,000 matched pairs as the test database.

The use of eight fingers from each individual was done with the implicit assumption that the eight prints in each ensemble would be independent, thus allowing the use of the 4,000 pairs as though they came from 4,000 individuals. This assumption was subsequently challenged by the test results, as will be discussed.

Test Requirements

Prior to receipt of the "test" images, the vendors were required to supply the binning results for the "training" images. Vendors were allowed to report either "hard" or "soft" binning results, provided that enough information was supplied for analysis. By "hard", we mean the assignment of each print to one or more discrete bins. By "soft", we mean the assignment of numerical values to each print representing in some way a probabilistic binning assignment.

Vendors were required to report the binning assignments of each print of the "test" set using the same format as used in reporting bins of the "training" prints. Then vendors were required to match the "test" to the "training" prints, again reporting either "hard" or "soft" results. A "hard" result was a "match" determination between files. A "soft" result was a numerical similarity measure between files. Vendors choosing to return "soft" results were required to submit a 4080 by 4128 matrix of similarity measures.

For reasons having to do with Republic of the Philippines government procurement rules, test results were required to be given a "pass/fail" evaluation against criteria announced prior to the test. Our announced criteria were: 1) a 0.1% system false

---

19 As discussed in Wayman(1), ibid.

match rate; 2) a 5% system false non-match rate[20]; 3) compliance with the required throughput rate, given the vendors' claimed hardware comparison rate. As explained in previous sections of this paper, these are competing goals. Vendors submitting "soft" results had their thresholds set by us during analysis so as to maximize their joint performance against these criteria.

VII. AFIS PERFORMANCE RESULTS

Prior to receipt of the "test" images, the vendors were allowed to protest any of the "training" prints they felt to be of unacceptable quality. Two vendors protested no prints. One vendor protested about 50 and one vendor over 400. Protests were primarily on the basis of low minutiae counts. We also noted that a significant number of the protested prints had no clearly captured core. All prints were reviewed by the senior fingerprint examiner at the National Bureau of Investigation in Manila. He ruled that all prints were adequate for matching by human experts, so none was removed from consideration. The vendors also pointed out to us 4 matches within the "training" database, indicating collection or record keeping errors on our part. The fingerprint examiner concurred with all of these matches.

Two vendors submitted soft matching results; two submitted hard. Those returning hard results were not able to choose thresholds simultaneously satisfying the competing requirements and were found to have failed to achieve the test performance requirements. These vendors immediately resubmitted "soft" results. The remainder of this paper deals with the comparative analysis of the soft results submitted by the four vendors. For technical reasons relating to the contracting procedures of the Philippine government, the relative placement in this paper of the vendors based on technical performance does not reflect the vendor placement in the competitive contracting process.

Matching Results

Matching results were evaluated first. Only one of the four vendors (Vendor B) submitted the complete 4080 by 4128 similarity matrix in unedited form. The other vendors chose to replace low entries with zeros, possibly not computing similarity scores where bin assignments were incompatible. One vendor returned zeros for scores below an extremely high threshold, ultimately returning less than 4500 non-zero scores[21]. We

---

[20] These performance goals were intended as target system performance bounds. One vendor protested that goals 1) and 2) could not be simultaneously met under the bounding decision policies reflected in equations 18 and 22. M correct matches require no incorrect non-matches in M trys, leading to a false non-match rate of $(1\text{-FNM})^M$ which is considerably greater than the system false non-match rate given in equation 18. Equation 22 represents system false match rate under a static threshold. It was always the intent that exact decision policy, perhaps to include more interesting decision criteria, such as variable thresholds, will be determined during implementation as more precise performance data becomes available.

[21] Total number of returned non-zero scores: A=24,480; B=16,777,216;C=4445;D=111,181

computed single comparison false match and false non-match errors as a function of threshold for all vendors using the secret key.

There was significant correlation between all vendors regarding about eleven false matches. Consultation with our fingerprint expert confirmed that eight of these were in fact correct matches, indicating errors in our collection/record keeping procedure. The remaining three false matches were interestingly the false match of fingers from the correct individual. Figure 2 shows such a case of two very closely matching fingerprints, indicating the failure of our underlying assumption of independence between the fingerprints of a single individual. For this reason, we chose to disregard false matches when made from the correct individual. On this point, all vendors were effected about equally.

We have no assurance that the editing of results by vendors was done without reference to print binning. Consequently, we made the decision to divide the number of false match errors at each threshold by the number of non-zero cross-comparisons actually returned. We had hoped to report the 95% confidence bound on the returned data. This confidence bound is expressed by[22]

$$\beta = \sum_{i=0}^{K} \binom{N}{i} p^{i} (1-p)^{N-i} \tag{23}$$

where β is the confidence bound, K is the number of errors found, N is the number of comparisons and p is the upper bound on the probability. In practice, equation (23) cannot be evaluated because of the required factorial values in calculating the cumulative binomial distribution that forms the right hand side. Instead, the cumulative binomial distribution is replaced by the incomplete beta function[23,24], I_p, as

$$\alpha = 1-\beta = \sum_{i=K}^{N} \binom{N}{i} p^{i} (1-p)^{N-i} = I_p(K, N-K+1) \tag{24}$$

Inversion to determine p given β is accomplished through numerical iteration on p, a highly unstable process large N and small p (near the domain limit of I_p at p=0), particularly for small α. Because of the numerical instability in evaluating results for vendors returning more than a few thousand results, we were unable to report error bounds on p.

Results, without error bounds, are shown as Figure 3. The results of Vendor A are interesting in that they were independent of chosen threshold over a large range of threshold values. Vendor A had no false matches at any reasonable threshold, so increasing threshold had no effect on the FMR . As has been noted elsewhere[25], genuine

---

[22] As discussed in Wayman(1), ibid.

[23] M. Abramowitz and I. Stegun, eds. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, (Wiley and Sons, New York, 1972), pg. 945

[24] W.H.Press, etal, "Numerical Recipes in C", 2nd ed. (Cambridge University Press, Cambridge, 1992)

[25] Wayman(2), ibid.

distributions are often bimodal, with the second mode coincident with the single mode of the impostor distribution. The distributions of Vendor A were disjoint, except for the overlap of the second mode of the genuine distribution with the mode of the impostor distribution. Therefore, decreasing the decision threshold had no impact on the false non-match rate until the threshold was well inside the impostor distribution, thus driving the false match rate sky high. So for all reasonable values of the threshold, the number of false matches remained at zero with about 2% false non-matches. Using equation (23) in application to the results of Vendor A, who returned 24,000 match results with no false matched, we can say with 95% confidence that the true false match rate is lower than $1.3x10^{-4}$.

Binning Results

Binning results were then evaluated. All vendors submitted "hard" binning results, with one vendor submitting results of two "hard" binning proceedures. Bins assigned to "training" prints were compared, using the secret key, to bins assigned the "test" prints and inconsistencies leading to binning errors were noted. Binning error rate as a percentage of the 3267 matching pairs was calculated. The two "hard" approaches for the single vendor were evaluated as though they were independent.

One vendor submitted "soft" results from a second binning approach, in addition to "hard" results from the first. These results were evaluated using a variety of thresholds.

Penetration rate was calculated empirically from the equation

$$P = \frac{\sum^{N_2} \text{test prnts} \sum^{N_1} \text{training prnts with common bin}}{N_1 * N_2} \qquad (25)$$

where N_1 is the number of training prints and N_2 is the number of test prints.

The results of the binning test are given as Figure 4. Bin error rates vary from about three per thousand to about 50 per thousand. Penetration rates vary from about 46% to 60% based on method employed. The single soft binning method was not considered independently, but rather appended to the hard results from the same vendor. Placement closer to the lower left hand corner of the graph indicates generally better performance .

VIII. SYSTEM PERFORMANCE PREDICTION

With this data, we can predict throughput and error rates for the final system, envisioned to have at its peak, 20,000,000 cardholders enrolled with 4 fingers, with an input rate of 20,000 applicants a day. First we will calculated the filter factor using gender-based filtering. The cardholder population is expected to be about 54% female. Applying equation (6), guessing that no more than 2% of the population will be of unknown gender and that these unknowns will be equally male and female, the gender filter factor becomes

F = 0.02 + 0.54*0.53 + 0.46*0.45 = 0.51.

The four collected prints will also be classified by right and left hand forefinger and thumb. Assuming no unknowns at the time of collection and the collection of 25% of each classification, the finger filter factor becomes, by equation (5), 0.25.

By equation (2), the total system filter factor becomes

F=0.51*0.25 = 0.13

Taking as the single finger penetration rate the value of 0.50, assuming (incorrectly) that finger classifications are independent and applying equation (1), the system penetration rate for ensemble classification over the four fingers becomes

$P = 0.5^4 = 0.063$

Taking 0.01 as the single finger bin error rate and using equation (10), the ensemble bin error rate becomes $\varepsilon_{\text{bin ensemble}} = 0.04$

Taking 0.1 as the single comparison false non-match rate, assuming a search against two fingers and that bin and false non-match errors are independent, we can apply equation (18)

$$FNM_{system} = (0.04+0.1)^2 = 0.02$$

Although violation of the assumption of independence of errors will cause this value to decrease, true system false non-match rate may be driven higher by the system decision policy. We believe this computed value to represent an optimistic estimation of system false non-match rate.

Taking 0.001 as the single comparison false match rate and applying equation (22), the bound on the total system false match rate becomes

$$FMR_{system} = 0.063*0.13*80{,}000{,}000* (0.001)^4 = 6.6x10^{-7}$$

With an input of 20,000 applicants per day, the average daily false match rate will be about 0.013. This represents 13 errors in 1000 days of operation, or about 4 false matches per year, as the best performance bound. Performance will degrade if decision policies other than that reflected in equation (22) are employed and with a higher than calculated penetration rate due to bin correlations. Regardless, this performance bound is significantly better than the target of 0.001, allowing us plenty of "headroom" when setting operational decision policy.

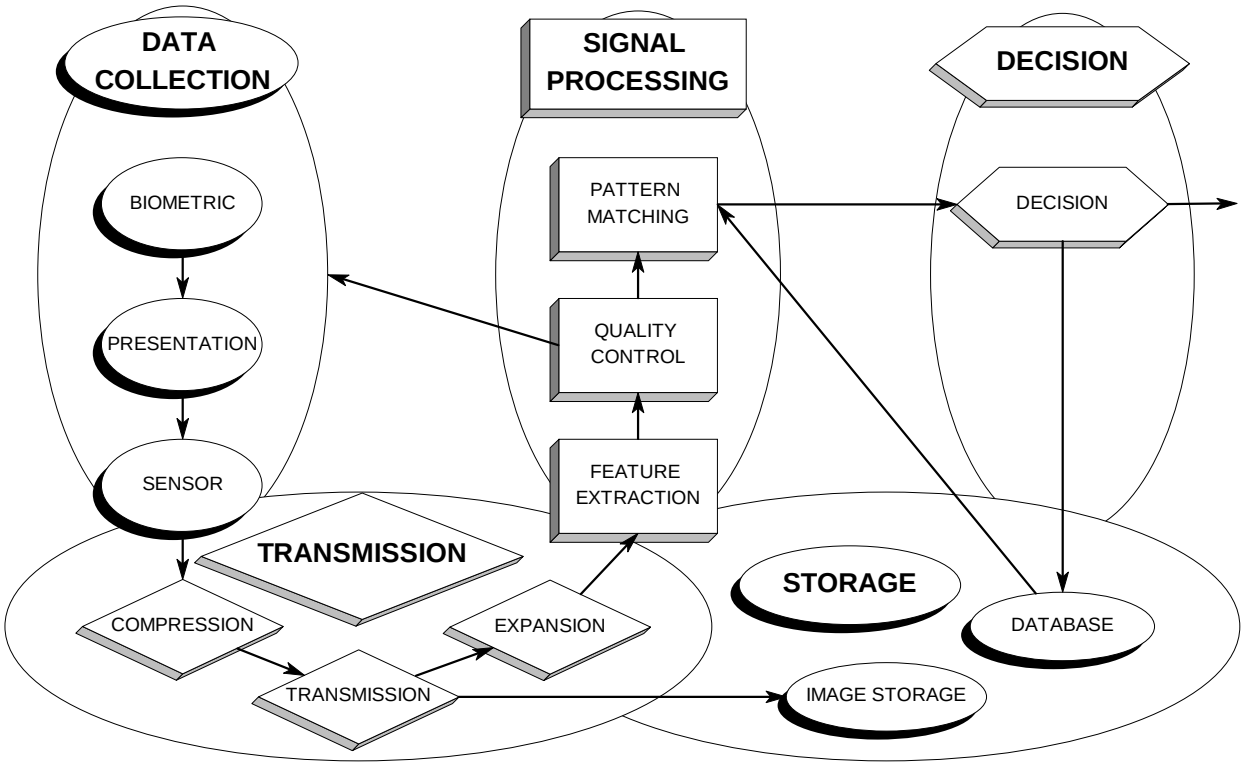
The hardware comparison rate required to support the peak throughput rate can be projected using equation (14). Assuming 80,000 operational seconds per day, the required comparison rate is

C = 20,000*0.63*0.13*80,000,000*2/80,000 = 330,000 comparisons per second

This is a lower bound on the true requirement, which will be greater owing to higher than calculated penetration rate caused by bin correlations and the additional computations required by matches.

XI. CONCLUSIONS

Legislative actions in many nations have resulted in a requirement for large-scale civilian biometric identification. In this paper, we have given an overview of civilian AFIS systems, developed mathematical performance models, and reported on experimental data collected in an international civilian AFIS benchmark test. Results demonstrate that a large-scale fingerprint identification system is feasible based on input data from multiple fingers. Systems of 20,000,000 enrolled persons, with input rates of 20,000 per day, should be able to operate within design limits of 5% false non-matches and 0.1% false matches under system decision policies tuned during implementation. A hardware processing rate of 500,000 matches per second should be sufficient for one-day processing in a four-finger system.



'FIGURE 1: THE GENERIC BIOMETRIC SYSTEM

Continuing Controversy Over the Technical Feasibility of Large-Scale Systems

James L. Wayman. Director
U.S. National Biometric Test Center

In May of 1997, IBM published a "challenge to the biometric industry" to show that large-scale identification systems (enrollment on the order of 25 million people) are technically feasible. Their concern revolved around a misunderstanding of a mathematical equation for the errors made by some biometric systems. In a series of papers and presentations, the National Biometric Test Center published the correct equations and showed that large-scale systems are indeed feasible with current technologies. Apparently, we have not been completely successful in correcting the public debate. In the last couple of months, at least two other papers and presentations have been given by scientists within the biometric industry using the mistaken equations. The purpose of these presentations has been to show the supposed advantage of their companies' specific technology in large-scale identification. Again, these scientists have misunderstood the equations for the error rates at large scale. The mathematics behind this continuing mistake is rather easy to understand, so let me take a few of lines to explain it.

If the probability of a "heads" in a coin toss is ½, then the probability of two "heads" in two independent coin tosses is $(½)^2 = ¼$. The probability of flipping three "heads" in a row will be $(½)^3 = 1/8$ and so on. The general expression will be:

Equation 1:

The probability of all "heads" in N tosses = (Probability of a single head)^N .

This equation is correct whether the coin is fair or not. If N is very big, the probability of all "heads" and no "tail" in N flips is very small, even if the coin is strongly biased in favor of heads. This, of course, is in keeping with our personal experience.

If rewritten to apply to biometric systems, the equation would read:

Equation 2:

The probability of all correct comparisons against a database of N previously enrolled individuals = (Probability of a single correct comparison)^N.

Following our previous logic, this means that the probability of all correct and no incorrect comparisons becomes very low when the database size, N , is very large, even if the probability of a single correct comparison is very high. Some scientists have argued that this equation shows that large-scale identification systems are impossible, even if the probability of a single correct comparison is very high. Where is the mistake?

The mistake is in presuming that all large-scale systems are created around the "coin toss" model. Large-scale fingerprinting systems depart from this model in two important ways: 1) they do not compare all input fingerprints to all prints already in the system; 2) they do not rely on the matching of a single measure.

Large-scale fingerprinting systems "bin" fingerprints according to pattern type, such as loop, arch or whorl. They might also "filter" the prints according to the age or gender of the person from whom they patterns came. Consequently, if a middle-aged male with loops on each finger, such as myself, applies for social service benefits, his

prints will be compared only to those from other middle-aged males with all loops. The prints will not be compared to all N prints in the database.

If a match is made on a single finger, so decision will be made until additional information, such as a second fingerprint, is compared. My application for social service benefits will only be denied if multiple fingerprints (or other pieces of information) match those of a previously enrolled recipient.

We can agree with the industry scientists that large-scale identification systems built around the coin toss model will not work. However, we understand that fingerprinting systems in social service applications are not based on this model. Consequently, we conclude that large-scale identification systems could be (and, in fact, are) possible when properly constructed.

We have developed equations giving the probability of no errors in this process as a function of database size, filtering and binning operations, and the single correct comparison probability, but they are far long to be included here. The complete paper will be sent upon request to the National Biometric Test Center.

The Philippine AFIS Benchmark Test Results

James L. Wayman
U.S. National Biometric Test Center

The Republic of the Philippines Social Security System (SSS) Identification Card Project AFIS benchmark test was conducted in May, 1997, with four international AFIS vendors, a fifth vendor withdrawing immediately prior to the test. The goals of the test were to measure penetration rate, bin error rate, and single comparison false match and false non-match rates for each vendor. We did not attempt to measure hardware comparison rate, as the conditions of the contract require minimum throughput rate performance. The benchmark was undertaken in support of a national social security card project, with an anticipated eventual enrollment of 20,000,000 cardholders and an enrollment rate of 20,000 people per day when the full enrollment is reached. The system will use gender-based filtering and four fingers (left and right forefingers and thumbs) for enrollment.

Fingerprint Image Database

To facilitate the test, we collected three sets of images: "training", "practice" and "test". Recognizing that poor quality images would lead to performance degradation not attributable directly to the algorithms, we endeavored to collect the best images reasonably attainable by our staff from the volunteers. All images were taken from SSS adult employee volunteers, each giving eight prints at each session, thumb through ring finger of each hand. The volunteers were primarily managerial and clerical workers, although some volunteer laborers were solicited, as well. Each volunteer signed a consent form authorizing release of the collected data. The collection was supervised by three SSS employees. No personal data was collected with the prints other than gender; 55% of the volunteers were women. Database bookkeeping was accomplished by assigning each volunteer a collection number. Handwritten data sheets connecting volunteers with collection numbers was maintained by the supervising employees and have since been destroyed.

Prints were imaged with an Identicator DF-90 "flat" scanner, believed to be "Appendix G" compliant and an "MRT" frame grabber in a lap-top computer. Front-end quality control software from Identicator was employed. Database software was custom supplied by Identicator for this project. The prints were stored, using loss-less compression, as "TIFF" images. Some image quality loss, attributable to frame grabber noise, was noticed in the upper right hand quadrant of each image.

The "training" data consisted of 4080 prints taken from 510 volunteer employees of the Social Security System over a three-week period. It was our original intent that the "training" data set be "clean", meaning free from duplicate images. Subsequent analysis by the vendors indicated that there were, in fact, 4 repeated prints in the set made by inadvertently inaccurate finger presentations by volunteers which were not corrected by the supervisors.

In collecting the 4080 "training" images, we were allowed to physically touch the volunteers, manipulating their fingers on the scanner and applying slight pressure with the intent of obtaining the highest print quality possible as judged by the quality control software. Moisturizing compound was applied as needed. In general, three or more

prints were collected from each finger of each volunteer, although the training data set consisted of only one image from each collected finger.

A second, "test", data set of 4128 images was collected from 506 volunteers, 409 of whom were included in the training set Although image quality was checked and moisturizing compound applied as needed, somewhat less care was generally taken to provide high quality images. Collection of the "test" set commenced one week after completion of the "training" set collection and was finished within three weeks. Consequently, individual volunteers were imaged at an interval of one to six weeks for the test and training sets. The test set contained 10 duplicate ensembles (80 prints) imaged from 10 volunteers in a separate session several weeks after the completion of the collection of the other images.

The order of the files and the file names were scrambled to prevent a determination of correlation between "test" prints and "training" prints or correlations within the "test" set, and a highly-secret key was created linking the "test" and "training" prints.

The third, "practice" set of 80 images was taken from 10 volunteers whose prints were in the "training" data set. The file names given these images were adjusted to identify them with their matching image files in the "training" set.

Both the "training" and "practice" data sets were mailed to the vendors several weeks prior to the tests. The "test" data set was hand delivered to the AFIS vendors on the day of the benchmark test. In all cases, testing was completed within the day.

Test Requirements

Prior to receipt of the "test" images, the vendors were required to supply the binning results for the "training" images. Vendors were allowed to report either "hard" or "soft" binning results, provided that enough information was supplied for analysis. By "hard", we mean the assignment of each print to one or more discrete bins. By "soft", we mean the assignment of numerical values to each print representing in some way a probabilistic binning assignment.

Vendors were required to report the binning assignments of each print of the "test" set using the same format as used in reporting bins of the "training" prints. Then vendors were required to match the "test" to the "training" prints, again reporting either "hard" or "soft" results. A "hard" result was a "match" determination between files. A "soft" result was a numerical similarity measure between files. Vendors choosing to return "soft" results were required to submit a 4080 by 4128 matrix of similarity measures.

AFIS Performance Results

Matching results were evaluated first. Only one of the four vendors (Vendor B) submitted the complete 4080 by 4128 similarity matrix in unedited form. The other vendors chose to replace low entries with zeros, possibly not computing similarity scores where bin assignments were incompatible. One vendor returned zeros for scores below an extremely high threshold, ultimately returning less than 4500 non-zero scores[1]. We

---

[1] Total number of returned non-zero scores: A=24,480; B=16,777,216;C=4445;D=111,181

computed single comparison false match and false non-match errors as a function of threshold for all vendors using the secret key.

There was significant correlation between all vendors regarding about eleven false matches. Consultation with our fingerprint expert confirmed that eight of these were in fact correct matches, indicating errors in our collection/record keeping procedure. The remaining three false matches were interestingly the false match of fingers from the correct individual, indicating the failure of our underlying assumption of independence between the fingerprints of a single individual. For this reason, we chose to disregard false matches when made from the correct individual. On this point, all vendors were effected about equally.

We have no assurance that the editing of results by vendors was done without reference to print binning. Consequently, we made the decision to divide the number of false match errors at each threshold by the number of non-zero cross-comparisons actually returned.

Results are shown as Figure 1. The results of Vendor A are interesting in that they were independent of chosen threshold over a large range of threshold values. Vendor A had no false matches at any reasonable threshold, so increasing threshold had no effect on the FMR . Genuine distributions are often bimodal, with the second mode coincident with the single mode of the impostor distribution. The distributions of Vendor A were disjoint, except for the overlap of the second mode of the genuine distribution with the mode of the impostor distribution. Therefore, decreasing the decision threshold had no impact on the false non-match rate until the threshold was well inside the impostor distribution, thus driving the false match rate sky high. So for all reasonable values of the threshold, the number of false matches remained at zero with about 2% false non-matches. It can be stated with 95% statistical confidence that the false match rate of vendor A was under 0.01%. It might be that the false match rate is even lower, but lack of returned match scores prevent us from making that determination. Vendor B returned 16,000,000 cross comparisons with only one false match, indicating a 95% statistical confidence of a false match rate of under 3 in 10 million ($3x10^{-7}$), but with a false non-match rate approaching 20%.

Binning results were also evaluated. All vendors submitted "hard" binning results, with one vendor submitting results of two "hard" binning procedures. Bins assigned to "training" prints were compared, using the secret key, to bins assigned the "test" prints and inconsistencies leading to binning errors were noted. Binning error rate as a percentage of the 3267 matching pairs was calculated. The two "hard" approaches for the single vendor were evaluated as though they were independent.

One vendor submitted "soft" results from a second binning approach, in addition to "hard" results from the first. These results were evaluated using a variety of thresholds.

Penetration rate was calculated empirically from the equation

$$P = \frac{\sum^{N_2} \text{test prnts} \sum^{N_1} \text{training prnts with common bin}}{N_1 * N_2} \qquad (1)$$

where N_1 is the number of training prints and N_2 is the number of test prints.

The results of the binning test are given as Figure 2. Bin error rates vary from about three per thousand to about 50 per thousand. Penetration rates vary from about 46% to 60% based on method employed. The single soft binning method was not considered independently, but rather appended to the hard results from the same vendor. Placement closer to the lower left hand corner of the graph indicates generally better performance .

The results of the binning test are given as Figure 2. Bin error rates vary from about three per thousand to about 50 per thousand. Penetration rates vary from about 46% to 60% based on method employed. The single soft binning method was not considered independently, but rather appended to the hard results from the same vendor. Placement closer to the lower left hand corner of the graph indicates generally better performance.

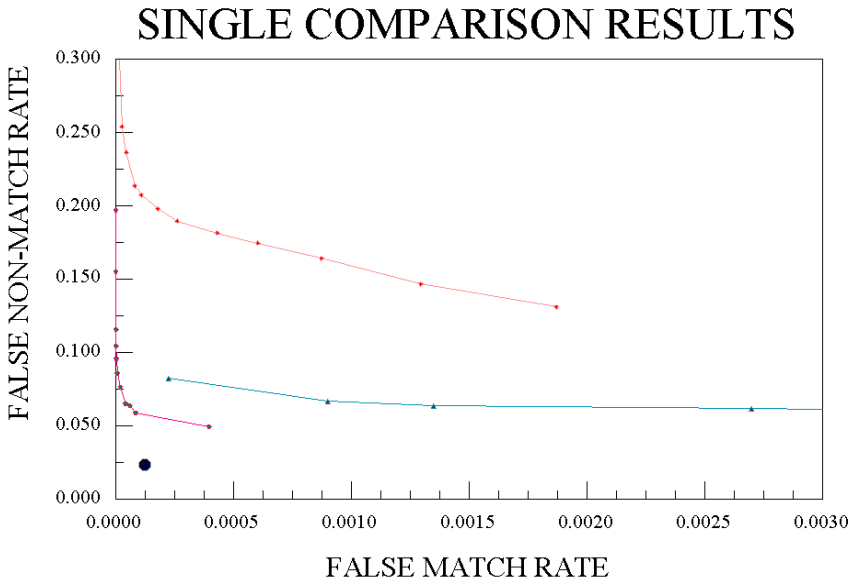


FIGURE 1

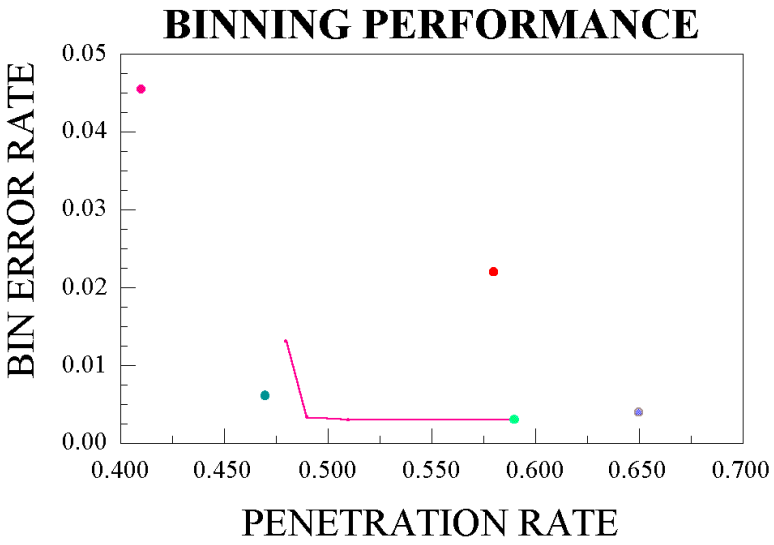


FIGURE 2

Performance Prediction

With this data, we can approximate throughput and error rates for a large-scale, centralized fingerprint system, assumed to contain records of 8.5 million people. Full mathematical development will appear in "Error Rate Equations for the General Biometric System", Automation and Robotics Magazine, Special Issue on Automatic Identification, scheduled for publication in January, 1999. Based on the Philippine test, we'll assume that a vendor can maintain a single finger false match rate of under one in one million while keeping the single finger false non-match rate below 10%. Further, let's assume a single finger penetration rate of 50% with an associated bin error rate of 1% and use gender-based filtering

We will first show that a single finger system cannot meet the reasonable functional requirements. A searched print can be falsely not matched because of a binning error or because of a single comparison false non-match. The probability that a single searched print has neither can be expressed as

$$1-\mathrm{FNM}_{sf} = (1-\varepsilon_{\mathrm{ensemble}})(1-\mathrm{FNMR}) \qquad (2)$$

where FNM_{sf} is the probability that the system falsely does not match a single searched finger, $\varepsilon_{\text{bin ensemble}}$ is the probability of a binning error over the ensemble, and FNMR is the single comparison false non-match rate. The explicit dependency of FNMR and FNM_{sf} on threshold, τ , has been dropped for notational simplicity. Providing that both FNMR and $\varepsilon_{\text{bin ensemble}}$ are small, equation (2) can be rewritten as

$$\mathrm{FNM}_{sf} = \varepsilon_{\text{bin ensemble}} + \mathrm{FNMR} \qquad (3)$$

In a single print system, the ensemble bin error is just the single finger bin error rate. Therefore, in our hypothetical single finger AFIS system, the system false non-match rate is 1% + 10% = 11%. In the verification mode, however, the probability of a false match is simply the false match rate, which in this case is 10%.

The probability that a single searched print is not matched against any print in the searched database is

$$1-\mathrm{FM}_{sf} = (1-\mathrm{FMR})^{P_{sys}*N} \tag{4}$$

where FM_{sf} is the system single comparison false match rate, FMR is the single comparison false match rate, P_{sys} is the penetration rate, and N is the size of the database. Again, explicit dependency of FM_{sf} and FMR on threshold, τ, has been dropped for notational simplicity.

Equation (4) can be rewritten as

$$\mathrm{FM}_{sf} = 1-(1-\mathrm{FMR})^{P*N} \tag{5}$$

In a single print system with gender-based binning, the system penetration rate is simply the single finger penetration rate of 50% times the gender-based binning of 50%, which in this case yields 25%. With a false match rate assumed to be 10^{-6} and N at 8.5 million, the system false match rate becomes, by (5) above, about 88%. Even worse, the expected number of false matches per search can be given by

$$E[\mathrm{FM}] = P_{sys} * N * \mathrm{FMR} \tag{6}$$

which computes to about 8 in this example. Clearly, a single finger system cannot meet the functional requirements.

Now we will show that a two-finger system can meet reasonable functional requirements. With two fingers, we can use ensemble binning, which will give us a penetration rate of 0.5*0.5=0.25. Multiplying by the gender-based filtering, we will have a total system penetration rate of 0.125. The ensemble bin error rate is approximately twice the single bin error rate, or 2% in our example. If we use a two-finger search with a policy declaring a match if either finger is found to match a record in the database, then no system false non-match requires no ensemble binning error and no false non-match on both fingers.

Under the assumption of error independence, the probability of that happening can be given by

$$1-\mathrm{FNM}_{sys} = (1-\varepsilon_{\text{bin ensemble}})(1-\mathrm{FNMR}^2) \tag{7}$$

which in our case yields a system false non-match rate of less than 3% in the recognition mode. For verification, the availability of two fingers implies a false non-match rate of $FNMR^2$, computing to 1%. It is noted that the false non-match rate is independent of database size, N.

If the bounding system match policy requires a match on both fingers for a system match to be declared, the probability of a system false match, FM_{system} ,occurring is

$$\mathrm{FM}_{sys} \approx P_{sys} * N * \mathrm{FMR}^2 \tag{8}$$

In our case, the recognition false match rate computes to $0.125*8.5*10^{6} *10^{-6} *10^{-6} = 1*10^{-6}$, or one false match in a million searches.

Equations (7) and (8) represent bounding approximations only, as neither accounts for the case of one match and one non-match against a fingerprint set in the database. A consistent set of equations is given in the “Error Rate Equations for the General Biometric System” previously referenced.

The throughput rate, S, of the system can be calculated by

$$S = \frac{C}{m * P_{ensemble} * N} \tag{9}$$

where C is the hardware comparison rate, and m is the number of fingers used. We can use (9) to calculate the required hardware comparison for a given throughput rate.

Suppose that a throughput rate of 2.5 million persons per year is projected for our 8.5 million person system. Assuming 80,000 operational seconds per (22 hour) day and 250 operational days per year, the required comparison rate is C = $2.5*10^6/250/80{,}000*0.125*2.5*10^6$ *2= 78,000 comparisons per second

This is a lower bound on the true requirement, which will be greater owing to higher than calculated penetration rate caused by bin correlations and the additional computations required by matches. Nonetheless, this value is well within the capabilities of current hardware systems. Using a current "rule of thumb" of many dollars per match per second, central processing hardware costs would approximate a million dollars for the system.

Conclusions

This report has documented a scientific test of the world's leading AFIS vendors and has demonstrated, using these test results, the feasibility of an 8.5 million person, 2-finger AFIS system. More complete mathematical development of the system equations will appear in upcoming publications.

Philippine Social Security System Inaugurates Huge Civilian ID Card/AFIS System

James L. Wayman, Director
U.S. National Test Center

On Tuesday, Nov. 17, 1998, the Philippine Social Security System (SSS) officially inaugurated its long awaited SSS ID card system. After an invocation and blessing by the priest of the SSS parish, site tours of the 9,000 ft^2 facility were given to over 100 invited international guests. In his keynote address at the reception following, Dr. W.G. Padolina, Secretary of the Department of Science and Technology, said (*full text of Sec. Padolina's remarks follows in the issue –ed*):

"It is well to note that the Estrada administration has placed highest priority towards poverty alleviation. However, we note that projects, such as the one we are inaugurating today, are sometimes intimidating especially to the poor. High tech projects are not for the poor, so we say… .This is where I must disagree, and I must say that projects, such as the new SSS ID, because it improves efficiency of governance, will be a contribution to the process to free many of our countrymen from the bondage of poverty…. I consider this biometrics project a vital cog in the poverty alleviation program of the Estrada administration. This intervention will definitely make available resources to those who will need them most because fraud will be minimized."

Over 250,000 SSS members have been enrolled over the last 6 months, even before installation of the card printing and mailing equipment was complete. Issuance of ID cards is expected to exceed 4.2 million by the end of 1999, making this system the world's largest civilian AFIS-based card system. Over 35 million members, beneficiaries and dependents will be enrolled by the end of 2004. "The purpose of the system is multi-fold", said May C.Ciriaco, Vice President of the SSS. "Our primary goals are to clean up our own database of members, pensioners and dependents, to make service delivery more efficient, and to eliminate fraud and abuse." After a failed bid process in early 1996, the SSS invested approximately US$300k in the following 18 month period to put together a second "Request for Proposal" (RFP) and evaluate responses. The resulting RFP is now on the SSS web page at www.sss.gov.ph. The contractual award, worth approximately US$40M, was made in January, 1998, to the prime contractor, Ayala Systems Technology, Inc.

The ID Card and the Mission of the SSS

The SSS is a Congressionally-chartered government financial institution with its own budget independent of the Philippine treasury. SSS assets, which currently exceed US$4B, are used to provide member loans and loans for housing and business development. Although the primary mission of the SSS is to provide social insurance to members and dependents in the form of retirement, maternity, unemployment, disability and death benefits, the institution has always taken seriously its secondary role of promoting economic development in the Philippines. Controversy over a government-backed national ID project, unrelated to the SSS ID, erupted in early 1997 when the Philippine Supreme Court ordered the National Statistical Office to cease using the Personal Reference Number to link personal data on citizens. The Supreme Court has

since ruled that Congress does not have the constitutional authority to order creation of a national ID card. (*See John Woodward's article in BHSUG Newsletter #10 – ed.).* The SSS is quick to point out that the ID Card project is not a national identity card and that cards will ultimately be issued to less than 60% of the 60 million Filipino population. As only 8% hold a driver's license, the SSS ID card may be the only photo ID possessed by most citizens.

Card Issuance Procedures

Cards are currently being issued only to active members and pensioners. Enrollment of dependents will commence in about 2002, after all 20M active members are enrolled. Members are generally enrolled through their employers, either at the employer's site using mobile equipment or during appointments at the SSS offices as arranged by the employers. Early demand for the cards has been enthusiastic, leading to long queues at the employer sites and the 45 SSS offices currently enrolling members. Card enrollment at all 200 offices will commence by January of 1999. Additionally, 20 mobile data collection units have been purchased for use at major places of employment and in remote areas not serviced by a local office. Card issuance at Philippine embassies for oversees workers is also being discussed. Cards are distributed by mail to the employers, who transfer the cards to the individual members. Cards of the self-employed are sent via registered mail.

Reduction in the Incidences of Multiple Identities

The database of the new system is fully integrated with the previous member database. One of the immediate benefits accrued by the new system is in pointing out errors in this previous data, such as the assignment of multiple or unused names to the same individual (often related to married/maiden name confusion), or errors in other information, such as addresses and birth dates. There have already been many attempts by single individuals to receive multiple cards. The presumption is that most, if not all, of these cases represent either bureaucratic error or a legitimate misunderstanding of the "one member, one ID number" policy. Ms. Ciriaco stated, "With regard to multiple identity claims by the same individual, our goal has always been reduction through discovery and deterrence, not prosecution. No member will be prosecuted simply because the computer indicates a duplicate record in the database." All cases of duplicate records will be investigated on a case-by-case basis by the Anti-Fraud Office (AFO) using the same procedures in place under the previous system. Ms. Ciriaco indicated that the SSS will continue exactly as before in seeking tough criminal penalties against deliberate fraudsters uncovered by the AFO investigators. "We have a strong obligation to our members to protect their assets against the unscrupulous", she stated.

Technical System Performance

The detailed technical design and specification of the system was made by the SSS during the RFP development process. The Automatic Fingerprint Identification System (AFIS) component of the ID card requires collection of both index and both thumb prints of each enrollee. An initial search is made on the two index fingers. With the current small database size, thumb prints are compared by human examiners when matches are found on one or more of the index fingers. As database size grows, thumb print comparisons may be routinely incorporated into the search protocol. Although

thumb prints are not currently searched by the system, their classifications are used to partition the database, thereby decreasing penetration rate and the required computer hardware speed. Testing during the vendor benchmarking phase of the proposal evaluation process revealed a single-finger penetration rate of slightly under 50%. (*See Jim Wayman's article in BHSUG Newsletter #9 –ed*) If the fingerprint patterns over the four fingers were uncorrelated, the four finger penetration rate would be about $(0.5)^4$ = 6%. To account for the anticipated classification correlation across fingers, the system was designed around an estimated penetration rate of 12%. Early operational data supports this estimation. Additional filtering schemes (if any) have not been publicly disclosed for security reasons.

During the early years of the project, demand on the computer matchers will be low, owing to small database size. As the database grows, additional computer hardware modules will be added. To minimize hardware purchases, matchers will be run on a 22-hour a day schedule, 7 days a week. Card printing and mailing hardware has sufficient capacity to allow for 16-hour, 5-day operation when the operational goal of 25, 000 cards per day is reached sometime this year.

As required in the RFP, each enrollment station performs a quality control check on the fingerprints. Each enrollee is allowed 3 attempts to give a readable fingerprint in the supervised setting. No one is denied enrollment for lack of readable prints so, in difficult cases, the third attempt accepted regardless of quality. Two quality control stations have been set up at the Card Processing Facility in Manila for auditing (and correcting, in some cases) print quality. A preliminary examination of both index fingerprints of the first 113, 000 enrollees shows the percentage of poor quality prints to be about 2.5%, as defined by the quality-control module. The percentage of enrollees having poor quality prints on both index fingers is somewhat over 1%, which indicates a substantial correlation between print quality of the two fingers over all enrollees. No attempt is being made to deconvolve poor physiology from poor presentation, so the correlation may strongly reflect presentation errors, such as movement or excess pressure. Substantial system feedback on print quality to the enrollment station supervisors throughout the country is planned. The system was designed by SSS to have a false acceptance (non-match) rate of less than 5%, meaning that over 95% of all fraudsters will be caught. The false rejection rate, requiring manual record examination by the AFO team, will be kept to under 0.1% of all applicants by periodic adjustments to the system decision policy as the database size increases.

The SSS management has indicated a strong willingness to share experiences, operational statistics, and performance audit data with other countries. Email inquiries can be sent to sssemail@info.com.ph.

The "Penetration Rate" in Automatic Fingerprint Identification Systems

Kang James and Barry James
Department of Mathematics and Statistics
University of Minnesota at Duluth

Suppose that there are k categories ("bins") into which fingerprints are classified. If each print has only one classification, with p_i the proportion of prints in each bin ($\sum_{i=1}^{k} p_i = 1$), then each incoming print needs to be compared with at most a fraction p_i of all prints (those in its category). Thus a randomly chosen print needs to be compared with (at most) an expected fraction $\sum_{i=1}^{k} p_i^2$ of all prints. This number is the "penetration rate".

If prints can be placed into more than one category, because of ambiguities in the classification scheme, then we can have $\sum_{i=1}^{k} p_i > 1$. In this case, it is not necessary to check on average the $\sum_{i=1}^{k} p_i^2$ of all prints, *if* we assume that when the sampleprint is compared with print x in bin I, it does not need to be compared with print x again if print x also belongs to some bin $j \neq i$.

Question

What other information is needed to produce an analytic expression for the penetration rate? Here we interpret "penetration rate" to be the expected fraction of ll prints with which a sample print needs to be compared.

A Simple Answer

If the fraction of prints in each *combination* of bins is known, the formula extends easily. Let

p_i = proportion of prints in bin i;
$p_{i\,j}$ = proportion of prints in both bin i and bin j;
$p_{i\,j\,l}$ = proportion of prints in bins i, j and k ;
etc. ($p_{1\,2\,3\,\ldots k}$ = the proportion of prints in all bins).

With A_i the event "print is in the bin i", where we are thinking of a randomly chosen print, we have $P(A_i) = p_i$, $P(A_i \cap A_j) = p_{i\,j}$, etc. Before giving the formula for the penetration rate, consider the simple example of 2 bins.

Example

Suppose there are only two categories, say "whorl" and "not whorl". If a fingerprint is put in both bins, we take this to mean not that it is of both types, but that we cannot tell to which of the two types it belongs. Thus, if a subject print is classified in both bins, it needs to be compared to all prints in bin 1 or bin 2 (in this case, all prints).

A print classified in one bin and not the other need only be compared to the prints in its respective vin. Therefore, the penetration rate is

$$1\cdot p_{12} + p_1(p_1 - p_{12}) + p_2(p_2 - p_{12}) = p_1^2 + p_2^2 + p_{12}(1 - p_1 - p_2) = p_1^2 + p_2^2 - p_{12}^2$$

Theorem

The penetration rate is

$$\sum_i p_i^2 - \sum_{i<j} p_{ij}^2 + \sum_{i<j<l} p_{ijl}^2 - \ldots + (-1)^{k-1} p_{123\cdots k}^2$$

Proof. Since $\bigcup A_i = \Omega$ = a sure event, we have

$$1 = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<l} P(A_i \cap A_j \cap A_l) - \ldots + (-1)^{k-1} P(A_i \cap \cdots \cap A_k)$$

$$= \sum_i p_i - \sum_{i<j} p_{ij} + \sum_{i<j<l} p_{ijl} - \ldots + (-1)^{k-1} p_{123\cdots k}$$

The proof essentially duplicates the proof of the equation above, obtained from the inclusion-exclusion principle.

If we compare a print with all prints in bin *i* whenever it is classified in bin *i*, the expected proportion of comparisons is $\sum_i p_i^2$. This is too high, since a print classified in both bin *i* and bin *j* will be compared twice with those prints in both bins. To compensate, we subtract $\sum_{i<j} p_{ij}^2$. This has the effect of discounting the second comparison of a print classified in both bin *i* and bin *j*, for all pairs *i* and *j*. But then a print classified in 3 bins will find that all 3 of its comparisons with those prints in all 3 bins will have been discounted. So $\sum_i p_i^2 - \sum_{i<j} p_{ij}^2$ is too low. To compensate, add $\sum_{i<j<l} p_{ijl}^2$. Continue in this way, alternating signs.

Remarks

If obtaining the probabilities of p_{ij}, p_{ijl}, etc. is not feasible, the problem becomes one of estimating or approximating them.

Sample of the k Largest Order Statistics

Kang James and Barry James
Department of Mathematics and Statistics
University of Minnesota at Duluth

Problem

Suppose we have n independent samples of size m from the cumulative distribution F. Then we have a total of nm i.i.d. random variables $X_{ij}, 1 \le i \le n, 1 \le j \le m$, each with a cumulative distribution function (CDF) of F(x); If we are given the top k order statistics from each sample, what can we say about the underlying distribution F?

Comments

It depends on the relationship of k, m, and n. Intuitively, if we are to learn more about F than just the upper k/m fraction of the distribution, the number n of samples will need to be large in relationship to the sample size m. If m and n are of the same order and k is small , not much can be said about F below its 1-k/m quantile. We will consider some cases below: in each, we will assume that $m \to \infty$ and $n \to \infty$.

Case 1

Assume that k is fixed and $m \to \infty$ and $n \to \infty$ in such a way that $\log(n) = o(m)$ .Then, essentially, nothing is known about the distribution.

Proof. Let $U_1, \ldots, U_n$ be a random sample from UNIF (0,1), the uniform distribution on [0,1]. If $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ are the order statistics, then a theorem of Kiefer (Shorack and Wellner (1986), pg. 407-8) says that for L fixed

$$\limsup_{n\to\infty} \frac{\log(1/nU_{(1)})}{\log\log n} = \frac{1}{L} \text{ a.s.}$$

Among other things, this implies that for any $\varepsilon > 0$,

$$(*)\ P\left(U_{(L)} > \frac{1}{n(\log n)^{(1+\varepsilon)/L}} \right) \to 1 \text{ as } n \to \infty$$

If $X_{1,(m)}, \ldots X_{n,(m)}$ are the maxima of the n samples (i.e. $X_{1,(m)} = \max(X_{i,1}, \ldots X_{i,m})$), they form a random sample of size n from $F^m(x)$ and $F^m(X_{1,(1)}), \ldots, F^m(X_{n,(1)})$ are the order statistics of a random sample from UNIF (0,1). If $X_{(1)}$ is the *minimum* of these maxima ($X_1 = \max(X_{i,(m)}, \ldots X_{n,m})$), then we have (set L=1 in $(*)$)

$$P\left(F^m(X_{(1)}) > \frac{1}{n(\log n)^{1+\varepsilon}} \right) \to 1 \text{ as } n \to \infty \text{ (and } m \to \infty),$$

or equivalently

$$(**)\ P\left(F(X_{(1)}) > \frac{1}{n^{1/m}(\log n)^{(1+\varepsilon)/m}} \right) \to 1$$

The assumptions on n and m imply that $n^{1/m} \to 1$ (take logarithms to see this) and $(\log n)^{(1+\varepsilon)/m} \to 1$. This means that

$$F(X_{(1)}) \xrightarrow{P} 1 \text{ as } m \to \infty \text{ and } n \to \infty .$$

Therefore, if k=1, so that we know the maximum of each sample, we know essentially nothing about the distribution (the smallest maximum is off in the right tail of the distribution).

The behavior of any fixed upper order statistic is similar. To illustrate the process, look at the second largest order statistic $X_{i,(m-1)}$ (we're thinking of k=2 here), which form a random sample of size n from the distribution

$$G(x) = P(X_{i,(m-1)} \le x) = mF^{m-1}(x) - (m-1)F^{m}(x).$$

If $Y_{(1)}$ is the minimum of the $X_{i,,(m-1)}$, (*) implies (again L=1)

$$P\left(mF^{m-1}(Y_{(1)}) - (m-1)F^{m-1}(Y_{(1)}) > \frac{1}{n(\log n)^{1+\varepsilon}} \right) \to 1 \text{ as } n \to \infty \text{ and } m \to \infty ,$$

That is

$$P\left(mF^{1/(m-1)}F(Y_{(1)})\left(1 - \frac{m-1}{m}F(Y_{(1)})\right)^{1/(m-1)} > \frac{1}{n^{1/(m-1)}(\log n)^{(1+\varepsilon)/(m-1)}} \right) \to 1$$

Since $m^{1/(m-1)} \to 1, n^{1/(m-1)} \to 1$, $\log(m)^{(1+\varepsilon)/(m-1)} \to 1$, and $1 - \frac{m-1}{m}F(Y_{(1)}) < 1$, we conclude that $F(Y_{(1)}) \xrightarrow{P} 1$ as $n \to \infty$ and $m \to \infty$. This means that the smallest second-largest order statistic is still too large to effectively tell us anything about the distribution.

Case 2

Assume that k is fixed, and $m \to \infty$ and $n \to \infty$ in such a way that log(n) = O(m) but not o(m). To fix ideas, suppose that $n \sim be^{cm}$ as $m \to \infty$, for some constants b>0 and c>0. Then we can estimate the upper $1 - e^{-c}$ fraction of the distribution F ; i.e., from the e^{-c}-quantile and above (and we cannot estimate F below its e^{-c}-quantile).

Proof. Let x_t be the t-quantile of F ($F(x_t)=t$). Let N_t be the number of n maxima which are less than or equal to t:

$$N_t = \#\{i : X_{i,(m)} \le x_t\}.$$

N_t has a binomial distribution, $N_t \sim BIN(n, F^m(x_t)) = BIN(n, t^m)$. If $n \sim be^{cm}$, then $E(N_t) = nt^m \sim b(te^c)^m$, so that

$$E(N_t) \to \infty \text{ if } t > e^{-c}$$

$$\rightarrow b \text{ if } t = e^{-c}$$

$$\rightarrow 0 \text{ if } t < e^{-c}$$

Therefore, the sample maxima can be used to estimate F above its e^{-c}-quantile.

A similar result holds for any fixed k. If, for example, k = 2, consider the sample of second-largest order statistics $X_{i,(m-1)}$. Since

$$G(x_t) = mF^{m-1}(x_t) - (m-1)F^m(x_t) = mt^{m-1} - (m-1)t^m,$$

the expected number of second-largest order statistics less than or equal to x_t (call this number M_t) is

$$E(M_t) \rightarrow n[mt^{m-1} - (m-1)t^m] \sim mbe^{cm}t^{m-1}(1-t)$$

$$\rightarrow \infty \text{ if } t \geq e^{-c}$$

$$\rightarrow 0 \text{ if } t < e^{-c}$$

Therefore, the top two order statistics can be used to estimate F above its e^{-c}-quantile.

Conclusion

For fixed k, one needs log(n) to be at least the order of m to be able to say much about the distribution of F.

Reference

Shorack and Wellner (1986) Empirical Processes with Applications to Statistics. John Wiley and Sons

Multi-Finger Penetration Rate and ROC Variability for Automatic Fingerprint Identification Systems

James L. Wayman, Director
U.S. National Biometric Test Center

1. Introduction

In previous papers [1-7], we considered performance estimation of biometric identification systems based on assumptions of measurement independence between measures. We noted in those papers that such assumptions are generally incorrect, but lacking any data on measure correlations, no quantitative estimates of the effect on system performance were offered. Although measurement correlations effect error rates and throughput of all biometric systems, it is the performance of large-scale identification systems that is most critically effected by data correlations because of the large number of measurement comparisons generally made.

Currently operational, large-scale biometric identification is restricted to Automatic Fingerprint Identification Systems (AFIS). In this paper, we will estimate various measure correlations for AFIS from new fingerprint test data. The multi-finger test data is available for both false match/false non-match comparison errors and binning error/penetration rate estimation. Specifically, in this paper we will estimate penetration rates for single finger systems based on thumb, index, middle and ring fingers, and multi-finger systems for two thumb, two index finger and combined four thumb-index finger systems. Penetration rates calculated from test data are compared to theoretical calculations based on recent finger-dependent pattern classification statistics from the FBI [8].

We will show Receiver Operating Characteristic (ROC) curves computed with non-matching comparisons differentiated between fingers in communicating and non-communicating bins. Further, we will develop different ROC curves for thumb, index, middle and ring fingers of right and left hands. Finally, the variability of the "impostor" distribution across test samples will be discussed.

2. Test Data

The electronically "live" scanned Philippine fingerprint test data base [3] was used in this test. The data consisted of two sets, enrollment or "training", and "test" data. The training set, consisting of 4080 distinct fingerprints, was taken from 510 individual adult volunteers, each giving eight fingerprints (thumb through ring fingers on both hands). All volunteers were employees of the Social Security System of the Republic of the Philippines. Most were office and administrative workers and 55% were women. The test set of 4128 prints was collected one to six weeks after the training set from 506 individual volunteers. Of these 506 volunteers, 409 were common to both test and training data sets. Ten volunteers in the test set donated two sets of 8 prints each. 97 volunteers in the training set were not represented in the test set.

A third "practice" set of 80 images from 10 volunteers, whose images were in both test and training sets, was taken 6 weeks after the test database was completed.

Prints were imaged with an Identicator DF-90 "flat" scanner, believed to be "Appendix G" compliant and an "MRT" frame grabber in a lap-top computer. Front-end quality control software from Identicator was employed. The Identicator "Biometric Enrollment System" collection and database management software was used for this

project. The prints were stored, using loss-less compression, as "TIFF" images. Some image quality loss, attributable to frame-grabber noise during collection, was noticed in the upper right hand quadrant of most images.

3. Vendor Testing

To date, six AFIS vendors have had their algorithms evaluated against this data. The current test procedure is to send any requesting vendor training, test and practice data sets. The ordering of the test data image files has been randomly scrambled, but the practice images are clearly linked to their corresponding training set images. These practice images allow the vendors to tune any internal parameters required by our data quality or format. Any vendor can request testing of matching and/or binning algorithms.

For the matching test, the vendor returns a 4128x4080 matrix of comparison scores for all test prints compared to all training prints. For the binning test, the vendor returns the bin assignments for all test and training prints, and the rules by which bins are determined to be "communicating" or "non-communicating". In large-scale AFIS system, prints in "communicating" bins are similar enough that they must be compared for possible matching. Upon receipt of all of this data, we release to the vendor the "key" linking the test and training sets.

In this analysis, we used the score matrix from the "best" matching vendor tested to date, meaning the score matrix that produced the generally lowest ROC. We used the binning results from the "best" binning vendor tested to date, meaning the data that we judged presented the best trade-off between penetration and bin error rates. Binning and matching data used here was not from the same vendor. Precise matching values and binning assignments are not discussed here to protect the identity of the vendors.

4. Finger Dependency of Penetration Rate

It is well known that print classification statistics are finger-dependent. Table 1 shows classification statistics by finger from recent FBI data [8].

TABLE 1: SINGLE FINGER CLASSIFICATION STATISTICS

Pattern Type	Finger Position										
	1	2	3	4	5	6	7	8	9	10	Ave
Arch	3.01%	6.09%	4.43%	1.24%	0.86%	5.19%	6.29%	5.88%	1.78%	1.15%	3.59%
Tented Arch	0.40%	7.72%	3.20%	1.03%	0.72%	0.58%	7.96%	4.53%	1.45%	1.10%	2.87%
Right Loop	51.26%	36.41%	73.38%	51.20%	83.03%	0.63%	16.48%	1.66%	0.51%	0.12%	31.47%
Left Loop	0.46%	16.96%	1.47%	1.10%	0.26%	58.44%	39.00%	70.30%	61.47%	86.11%	33.56%
Whorl	44.77%	32.45%	17.21%	45.24%	14.96%	35.04%	29.93%	17.30%	34.57%	11.33%	28.28%
Scar	0.03%	0.17%	0.13%	0.06%	0.06%	0.04%	0.14%	0.12%	0.06%	0.06%	0.09%
Amp	0.07%	0.20%	0.18%	0.14%	0.12%	0.09%	0.20%	0.20%	0.16%	0.13%	0.15%

When each print can be classified only into a single bin, the equation for calculating penetration rate from classification statistics is given in [1] as

$$Pn = p_K + \sum_{i=1}^{K-1} (p_i + p_K) p_i \qquad (1)$$

where Pn is the penetration rate, p_i is the probability that the print is of the i^{th} classification and the k^{th} classification is considered as "unknown". This equation was applied to the data of Table 1. Scarred fingers where considered of "unknown" classification and the data was re-normalized after removal of the amputated finger statistics. Table 2 shows the resulting penetration rates for this approach when fingers in each position are compared to corresponding fingers, right to right, left to left, right to left (or left to right), or all to all.

TABLE 2: SINGLE FINGER PENETRATION RATES FROM FBI STATISTICS

Finger	Penetration Rate			
	Right->Right	Left-> Left	Right->Left	All -> All
Thumb	0.54	0.56	0.20	0.37
Index	0.44	0.44	0.37	0.40
Middle	0.85	0.83	0.09	0.47
Ring	0.63	0.70	0.23	0.45
Little	0.92	1.0	0.03	0.49

By equation (1), penetration rate will generally decrease with increasing number of classifications of non-zero probability. The 5-type classification system of Table 1 does not represent an optimal approach by any measure and AFIS classification algorithms do not generally use this system. Further, AFIS can place prints in multiple classifications, so penetration rate cannot be determined from classification probabilities using equation (1). The values in Tables 1 and 2 simply make for an interesting comparison when testing AFIS classification algorithms.

To test AFIS penetration rate, we compared the classifications of each training print to those of all other training prints. Using the vendor's rules of "communication", we calculated the percentage of all comparisons that showed communicating bins. Results were differentiated by finger. As mentioned, 409 volunteers were represented in both training and test data sets. Because of errors in the data collection process, there were only about 404 training-test pairs for any particular finger. All comparisons are symmetric. Therefore, there were about 404x403/2= 81,406 non-independent comparisons made for penetration rate.

The penetration rate benefits of fingerprint classification come at the cost of classification errors. If the individual test and training prints of a matching pair are placed in non-communicating bins, the prints will not be matched. To test bin error using the AFIS binning algorithm, we compared binning assignments for each training-test pair based on the bin communication rules. There were about 404 matching pairs for each finger.

Table 3 shows the bin error and penetration rates individually for thumb, index middle and ring fingers. The binning error rate is best for thumbs and left index fingers and worst for right middle and ring fingers. None of the error rate differences between fingers is statistically significant at even the 90% confidence level[1].

TABLE 3: SINGLE FINGER BINNING ERROR AND PENETRATION RATES FROM TEST DATA

Finger	Error Rate		Penetration Rate			
	Right	Left	Right-> Right	Left-> Left	Right->Left	All -> All
Thumb	0.002	0.002	0.70	0.67	0.26	0.47
Index	0.005	0.002	0.46	0.43	0.40	0.42
Middle	0.012	0.007	0.74	0.66	0.29	0.49
Ring	0.010	0.007	0.74	0.66	0.40	0.55

5. Penetration Rates of Multi-Finger Systems

In Reference [1], prediction of penetration and bin error rate performance for systems using multiple fingerprints was discussed under the assumption that the errors and penetration rates are independent. The general equation for multiple-finger penetration rate can be written as

$$P_{ensemble} = \prod_{i=1}^{T} P_i \tag{2}$$

where P_i is the penetration rate of the i^{th} finger and $P_{ensemble}$ is the total penetration rate of the multi-finger "ensemble". In reality, the binning assignments for thumb, index, midddle, or ring fingers of a person are not independent, but are highly positively correlated. Therefore, we would expect a true penetration rate less than that calculated from equation (2).

Binning error rate for the multi-finger case, again under the assumption of error independence, is given in [1] by

$$1-\varepsilon_{ensemble} = \prod_{i=1}^{T} (1-\varepsilon_i) \tag{3}$$

[1] This is established by testing with a cumulative binomial distribution the null hypothesis that observed errors for each finger could have come from the same error probability.

where ε_i is the bin error rate of the i^{th} finger and $\varepsilon_{ensemble}$ is the total error rate for the ensemble. If errors are positively correlated, the value $\varepsilon_{ensemble}$ of will be smaller than calculated using (3).

Using the same AFIS binning algorithm, we tested about 404 finger pairs for left-right thumb, index, middle and ring fingers with every other similar pair in the training data set. Again, these were symmetric comparisons, so there were about 81,406 non-independent comparisons. Both binning errors and penetration rates were measured and are given as Table 4. Included in Table 4 are the error rates calculated from the test data in Table 3 by equation (3) under the assumption of error independence. Test and calculated error rates are identical except for the case of middle fingers. The middle finger test error rate is slightly smaller than that calculated by (3). In the test data of about 404 pairs, there were two instances of classification errors occurring on both left and right middle fingers of the same volunteer. Again, the error rate differences between fingers is not statistically significant.

Also included in Table 4 are the penetration rates calculated from both test and FBI data in Tables 2 and 3 by equation (2) under the assumption of classification independence. Test penetration rates are somewhat (10-20%) higher for all fingers than those calculated using equation (2) from the test data of Table 3, indicating some positive classification correlations between left and right fingers. Test penetration rates are also higher than calculated using (2) with the FBI data from Table 2, except for the middle finger.

Table 5 shows error and penetration rates for four-finger (both thumbs and both index) and eight-finger binning systems. While binning error rates behave as though independent, penetration rates do not. The penetration rate on the four-finger system was found to be15%, while an assumption of finger classification independence would have lead to a 9% penetration rate based on the single-finger values. The eight-finger system showed a penetration rate of 8%, with a predicted value of 2%.

TABLE 4: TWO-FINGER BINNING STATISTICS

Finger	Error Rate	Error if independent	Penetration Rate	Penetration if independent	
				FBI Data	Test Data
Thumb	0.005	0.005	0.52	0.30	0.47
Index	0.007	0.007	0.25	0.19	0.20
Middle	0.015	0.019	0.55	0.71	0.49
Ring	0.017	0.017	0.55	0.44	0.49

TABLE 5: MULTIPLE-FINGER BINNING STATISTICS

Fingers	Error Rate	Error if independent	Penetration Rate	Penetration if independent	
				FBI Data	Test Data
Four: Thumb and index	0.012	0.012	0.15	0.059	0.093
Eight: Thumb index, middle, ring	0.040	0.048	0.08	0.018	0.022

6. ROC Curves for Communicating and Non-Communicating Impostor Comparisons

In an AFIS system, submitted fingerprints are binned, then compared only to enrolled prints placed in similar (communicating) bins. We might hypothesize that there is a greater probability for prints in communicating bins to be falsely matched than for prints in non-communicating bins. We computed the ROC for the test fingerprints in three ways: comparing communicating impostors only, comparing non-communicating impostors only, and comparing all impostors. Figures 1 and 2 show three ROCs each for right and left thumb comparisons. We note that the false match rate for the communicating comparisons is almost an order of magnitude greater than for the non-communicating comparisons at some points in the ROC.

FIGURE 1:

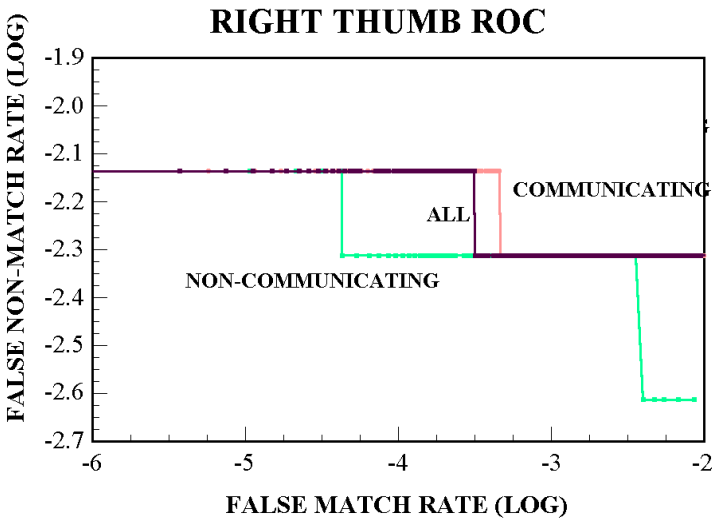
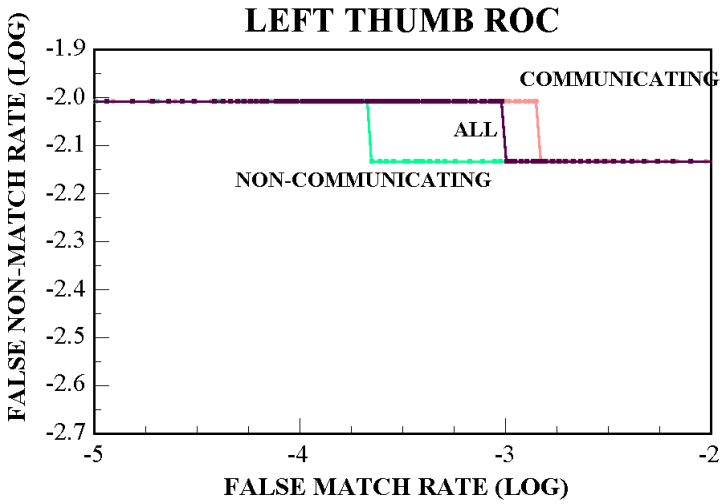


FIGURE 2:



7. Finger Dependency of ROC

Does the ROC vary depending upon which finger is used? We calculated the ROC for thumbs, index, middle and ring fingers using impostor comparisons only with the same fingers from communicating bins. For example, impostor scores for thumbs were developed by comparing right thumbs only to other right thumbs, and left thumbs only to other left thumbs, with communicating classifications. In all, about 410 genuine comparisons and between 100,000 and 200,000 impostor comparisons were made for each finger. Figures 3 and 4 show right and left hand ROCs for each finger position. Both graphs show generally increasing error rates as we move from thumbs through ring fingers.

FIGURE 3:

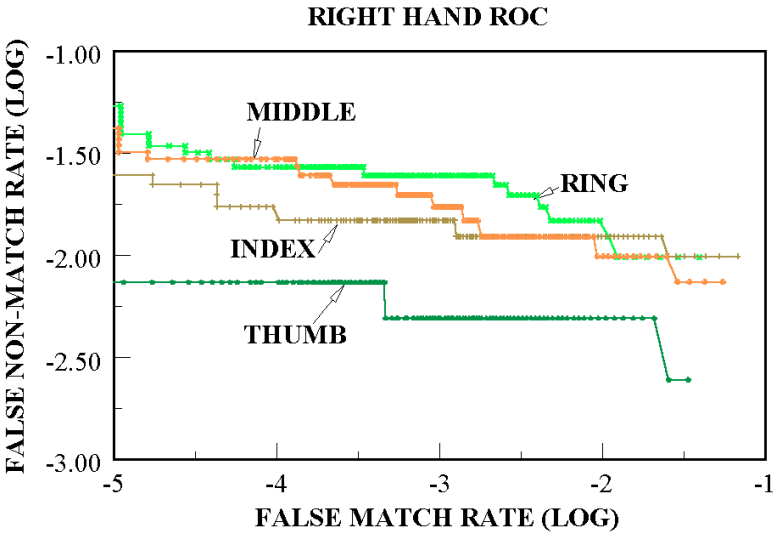
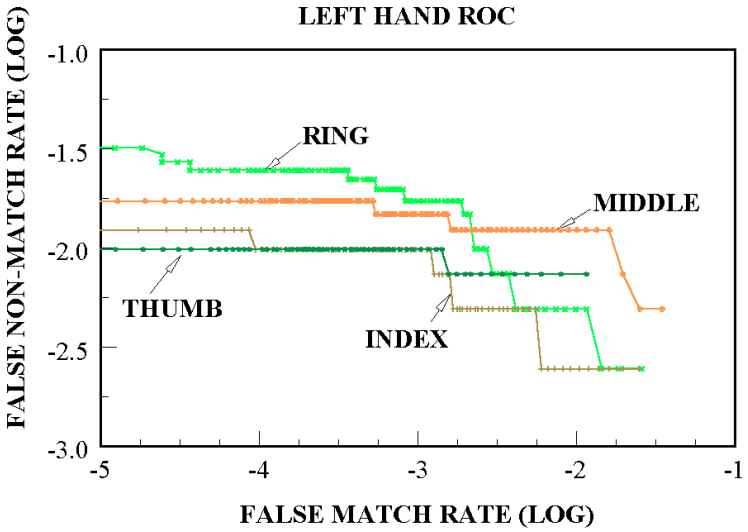


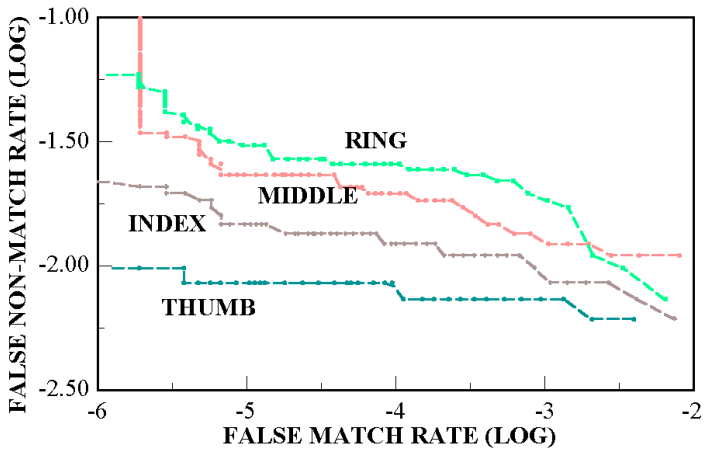
FIGURE 4:



The most notable difference between the right and left hand ROC curves is the difference in thumb error rates, with left thumbs showing worse performance than right thumbs. Figure 5 combines ROCs of both left and right for each finger position and clearly shows increasing errors as we move from thumbs through ring fingers.

FIGURE 5:

ROC BY FINGER



We also tested to see if a correlation exists between left and right finger scores for thumb and index fingers of the individual users. Using the non-parametric Kendall's Tau test [9] over about 409 volunteers with eight fingers in both enrollment and test sets, τ = 0.33 and 0.26 for thumbs and index fingers respectively. Comparing ranks of right thumbs to right index fingers, τ=0.28. None of these measures is statistically significant at any significance level, indicating that individual users do not generally have correlated finger scores.

8. Impostor Distribution Variation Across Test Samples

Researchers in biometric identification talk about "sheep", "goats", "wolves" and "lambs" to indicate the variability of error rates of a specific biometric system across various users [10]. Most users are "sheep" who can use the system consistently well and are not easily impersonated. "Goats" are those users who cannot consistently be identified. "Wolves" are users who can be easily mistaken for another user in a "zero effort"[2] attack. "Lambs" are users easily preyed upon by "wolves".

In the comparison matrix, the fingerprints in the rows can be considered as attempted attacks on the fingerprints of the columns. Because we have only two samples of each finger, we cannot test for "goats", those consistently not matched to their own enrollment template. We can, however, test for "wolves" and "lambs". Because of the lack of score correlation between prints of an individual user, we have chosen to test for "wolves" and "lambs" at the single print level. A "wolf" row will have consistently higher scores across the columns of enrollment prints, not considering, of course, the genuinely matching enrollment image, while a "lamb" column will have higher scores across the rows. Again, we limited our comparisons to prints in communicating bins. Therefore, for each row we summed the scores across all columns that communicated with the row print and for each column we summed scores across the rows. Because the

---

[2] The term "zero effort attack" means that the attack is passive and does not involve active efforts at impersonation.

number of communicating comparisons will vary, these results must be normalized against the number of comparison scores used for each "wolf" row or "lamb" column. This produces the mean communicating impostor score.

If the comparison matrix were symmetric, each "wolf" row mean would be identical to the matching print's "lamb" column mean. The comparison matrix is not symmetric, however, for two reasons. Firstly, the prints represented in the columns are images acquired at a different time from the prints represented in the rows. Secondly, fingerprint comparison scores are not symmetric. The score of the comparison of print A to print B is not generally equal to score of the comparison of print B to print A. Therefore, we computed both the row and the column sums.

Using a one-way analysis of variance [11], we tested the null hypothesis that all the communicating scores in the matrix came from the same distribution against the opposing hypothesis that the distribution was row dependent. Combining results for right and left thumbs, 1420 thumbs were in about 336,000 communicating comparisons . The "F" statistic was calculated at 6.7, which is much larger than the critical value of nearly 1 for this number of "degrees of freedom". Thus, the alternate hypothesis was accepted. This shows that there are "wolves".

Then we repeated this test for column dependency in the thumb data, calculating the "F" statistic as 9.0, with 1437 columns in about 278,000 comparisons. We again accept the alternate hypothesis that the data is column dependent, showing that there are also "lamb" fingerprints.

Figure 6 shows a histogram of the mean row impostor thumb scores. Also graphed is the histogram of the mean column thumb scores. Because these distributions are nearly identical, they are not individually labeled. If all the means were nearly identical, Figure 6 would show a sharp spike. If there were strictly "sheep" and "wolves", there would be two spikes, one at a low and one at a high score value. Figure 6 shows both "lamb" and "wolf" distributions to be smoothly spread. This indicates that there are "sheep" and "wolves", and "sheep" and "lambs", but the boundary between them is not well defined.

Figure 7 shows the same study done on index fingerprints. Results are seen to be the same. Analysis of variance of the index finger rows gave an "F" statistic of 7.5. The "F" statistic for index finger columns was 10.1. With the 1420 relevant rows or columns and the 232,000 communicating comparisons, both of these "F" statistics are significant at all reasonable significance levels.

The existence of lambs and wolves calls into question the suitability of system false-match error rate equations [1] based on the assumption that all stored templates have the same probability of being falsely matched. Equations of the type

$$\mathrm{FMR}_{\mathrm{sys}} = 1 - (1 - \mathrm{FM})^{\mathrm{N}} \tag{4}$$

where $\mathrm{FMR}_{\mathrm{sys}}$ is the system false match rate, FM is the false match rate of a single comparison (assumed to be uniform) and N is the number of stored templates, should be more reasonably replaced with the form

$$\mathrm{FMR}_{\mathrm{sys}} = 1 - \prod_{\mathrm{i}=1}^{\mathrm{N}} (1 - \mathrm{FM}_{\mathrm{i}}) \tag{5}$$

yielding higher estimates for the system false match rate, FMR_{sys}, if $FM_i \neq \text{constant}$.

FIGURE 6:

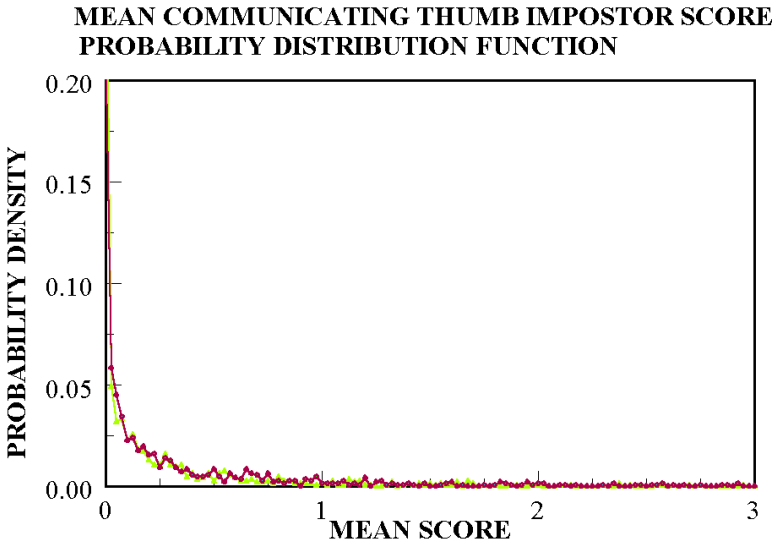
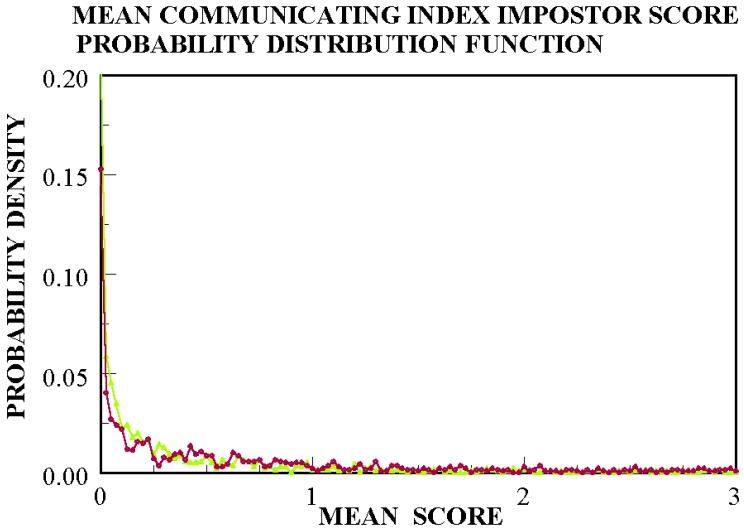


FIGURE 7:



9. Conclusions

We can make the following conclusions:

1) ROCs developed from images in communicating bins show worse performance than those developed without consideration of the binning.
2) Thumbs have lower binning and comparison error rates, but index fingers have better penetration rate.

3) Both binning and comparison error rates increase as we move from thumb, through index to ring fingers.
4) Because of pattern correlations across individual users, penetration rates for multiple finger systems cannot be accurately estimated from single finger penetration rates.
5) Matching scores and binning errors are not correlated across the fingers in the general individual user.
6) "Wolves" and "lambs" exist, but there is a gradual transition between sheep and these populations.
7) The existence of population variability in error rates calls into question the validity of system false match rate equations based upon the assumption that error probabilities are consistent across the population.

10. References

[1] J.L. Wayman, "Error Rate Equations for the General Biometric System", IEEE Robotics and Automation Magazine, March 1999.

[2] J.L. Wayman, "A Scientific Approach to Evaluating Biometric Systems Using a Mathematical Methodology", Proc. CTST'97, pg. 477-492

[3] J.L. Wayman, "Benchmarking Large-Scale Biometric System: Issues and Feasibility", Proc. CTST Government'97, Sept. 1997

[4] J.L. Wayman, "The Science of Biometric Technologies: Testing, Classifying, Evaluating", Proc. CTST'97, pg. 385-394

[5] J.L. Wayman, "Testing and Evaluating Biometric Technologies: What the Customer Needs to Know", Proc. CTST'98, pg. 385-394

[6] J.L. Wayman, "A Generalized Biometric Identification System Model", Proc. of the IEEE Asilomar Conference on Signals, Systems, and Computers, Nov.,1997

[7] J. L. Wayman, "Technical Testing and Evaluation of Biometric Devices" in A. Jain, etal, eds. Biometrics: Information Security in a Networked Society, (Kluwer Academic Press, 1999)

[8] Unpublished 1995 report by Frank Torpey of Mitre Corporation using data extracted from the FBI's Identification Division Automated Services database of 22,000,000 human-classified fingerprint records.

[9] W.H. Press, etal Numerical Recipes in C, 2nd ed, (Cambridge University Press, New York, 1992)

[10] G. Doddington, etal " Sheep, Goats, Lambs and Wolves: An Analysis of Individual Differences in Speaker Recognition Performance", ICSLP'98, Sidney, Australia, November 1998

[11] A.L.Edwards, Experimental Design in Psychological Research, 4th ed. (Holt, Reinhart, and Winston, New York, 1972

A Survey of Face Recognition Algorithms and Testing Results

William A. Barrett
U.S. National Biometrics Test Center

Abstract

Automated face recognition (AFR) has received increased attention in recent years. We describe two general approaches to the problem and discuss their effectiveness and robustness with respect to several possible applications. We also discuss some issues of run-time performance.

Introduction

A formal method of classifying faces was first proposed by Francis Galton in 1888 [GALT88]. He proposed collecting facial profiles as curves, finding their norm, and then classifying other profiles by their deviations from the norm. The classification was to be *multi-modal*, i.e. resulting in a vector of (hopefully) independent measures that could be compared with other vectors in a database.

Automated face recognition (AFR) has been of interest to a growing number of research groups since 1990. Driving the recent development have been improvements in the technology of neural networks, wavelet analysis, computer graphics and machine vision.

As in most other biometric measurement systems, a general goal of AFR is to achieve a high level of performance in *matching a given face against a database of faces.* The performance of an AFR system will be judged by some combination of precision of matching (low level of false negatives and false positives), robustness against adverse factors, high speed, and low cost of the equipment. Adverse factors in AFR include lighting conditions, noise in the image, facial expression variations, glasses, hirsute changes, and posture.

The matching performance in current AFR systems is relatively poor compared to that achieved in fingerprint and iris matching, yet it may be the only available measuring tool for an application. Error rates of 2-25% are typical. It is also effective if combined with other biometric measurements.

A survey of commercial AFR systems is given in [BIOM97].

AFR Technology Categories

The AFR technology falls into three main subgroups, which represent more-or-less independent approaches to the problem: *neural network solutions, eigenface solutions,* and *wavelet/elastic matching solutions*. Each of these first requires that a facial image be identified in a scene, a process called *segmentation.* The image should be normalized to some extent. Normalization is usually a combination of linear translation, rotation and scaling, although the elastic matching method includes spatial transformations. If the eyes and the mouth can be located, these reference points can be used to drive normalization to yield a standardized facial image. This of course, supposes that a nearly frontal view is provided. Few AFR systems work effectively with profile views, if the database consists of frontal views.

Segmentation

Segmentation (locating a face in a busy scene) is often considered a pre-processing step. However, it isn't necessarily simple. The images of many common objects resemble faces, and they may have to be rejected later.

An common segmentation approach uses video motion sequences. A video camera in a fixed location simply watches for moving targets against a stationary background. Then finding the head and something resembling a face is relatively simple. However, this can also be fooled by viewing a television set or some other moving object.

Elastic matching [LADES93] provides some built-in segmentation, and some work by Pentland [PENT94] suggests that eigenfaces can be effective in segmentation.

Applications

A short list of applications is given below. This does not include face recognition problems commonly performed by humans, for example, the use of Identikits, witness testimony, etc.

Table 1

Application	***Prospects of AFR***	***AFR problems***
Credit card, driver's license, passport, personal ID: verification	Very good For accurate verification, should be augmented with other measures	Expanding card code for image Image coding standards Potentially large database
Mug shot matching - yield a smaller list of suspects: identification	Good. Controlled segmentation	Digital conversion of mug shot library Candidate photo required
Bank/store security – identifying a suspect	Good Motion video segmentation	Image may be poor quality - few pixels, varying lighting, expressions
Crowd surveillance – searching for wanted persons	Fair to good	Poor image quality Segmentation difficult Real time performance
Smart room - identifying and tracking people in a meeting room	Good Motion video segmentation	Uncontrolled position and expression

Static matching

In *static matching*, we have a single facial photograph, and are required to find any or all matching faces in a database. The database will typically contain *mugshots* taken under controlled lighting conditions with deadpan expressions. A typical database will already be segmented, whether by manual or automatic methods, with the eye and mouth locations identified.

A candidate photo is often taken with uncontrolled lighting conditions, pose and expression. The subject may attempt a disguise.

An AFR should supply a measure of "closeness" between the candidate and each of the database members. Most AFR systems produce a many-dimensional vector that characterizes a face. Two such vectors can then be compared by reducing their difference to a single linear measure. For example, the Cartesian distance between two such vectors yields an easily computed linear difference measure $d'(i,j)$ between a candidate i and a database member j.

Ideally, d' will be zero for a match and large otherwise. In fact, for a large set of candidates, there will be a double distribution of d', one for the expected matches and another for the expected non-matches. By setting a threshold criterion for d' sufficiently small, we can minimize the rate of accepting impostors, but at the expense of also rejecting authentics. By setting the threshold larger, we can minimize the rate of rejecting authentics, but at the expense of accepting impostors.

The relationship between false matches and false acceptances is commonly expressed as a *Receiver Operating Characteristic,* or *ROC* curve.

The FERET tests

Another way to describe the quality of an AFR system is by *rank-ordering.* For each candidate face, the system is asked to rank-order the faces in the database by the quality of the match. If the system develops a linear measure d' in this process, d' is merely used to produce a simple rank position. A large number of candidates, some in the database and others not, are classified this way.

This works reasonably well in comparing AFR systems provided that the candidate set is large and diverse. However, the performance of a particular AFR on a particular task is not predictable from rank-ordering.

Comparative tests were performed by Phillips, [PHIL96, PHIL97] on systems provided by research teams at MIT (eigenfaces), USC (elastic matching), Rutgers, the Rockefeller Institute, and others. The reported results are mostly rank-ordered, but ROC tests are provided in recent reports .

Variations included pose (full frontal vs. quarter profile, half profile and full profile), glasses/no glasses, image brightness, image scale, and different capture times.

The database includes some lighting variations and changes in facial expression. (Candidates were asked to "make a face" for certain shots). Many of the candidates were photographed over a period of two years, in order to studying aging effects.

The results show that image size was easily overcome by all the methods. Illumination level was a problem for the USC system, but not MIT. Rotation negatively impacted all the methods, but to a somewhat different degree. The USC system was more robust to head rotation than the MIT system. Two years of aging significantly reduced recognition.

The most recent results show the highest scores for Joseph Attick's FACEIT system [FACE97]. Faceit is a commercial product. Details on its algorithm are lacking.

Preprocessing

Segmentation is most easily achieved in a typical surveillance situation through motion video. We merely look for changes from one frame to the next, which usually indicates the motion of a person. Pulling a facial image can be done in a number of ways.

Given a rough outline of a face, the eyes can usually be found by examining the horizontal intensity signature, and correlating it to that from a typical face. The eyes and the mouth will usually be darker than the other areas. Some rough pattern matching of dark circles to find the eye position can then be followed by a triangulation, using a typical mouth position, possibly refining this against the horizontal signature.

Scaling and rotation of the face can next be done by classical methods of pixel averaging. The eye-mouth triangle form the basis of a transformation by which all the pixels can be mapped into a standard orientation.

Neural Networks

A back-propagation neural network can be trained to recognize face images. This is in principle an associative memory problem, for which neural networks offer efficient solutions. However, a simple network can be very complex and difficult to train.

A typical image recognition network requires $N = m \times n$ input neurons, one for each of the pixels in an $m \times n$ image. For example, a low resolution image of 128 pixels square requires $N = 16{,}385$ input neurons. These are typically mapped to a number of hidden-layer neurons, p in number. These in turn map to n output neurons, at least one of which is expected to fire on matching a particular face in the database. It happens that p can be much less than N. The hidden layer is considered to be a *feature vector*. Roughly speaking, it expresses the facial features in a condensed way.

Such a network is difficult to train. To reduce the complexity, Cottrell and Fleming [COTT90] introduced a second back-propagation net as a *classification* net. The autoassociation net is used to train the network, and the classification net yields the matching information.

Although neural networks are used for many image recognition problems, Cottrell and Fleming show in their paper that, "under the best circumstances", a neural network of this design is no better than an eigenface feature network.

Eigenfaces

Eigenface recognition was first proposed by Sirovich and Kirby [SIRO87] as an application of principal-component analysis (PCA) of an n-dimensional matrix. They also present some simple experiments that illustrate the power of their method.

Start with a preprocessed image $I(x, y)$, which is a two-dimensional N by N array of intensity values (usually 8 bit gray scale). This may be considered a vector of dimension N^2, so that an image of size 256 by 256 becomes a vector of dimension 65,536, or, equivalently, a point in 65,536 dimensional space. An ensemble of images then maps to a collection of points in this huge space. The central idea is to find a small set of faces (the *eigenfaces*) that can approximately represent any point in the face space as a linear combination. Each of the eigenfaces is of dimension $N \times N$, and can be interpreted as an image.

We expect that some linear combination of a small number of eigenfaces will yield a good approximation to any face in a database, and (of course) also to a candidate for matching. An image can therefore be reduced to an *eigenvector* $\vec{B} = b_i$ which is the set of best-fit coefficients of an eigenface expansion. Now we can compare a candidate's eigenvector against each of those in a database through a distance matching, for example, a Cartesian measure. The distances found against the database yield both a rank-ordering and a linear closeness measure.

Sirovich and Kirby used an ensemble of 115 images of Caucasian males, digitized and preprocessed in a controlled manner, and found that about 40 eigenfaces were sufficient for a very good description of their set of face images. The root-mean-square pixel-by-pixel errors in representing cropped images (background clutter and hair removed) were about 2%.

Turk and Pentland [PENT91] refined their method, by adding preprocessing and expanding the database statistics. They, too, found that a relatively small number of eigenfaces drawn from a diverse population of frontal images is sufficient to describe an arbitrary face to good precision.

The runtime performance of an eigenface system is very good. The construction of a set of eigenfaces is computational intense, but need only be done infrequently. A set could in theory be developed once and for all time that adequately describes all of man and womankind, including persons yet unborn.

Given a candidate image, the task is finding its characteristic eigenvector, which is computationally equivalent to solving a least mean-squares minimization problem, albeit with N^2 datapoints and p unknowns. This is a matter of a few seconds work on a modern machine.

The final task, of matching an image against a database, is a matter of computing distances between the candidate's eigenvector and those of the database. Using a Cartesian distance, the unit computation is one of adding the squares of p variables, each of which is a difference between two eigenvectors. Even without special hardware, this can be reduced to a few dozen microseconds per comparison, making possible the search of a database of 100,000 images in a few seconds. Pentland claims that a match against a more modest database (a few hundred images) can be achieved on standard hardware (Sun Sparc stations) at frame-rate of the capturing video camera.

The robustness of eigenfaces to facial distortions, pose and lighting conditions is fair. Although Sirovich and Kirby were pleased to discover that their system found matches between images with different poses, the quality of matching clearly degrades sharply with pose, and probably also with expression, as Phillips discovered.

Wavelets and Elastic Matching

Wavelets were first proposed by Dennis Gabor as a tool for signal detection in noise.

A complex Gabor wavelet is described by the equation

$$\psi_{\vec{k},\sigma}(\vec{x}) = \exp\left(-\frac{k^2|\vec{x}|^2}{2\sigma^2}\right)\exp(i\vec{k}\vec{x})$$

k determines the oscillating frequency of the wavelet, and the direction of the oscillation. σ describes the rate at which the wavelet collapses to zero as one moves from its center outward. One can view a wavelet as a continuous wave (the second *exp()* function) propagating in the *k* direction, modulated by a Gaussian envelope (the first *exp()* function).

The general idea is to describe an arbitrary two-dimensional image function *I(x, y)* as a linear combination of a set of wavelets. In image applications, the *x, y* plane is first subdivided into a grid of non-overlapping regions, which may or may not be rectangular. At each grid point, the local image is decomposed into a set of wavelets chosen to represent a range of frequencies, directions and extents that "best" characterize that region. Each grid point will then be characterized by a set of wavelets varying in *k*, but with a constant σ. By limiting *k* to a few values, the resulting coefficients become largely invariant to translation, scale and angle, though not completely. Of course, the initial choice of the subdivision grid implies an arbitrary translation.

The finite wavelet set at a particular grid point forms a feature vector called a *jet.* The set of jets will now characterize the image. These comprise a relatively small set of numbers by which two images may be compared.

Application to face recognition

In the work of Lades, von Malsburg and others [LADES93], each jet consists of 5 logarithmically spaced frequency levels and eight orientations. Their initial grid had 7×10 points spaced by 11 pixels each. Thus an image covered by the grid will be characterized by a total feature vector of 110 values. These are not necessarily statistically independent, owing to some overlap between the grid regions, and other correlations found in face images.

Unfortunately, Gabor wavelets are not orthogonal and complete. The lack of orthogonality implies a computational overhead in finding an optimal decomposition. However, as in the PVD method, this need only be done once for each face in a database. Since the operation is local, only a small number of pixels are involved in each convolution.

Elastic grid matching

Matching a jet feature set with a fixed grid will only be effective if the face image is carefully preprocessed, and the face is reasonably expressionless. Gabor filtering relieves some, but not all, of the burden of preprocessing. In order to accommodate different scales, translations, and even facial expressions and pose variations, Lades and von Malsburg [LADES93] discovered that the grid could be elastically distorted (within constraints), in order to find a best match between two images.

The graph matching is first performed by large translations in order to center the grid on the face. It is followed by small local distortions, chosen in such a way to maintain a planar grid. At each stage, certain of the jets must be re-computed, and their feature vectors combined to obtain a quality measure. The matching is improved by changing the local coordinate system of the jets to correspond to the graph distortion, i.e. when the new grid points are closer together, the coordinate axes are similarly compressed.

It is perhaps not surprising that elastic matching is quite effective in dealing with changes in posture and expression. Small rotations of the face image around any axis

result in what might be considered to be the same face, except for local scale and rotation transformations. The jets therefore should be nearly alike. Changes in expression will affect the jets somewhat, but the grid distortion will to some extent track these changes, and after all, the face is essentially an elastic membrane pushed about by a complex of distributed muscles. No one can remove a freckle or wrinkle through a change in facial expression, though its relative position changes to some extent, and the elastic matching approach seems consistent with this observation.

Performance

The time performance of a *rigid* grid Gabor filtering system is comparable to that of the eigenfaces. Given that each image is characterized by a small set of jet vectors, the matching problem is the same, i.e. one of comparing Euclidean distances between the vectors of a candidate and each of the database members. If the grid can be positioned, scaled and rotated into a canonical position (for example, by first locating the eyes and mouth) by a preprocessor, then a high matching performance on conventional processors can be expected.

However, if elastic matching is employed, the time performance will be relatively poor. Elastic grid matching must be performed on the candidate against *each* database element, a task which requires high speed, and preferably parallel, processors. Lades [LADES93] used a system consisting of 23 transputers operating in parallel. Each transputer is a microprocessor with integrated support for message-passing and distributed memory. The convolution of a 128×128 pixel image with 40 wavelet filters requires less than 7 seconds. Comparison of an image to a stored object takes between 2 and 5 seconds on one transputer. A recognition run, comparing one image against a gallery of 87 stored objects (which is amenable to parallel computation) then takes about 25 seconds, a matching rate which is an order of magnitude smaller than with eigenfaces. But note that elastic matching is more robust with respect to pose and expression.

Bibliography

An extensive bibliography, including many online papers and tutorial materials, can be found online, thanks to the work of Peter Kruizinga:
http://www.cs.rug.nl/~peterkr/FACE/frhp.html
Specific references cited in this paper are as follows:

BIOM97: "Survey: Face Recognition Systems", in *Biometric Technology Today*, July/August 1997. Contains a list of commercially available facial systems and estimates of their effectiveness.

COTT90: G. W. Cottrell and M. Fleming, "Face recognition using unsupervised feature extraction", in *Proc. Int. Neural Network Conf.* Vol. 1, Paris, France, July 9-13, 1990, pp. 322-325.

FACEIT97: See the Web page for papers by J. Attick at the Rockefeller Institute. Details on Faceit are lacking.

GALT88: Francis Galton, "Personal identification and description", in *Nature,* June 21, 1888, p 173-177.

LADES93: M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. V. D. Malsburg, and R. Wurtz, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Trans. Comput.*, vol. 42, no. 3, pp. 300-311, 1993. Describes Gabor wavelet filtering and elastic matching.

LADES97: Jun Zhang, Yong Yan, and Martin Lades, "Face Recognition: Eigenface, Elastic matching, and neural nets", in *Proc. IEEE*, vol. 85, No. 9, Sept. 1997, p 1423-1435. Follow-on to LADES93.

PENT91: Matthew Turk and Alex Pentland, "Eigenfaces for recognition", in *Journal of Cognitive Neuroscience*, vol. 3, No. 1, 1991, pp 71-86. A fundamental paper on the eigenface approach.

PENT94: M. Bichsel and A.P.Pentland, "Human face recognition and the face image set's topology", in *CVGIP: Image Understanding*, vol 59, No. 2, March, pp 254-261, 1994, Academic Press.

PHIL96: P. Johnathon Phillips, Patrick J. Rauss, and Sandor Z. Der, "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results", *Army Research Laboratory, ARL-TR-995*, October 1996. Contains a large number of comparative results, mostly using rank-ordering.

PHIL97: P. Johnathon Phillips, Hyeonjoon Moon, Patrick J. Rauss, and Syed A. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms", to appear in *Proc. IEEE Conf. On Computer Vision & Pattern Recognition 97*, June 17-19, 1997.

SIRO87: L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces", in *J. Opt. Soc. Am. A*, Vol. 4, No. 3, March 1987, pp 519-524. A key paper on eigenfaces.

"Degrees of Freedom" as Related to Biometric Device Performance

James L. Wayman, Director
U.S. National Biometric Test Center

Recently at least 4 articles and papers [1-4] have proffered "degrees of freedom" as a figure of merit for biometric devices. The purpose of this article is to demonstrate that single measures, such as "degrees of freedom", cannot be generally used to compare biometric device performance. We will first, by use of a limiting case, establish that "degrees of freedom" is insufficient to determine performance. Then, we will show that the impostor distributions of at least two biometric devices cannot be characterized by "degrees of freedom". Finally, we will review other scalar performance measures to show that these also are not generally applicable as figures of merit for biometric devices.

A "probability density" is a mathematical function which allows us to compute the chances of a measure lying between two values. A "bell curve" is an example of such a probability density. If student test scores follow a bell curve, the probability of a student's score lying between 50 and 60, for instance, is equal to the area under the curve between the score values of 50 and 60. Figure 1 illustrates this example.

Strictly speaking, the bell curve is only appropriate if the measures can take on fractional values. A more appropriate probability density curve can be the "binomial distribution", in the case that the measures can only be whole numbers (as usually occurs with test scores). The mathematical expression for the binomial distribution is

$$y(i)=N!/(i!(N-i)!)\ p^{i}\ (1-p)^{N-i}$$

where *p* is the probability of a particular outcome and *N* generally describes the number of independent trials or repetitions of a random experiment [7]. In the case where p is close to 0.5, this distribution resembles a "bell curve" for any value of N.

The exclamation mark, as in *N*!, has a special meaning in mathematics. It is called "factorial" and means that all integers from 1 to *N* are multiplied together. For instance, 3!=1x2x3=6. To make a graph of the above equation, we need to choose values for both *N* and *p*. References [1,2] call *N* the "degrees of freedom". Reference 5 claims that the binomial distribution is a good fit for observed impostor scores in iris scanning. References [1-3] indicate that the binomial distribution should be chosen with N= 266 and p= 0.499 [1,3]. We hypothesize that other distributions fitting the observed values are very possible. Figure 2 shows this curve with the i-axis (labeled "SCORE") normalized by division by *N*. The area under the curve between any two points on the "score" axis is the probability that an impostor score lies on that interval.

What is the performance of a device with such an impostor distribution? We can't yet say. Device performance is related to both impostor and genuine distributions[6,7]. Without knowledge of the genuine distribution, no statements regarding "false match" and "false non-match" rates can be made, which are the undisputed measures of biometric system performance. Consider a biometric device where genuine and impostor distributions are both characterized by a binomial distribution with N=266 and p= 0.499. The genuine and impostor distributions are identical. We use this extreme case to demonstrate that the "degrees of freedom" is insufficient to characterize performance. The "false match" and "false non-match" error rates in this case are graphed as a "Receiver Operating Characteristic" curve in Figure 3. In this case, the equal error rate is 50%. Increasing the "degrees of freedom" for both

densities does not increase the accuracy of the device. Figure 4 shows identically overlapping genuine and impostor distributions with 532 "degrees of freedom". The "Receiver Operating Characteristic" curve is absolutely unchanged and is still given by Figure 3. The equal error rate is still 50%. Clearly, in this limiting case, error rates are independent of "degrees of freedom".

We might seek to correct the assertion that "degrees of freedom" is a measure of device performance by incorporating the difference in the parameter p between the genuine and impostor binomial distributions. This difference is also important and that together N and the two p values characterize the performance of any device. Even this would not be generally correct because most devices do not have probability densities characterized by the binomial distribution. Figure 5 shows the impostor distribution developed in one test of fingerprinting [8]. Figure 6 shows the impostor distribution developed in an unpublished test of a hand-based biometric device. Neither of these densities is not adequately modeled by the binomial distribution, so the concept of "degrees of freedom" cannot be applied in a discussion of error rates for these devices.

Several related concepts in classical pattern recognition theory are well established for measuring the separation of two distributions of classes, when the classes are "bell" shaped. Let's first model the genuine distribution grossly with mean probability p_1 and variance σ_1^2 and the impostor distribution grossly with mean probability p_2 and variance σ_2^2. This gross modeling of the distributions, however, has been shown wrong in many large-scale, realistic applications.

The first of these separation measures is the Fisher ratio [10], which attempts to measure the overlap between two distributions. Overlap is based on two things: the separation between the distributions and the width of the distributions. Fisher's discriminant is based on the ratio of inter-class separation over the average intra-class spread.

$$F = (p_1 + p_2)^2 / (\sigma_1^2 + \sigma_2^2)$$

This is clearly related to the "decidability" index, defined originally in [11] as,

$$d' = |p_1 - p_2| / \sigma$$

where $\sigma_1^2 = \sigma_2^2 = \sigma^2$. This concept was expanded by [12] to include cases where the distributions have different variances as

$$D^{1/2} = |p_1 - p_2| / ((\sigma_1^2 + \sigma_2^2)/2)^{1/2}$$

Information theory provides yet another dissimilarity measure between two classes, known as divergence, I. Divergence is the total average information for discriminating one class from another [13]. This measure has been successfully applied already in biometrics for speaker identification [14]. Interestingly, when the two distributions are bell curves with same widths, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the divergence measure, I, is similar to the previous measures, and is given by

$$I = (p_1 - p_2)^2 / \sigma^2$$

Two biometric devices can be compared using these measures only if the genuine and impostor distributions for both devices have similar shapes. If the distributions of the two devices are dissimilar, none of these provides a comparative measure of device performance. Figures 5 and 6 show impostor distributions that depart significantly from the binomial distribution used to model the impostor distribution of iris scanning. In general, due to the difference in distributional shapes, biometric devices cannot be compared using any of these measures.

We conclude that the concept of "degrees of freedom" only applies to devices whose distributions are well modeled by the binomial. Even when distributions are binomial, devices with more "degrees of freedom" may not have lower error rates. Further, the concept of "degrees of freedom" is not applicable to most biometric devices . Other classical measures of distribution separation can be used to compare device performance only in the unusual case where the devices have distributions which are similarly shaped. Consequently, the relative performance of a biometric device cannot be expressed in any single number.

References

[1] G.O. Williams, "Iris Recognition Technology", IEEE AES Systems Magazine, April 1997, pg. 23-29

[2] J. Daugman, "Continuing Debate on Issues and Strategies in Large-Scale Biometric Searches", Biometrics in Human Services User Group News Letter , volume 3, no.1 ,January 1999, downloadable from www.dss.state.ct.us/digital.html

[3] C. Seal, M. Gifford, D. McCartney, "Iris Recognition for User Validation", British Telecommunication Engineering, vol.16, July 1997

[4] C. Wu, "Private Eyes", Science News, Vol.153, No.14, April 4, 1998

[5] R.P. Wildes, "Iris Recognition: An Emerging Biometric Technology", Proc. IEEE, Vol. 85, No. 9, Sept. 1997

[6] J.L.Wayman, "Testing and Evaluating Biometric Technologies: What the Customer Needs to Know", Proc. CTST'97, pg. 329-348 , also on www.engr.sjsu.edu/biometrics

[7] J.L.Wayman, "Technical Testing and Evaluation of Biometric Devices", in A. Jain, etal (ibid)

[8] J.L Wayman, "Biometric Identification Standards Research—Final Report: Vol. 1" (1997), report to the Federal Highway Administration, available for downloading at www.engr.sjsu.edu/biometrics/fhwa.htm

[9] Robert V. Hogg and Allen T Craig, "Mathematical Statistics", 4th Edition, Macmillan Publishing Co, Inc., New York, 1978.

[10] Thomas Parsons, "Voice and Speech Processing", McGraw-Hill Book Co, New York, 1987.

[11] Tanner, W.P. and Swets, J.A., "A Decision-Making Theory of Visual Detection", Psychological Review, Vol. 61, (1954), pg. 401-409

[12] Swets, J.A.(ed.), Signal Detection and Recognition by Human Observers (Wiley, 1964)

[13] J. T. Tou and R.C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley Publishing Co, Reading Mass, 1974.

[14] Joseph P. Campbell, "Features and Measures for Speaker Recognition", Oklahoma State University, Dec 1992.

FIGURE 1:

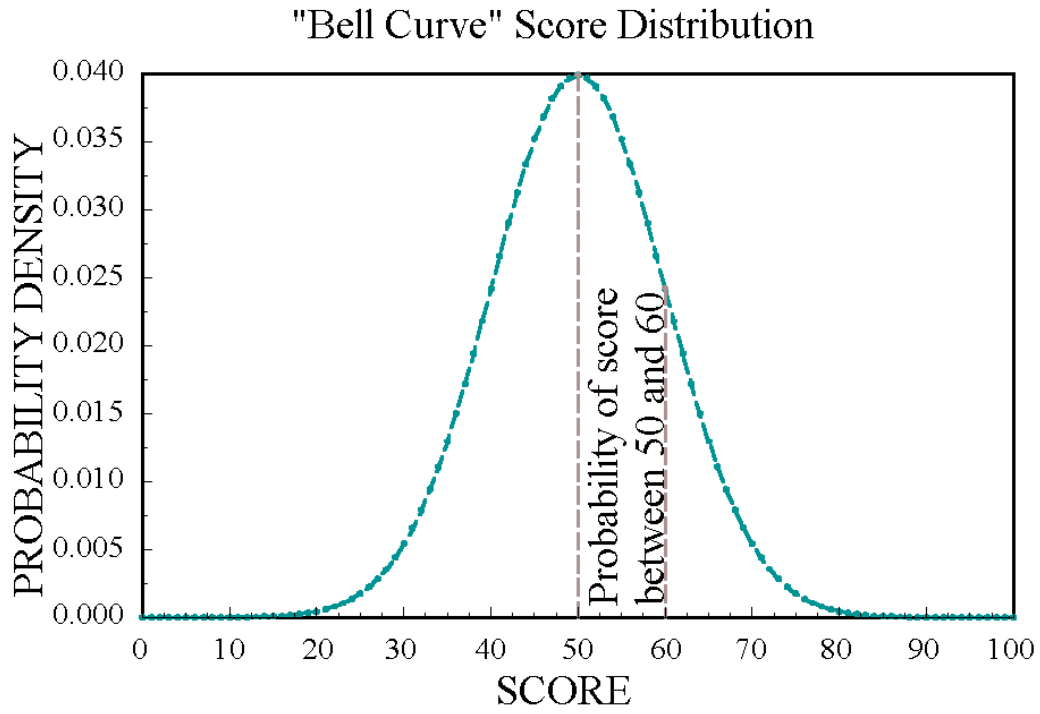


FIGURE 2:

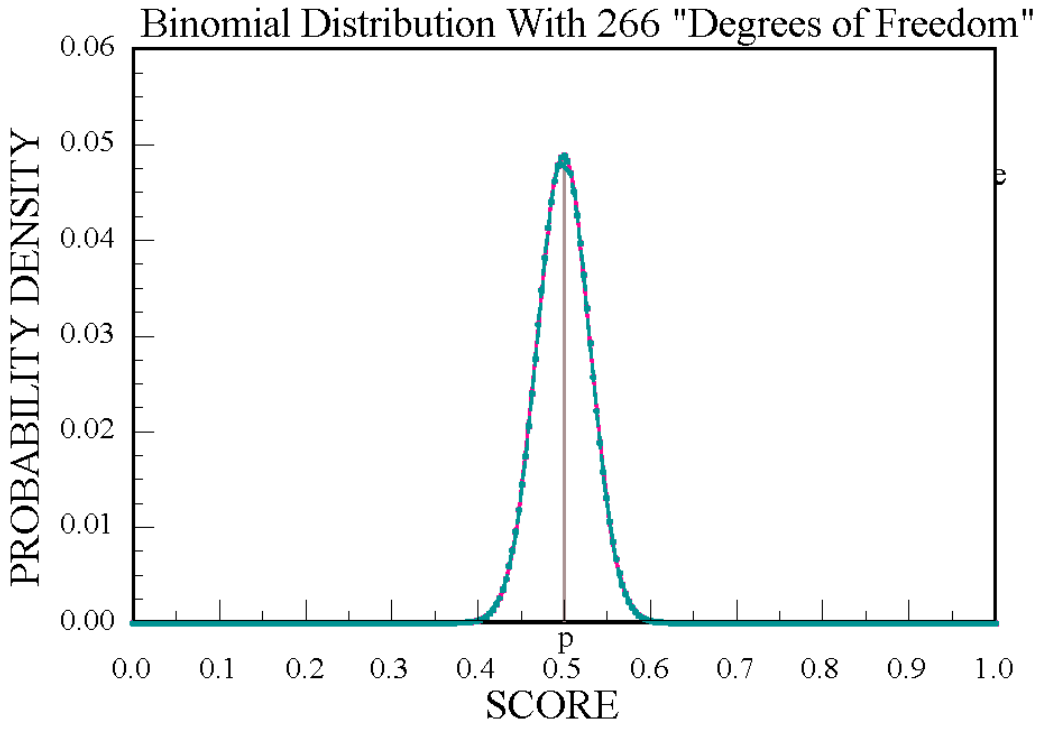


FIGURE 3:

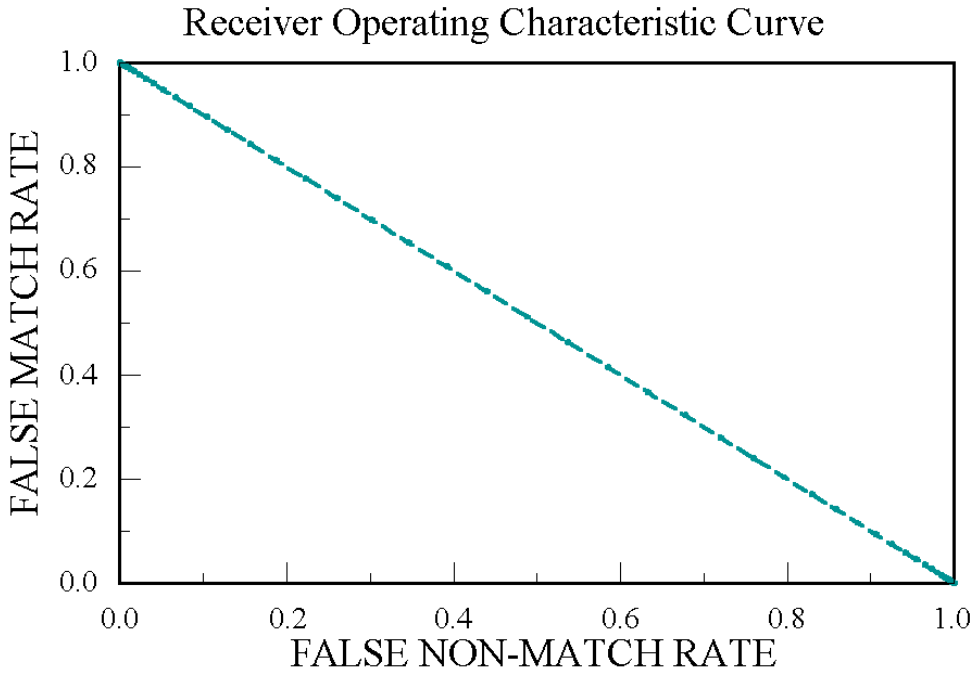


FIGURE 4:

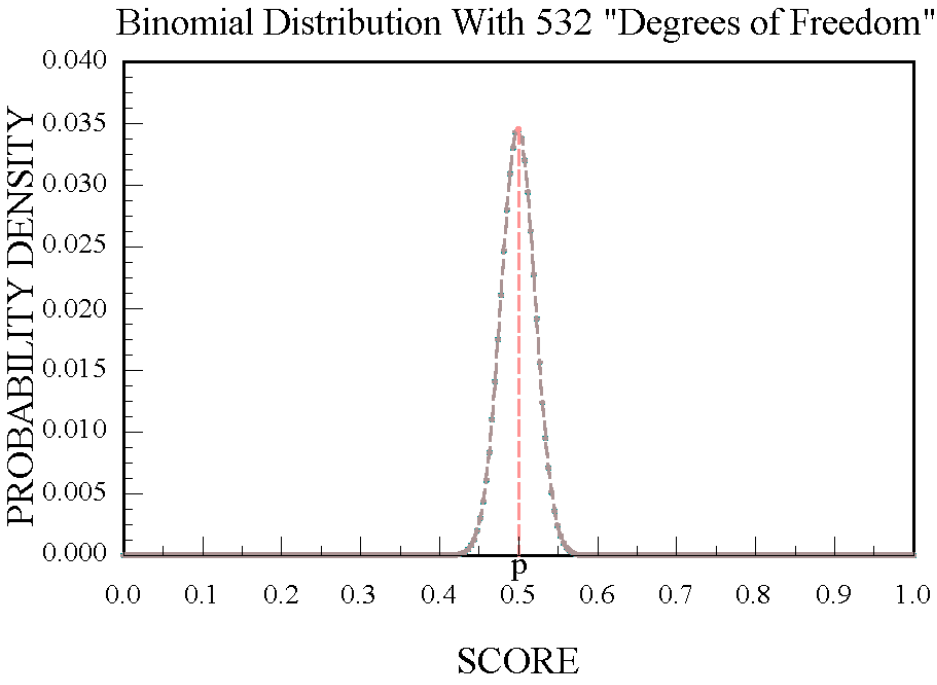


FIGURE 5:

Fingerprint Impostor Distribution

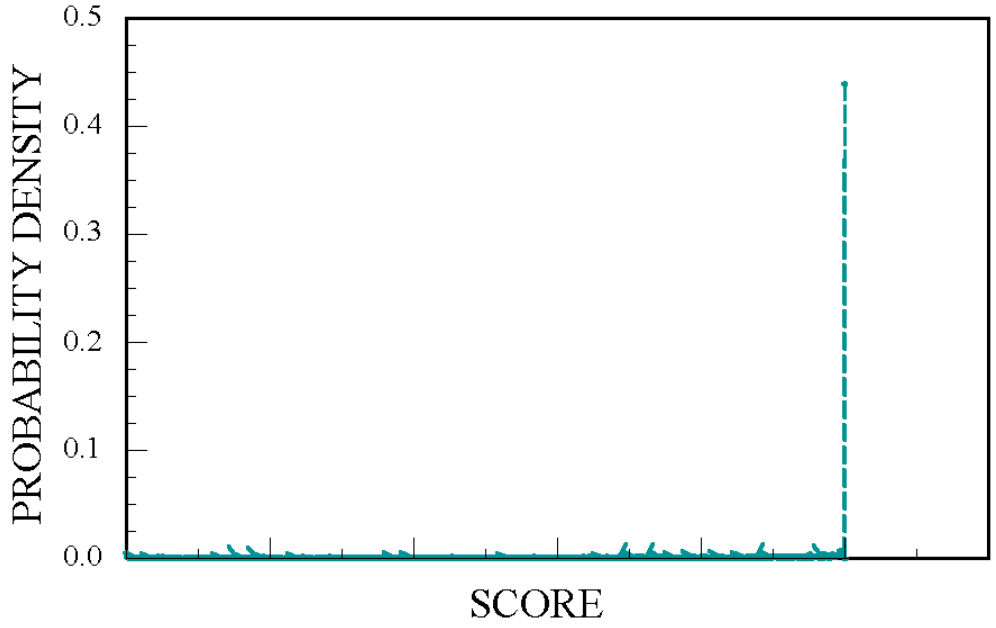
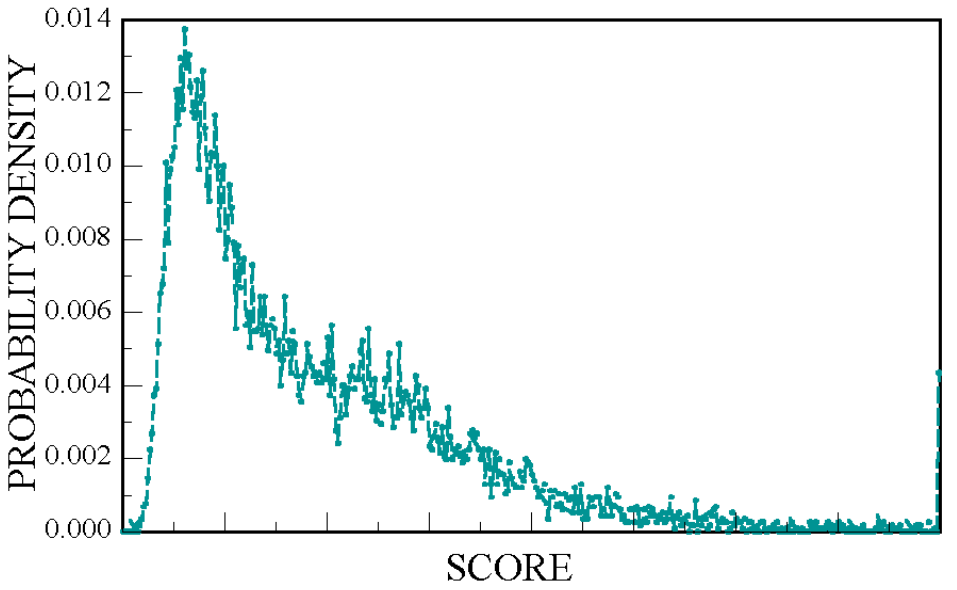


FIGURE 6:

Impostor Distribution for a Hand-Based Biometric Device



Engineering Tradeoffs in Biometric API Design

John M. Colombi, Government Director
James L. Wayman, San Jose State University Director
U.S. National Biometric Test Center

I. Introduction

Over the roughly 25-year history of the industry, biometric identification devices have been seen as "high end" security tools, the stuff of James Bond movies, useful for controlling access to highly secure government facilities. Most of the investment by government and industry in development and testing of biometrics has been targeted at the creation of highly secure (and costly) tools for protection against facility intrusion by master criminals and terrorists. But the world is changing. Facility protection, while still important, is now seen as only one application of biometric technology. The recent creation of the Reno Commission to study national infra-structure security has underlined the need to protect computer, banking, utility and communication networks, as well. With the rise of communication networks, facilitating rapid data exchange and electronic commerce, information security is now understood to be of vital importance. Further, the decrease in price and increase in ease of use of biometric devices has created broader, government and consumer applications aimed at increasing convenience and lowering system costs, while maintaining adequate levels of security. Biometric systems are no longer "high-end", custom-made, "one-offs", but are becoming available as "commercial off-the-shelf" devices for a wide variety of government and consumer applications. Consequently, everyone stands to benefit by biometric Application Programming Interface (API) standardization activities aimed at facilitating "plug and play" biometric devices and applications.

During the past few years, various vendors have produced and marketed biometric software development kits. These typically provide access to the large number of functions which control operation of a specific biometric technology or, at a lower level, control a particular biometric device. Ideally, however, a biometric API should be more than a complete toolkit of vendor specific functions. The perfect API would allow maximum flexibility after system purchase and installation, allowing users to reconfigure and tune their systems to meet evolving application requirements. Broad acceptance of such an API standard would stimulate growth in the use of biometric technology in consumer, industrial and government applications, by breaking the "lock-in" inhibitions of application vendors, purchasing departments and end users.

While software standardization provides a multitude of benefits, including interoperability, modularity and quicker development cycles, the adage "you don't get something for nothing" appropriately applies to biometric API designs. This paper describes the many engineering trade-offs and design alternatives which should be considered in the development of a truly generic biometric API.

II. Level of Generalization and Abstraction

What, specifically, is an Application Programming Interface? Defining more precisely our ideal API, based on our projected application requirements, will be a primary goal of this paper. Determining the right level of abstraction and generalization directly relates to the philosophy and intent of the API development effort. For example, if the goal includes easy introduction of biometric capabilities to various application

developers, then a high-level, simple approach should suffice. However, if the goal includes standardization between algorithm/ engine developers and device manufacturers or interoperability of processes and data formats, a much lower level of abstraction is necessary. Both are equally useful and relate to the design philosophy.

A . How high?

At the highest level of abstraction, there are two only general functions: enrolling a customer's biometric pattern into a database (which might contain as few as one or as many as millions of other patterns or templates), and matching that biometric pattern against that database. Even at this high level, however, there are areas of uncertainty as to the required scope of the API. Biometric devices can be used for verification, in which the customer claims to match a specific template in the database, or identification, in which no specific template in the database is pointed to by the customer. In addition, there are two modes of identification: the customer claiming to match some unspecified template in the database, or claiming not to match any template in the database. This first mode of identification is sometimes called "PIN-less verification" as it allows the system to identify the customer without requiring a Personal Identification Number (PIN), identification card, or any other token. The second mode of identification is used at the time of enrollment in applications that prohibit multiple enrollments by the same individual, such as social service and driver's licensing applications.

Should the API allow for only verification, or also identification? If identification is recognized, should both forms be supported? Generally, access control mechanisms, even those employing biometric recognition, use passwords, PINs, mag-stripe cards, smart-cards or other identity tokens for the purpose of claiming an identity for the user. If biometric technology is to be embedded in applications as a substitute for these token-based methods, identification capability will be needed by the API. If our goal in developing an API standard is to promote the usability of biometric systems, then the broadest collection of applications should be supported at this "high level".

B. Technology Creep

A diverse collection of scientific disciplines, including image processing, multidimensional feature extraction, signal coding, information theory, neural networks, and statistical pattern recognition, are represented within the mainstream biometric technologies. The scope and diversity of the technologies provide a challenge for system integrators and application developers looking to provide general biometric solutions. The ideal API should shield not only the end users, but the application developers as well, from the complexities of learning the specifics of each biometric device in environments which have yet to be defined.

We introduce the term "Technology Creep" to describe the addition of technology specific functions into the API, required as the list of supported devices and applications increases. For example, "cepstral liftering" and "cohort normalization" are standard techniques, but only for speaker recognition technology. Again, the ideal API should shield the application developer from direct involvement with these techniques. Video and image processing functions may be applicable to face, iris, hand and fingerprint verification, but may not be required for speaker recognition, finger-based methods and dynamic signature analysis. The ideal standard biometric API should not let specific

requirements of any biometric technology creep into the design. Therefore, "low level" technology and device specifics should be avoided within the interface.

C. How Low?

While technology creep needs to be avoided, "low level" device control needs to be included. The range of device controls is immense and extremely vendor specific. We refer to the control of the "bells and whistles", and the many other technology specific calls, required for optimal capture and processing of biometric signals. For example, in speaker recognition applications, "end-of-signal" detection algorithms may require energy levels, noise estimates, or zero-crossing rates to be passed from the device to the API, or decision thresholds to be passed from the API to the device. The API may need to instruct fingerprint devices to tune contrast or lighting. Facial recognition systems may require adjustment, through the API, of camera contrast or color settings for optimal performance.

The API design should allow for the "low level" activities of defining range and type of device control, and setting or acquiring device thresholds in an extensible manner.

III. The Architecture and Model

The "High/Low" abstraction is not the only way of viewing the required engineering trade-offs. Figure 1 depicts the overall architecture of a generic biometric API and suggests functions and interfaces that can and should be standardized.

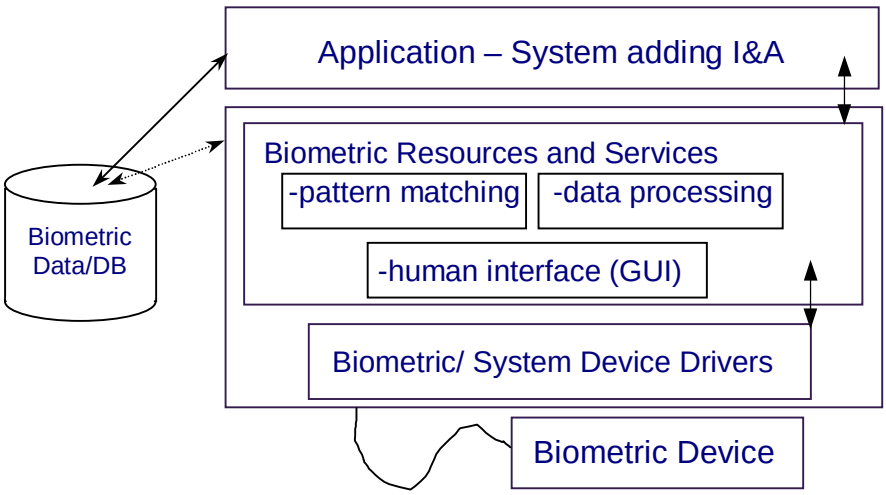


Figure 1: Generic Biometric Architecture

The four main components which would readily benefit from APIs include:

- Applications: all common applications, such as "Database Management Systems", or network servers, which might benefit by adding biometric authentication and identification (A&I) services;

- Biometric Resources and System Services: the "middleware" of biometric identification processing, algorithms, engines;

- Biometric Devices: sensors and pre-processor hardware;
- Biometric Data and Databases: stored biometric data , as either unprocessed images, or in processed, "template" form;

In addition, the interfaces between these groups should be addressed. These primarily include the interface between:

- Applications and biometric services;
- Biometric services and biometric devices;
- Biometric data and both applications and biometric services;

Standardizing each of these interfaces promises a number of unique benefits and should be considered in the context of the developing market.

A. Applications

It seems to be a rule that the availability of technologies creates applications for them. Only a very few years ago, for instance, few of us would have understood the need for Internet access from within a word processing application. We believe the same may prove true for biometric technology. When the ability to interface with these technologies becomes available, application developers and end-users will find uses for them not predictable now. Yet development and standardization of the API application/biometric services interface requires a prescience of what these applications might be. So which comes first, a definition of the API or an enumeration of the applications? The only conceivable solution is that the process will be iterative, with development of both APIs and applications occurring together. For this reason, we question whether a single standard API can or should be created now. At our current stage of development, competition between APIs should be encouraged.

Clearly, the development process could be accelerated if the software applications industry were to identify which functions are desired at the application level, or perhaps, which functionality should be provided by the API. As mentioned, biometric technologies vary greatly in their methods of capturing, processing, and matching the input data. Speech, dynamic signature, finger-geometry and retinal verification systems acquire and process one-dimensional data. Face, iris, hand and fingerprint methods acquire and process two-dimensional data. Three-dimensional data acquisition systems are on the horizon. Varied too are the uses within software applications. Simple verification might be all that is desired, with enrollment done through some other means, perhaps associated with the system's administration service. Or an application may require both enrollment and "PIN-less" verification, or even full-scale identification.

To clarify the discussion, we offer a multi-level conceptualization, as shown in Figure 2. If an API were to contain interfaces at all of these levels, applications developers and end-users could employ the level that matches their technical capabilities, resources, and functionality requirements, whether it was access control using hands or faces, or large-scale identification using fingerprints or eye scanning.

Thus, the high-level portion of the API could provide a few simple functions, or the developers could tailor their application to interface directly with data acquisition

devices, thus using the functions of the lower-level interface. Overall, based on vendor discussions, we believe that more than 90% of applications would not require the device-level functions of the bottom level of the interface.

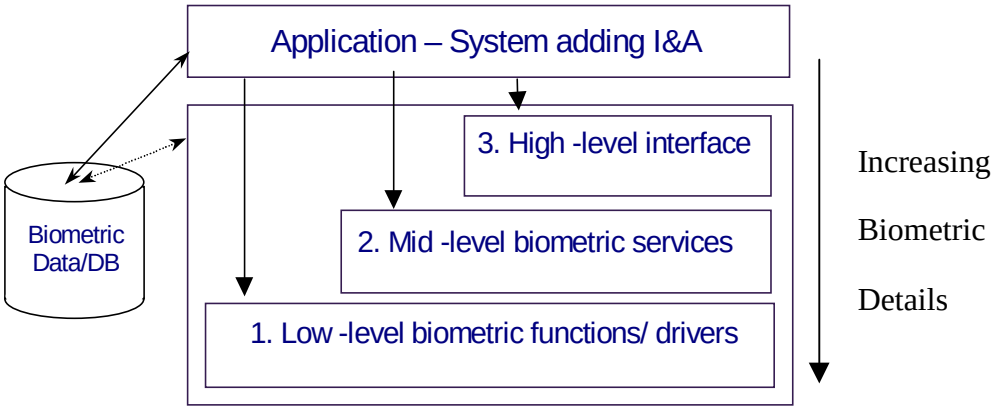


Figure 2: Multi-level Abstraction of API Design

B. Biometric Services and Device Manufacturers

API designs may include specific procedure calls and functions between biometric services middleware (processing, quality control, pattern matching, etc.) and devices. For applications which use non-biometric specific sensors, such as face or voice recognition, the middleware may be inherently flexible, easily accepting various cameras or microphones, frame grabbers or A/D boards, or easily allowing various standard system drivers (i.e. Video for Window (VfW)). Vendors who produce biometric-specific devices, such as an optical or chip-based fingerprint scanners, typically provide various drivers and algorithms with their product.

Such inherent relationships between devices and biometric middleware are not always the case. Integrators often license image processing and pattern matching software from other vendors, then supply their systems with hardware from a yet another source. In this latter situation, the middle-level of the biometric API would provide device abstraction, facilitating better integration between algorithms and devices.

C. Biometric Data

Biometric data can be of two types, processed "templates" reduced from the input samples and the raw samples themselves, although most biometric systems store only the templates. Another interface in the overall API design could be to control access to this data. This interface specification could include data header and formatting information and access privilege mechanisms. A significant trade-off in the design includes not only how access is controlled, but who manages the biometric data. Since the biometric template, like the password, enables access to the system, it should be protected to insure the integrity of the data authentication process. Protection of any stored raw samples is required to protect user privacy.

Consider the example of a large personnel database, where biometric data can be added to each user record as simply another field. In this model, the biometric data must

be extracted and passed across the API to the biometric services. Likewise, data captured from the biometric device must be passed back into the database. Vendors may choose to create a separate biometric database using standard query language access. In either case, biometric data and templates, and their security, are the responsibility of the application developer.

A second example would be the use of biometric verification to supplement existing username/ password directory services, say as an NT Secure Account Manager (SAM) database or UNIX password file. This method may prove complex and operating system dependent, yet provide better a safeguard of the user biometric templates. Further, this method may provide the only solution to the problem of requiring enrollment of a single user into each application database, and the subsequent storage of redundant user templates.

IV. General Functionality

All biometric systems perform a baseline series of services. These typically include:

- Enumeration of the services available
- Device control and configuration
- Data capture
- Processing and feature extraction
- Recognition
- Enrollment, re-enrollment, and template adaptation
- Data storage
- Cryptographic support
- Graphical User Interface considerations

A. Enumeration, Control and Configuration of Biometric Devices

As corporate, Government and DoD users integrate biometric services when upgrading existing networks or installing new networks, backward compatibility with current equipment, having different or no biometric technology, will be required. There may be situations where various versions of fingerprint scanners (keyboard or mice built-ins, or external units), cameras, microphones or other input sensors, may co-exist in the same system. The API must be able to specify a broad range of services and devices, either by machine and/or by user. Further, applications will need to control and configure the devices, and set thresholds, through the API.

B. Data Capture

The data capture functions might include auto-capture, liveness detection, and image quality assessment, in addition to the return of the captured data. Some capture

devices return raw signals, while some return processed features or templates. The API must account for both possibilities. Functionality of blocking or non-blocking should be considered. Even devices returning processed features might also return raw biometric data for storage or for future template adaptation. Returned raw data may be in a standard format such as TIF, WAV, Sun AU, MPG, AVI, BMP, JPG, etc. In addition to the ability to handle any format, the API might also inform the application of format type returned by the device.

C. Processing and Feature Extraction

Biometric systems do not compare raw data signals directly. Rather, a small number of "features" extracted from the signals are used as the basis for comparison of samples to stored templates. Systems reduce the size of the data set by extracting "features" that contain all the pertinent, user-unique information, while rejecting the "noise". Noise is defined here as the unimportant variations caused by changes in the pattern, presentation or sensor. Typically, feature extraction is done with propriety algorithms and results in proprietary feature templates. It would be difficult to enumerate all vendor proprietary template types in the API design. Whether feature extraction is a device function or a middleware function depends upon the device design of the sensor vendor. The ideal API needs to be able to handle either case.

Lastly, the processing or final processed content may be dependent on whether identification or verification is chosen, and may contain exogenous information, such as user's gender, presented finger or spoken phrase text.

D. Recognition

Recognition is based upon a quantitative value related to the degree of match between a submitted sample and a stored template or model. Scores, similarity measures, distance, distortion, or probability expressions are among the many ways to determine the degree of match between the features of a submitted biometric sample and the stored enrollment feature template or model. In some biometric systems, such as hand geometry and some facial recognition systems, matching requires only the calculation of the geometric distance between the sample and enrolled feature vectors. In other cases, such as iris scanning and some speech systems, adjustment in the alignment of the sample and enrollment feature vectors may be required before a distance can be calculated. In yet other cases, such as fingerprinting, the features are not vectors and can only be compared by sending sample and enrollment features into a "black box" for vendor proprietary computation.

There are some cases in which the features extracted from the sample are not in the same form as the enrolled model. For example, in speaker verification, a likelihood measure is computed between features, such as a time series of multidimensional cepstral coefficients, and a set of speaker-dependent Hidden Markov Models (HMM). These HMMs contain means, variances and transitional probabilities of the enrollment features, but are not the feature coefficients themselves. Lastly, there are "cohort" designs which compare sample features to stored models of the claimed identity in the context of stored models from other, non-claimed identities. The truly general API design must consider all of these comparison methods.

So, at this point, we have a score or scores as determined by the biometric service provider from the comparison sample input from the biometric device with templates

from the database. The design of the API now must determine where the decision policy is invoked. If the application makes final accept/reject decision, then the application needs to know the variation and approximate probability density functions of the vendor scores, as well as the policy set by the user, to knowledgeably make a decision. A trade-off would be to allow the application to set or pass the threshold for the current transaction and have the biometric technology return a match/no-match decision, which the application converts to an accept/reject decision.

Some biometric algorithms, even those based on single biometric methods, execute multiple feature extraction algorithms, then use multiple thresholds to reach final decision. For example, speaker verification may examine the signal-to-noise ratio, word error rate and speaker score to declare a match. If the signal is overly noisy, or the word error rate is too high, then a non-match is declared even if the speaker model matches. Other examples include a fingerprint system that measures a high comparison score, but declares a non-match because the incorrect finger was scanned (expecting index, got middle) or because a liveness detector fails (fake print). Thus, determining a match may be more complicated than simply establishing at the application level a single threshold.

E. Enrollment, Re-Enrollment and Template Adaptation

Lastly, the API should be able to handle multiple forms of enrollment. These should include batch or off-line enrollment, re-enrollment, and both biometric service and application initiated updating of previously enrolled models. Re-enrollment may require the original enrollment biometric data in addition to the current sample and/ or the current template. The number of templates and models and the number of samples collected are all vendor-specific and should not be limited by the API. For example, face recognition may use several poses in the creation of an optimal template, or may store several templates from multiple poses for each enrolled user. Fingerprint vendors may require multiple fingers and multiple samples of each (typically 1 to 3) in the creation of a user model. Speaker verification vendors may create templates from multiple samples of the same pass-phrase, one or multiple samples of multiple specified utterances (digits, for instance) or even from several seconds of user-determined, free text.

While enrollment and re-enrollment seem straightforward, handling template updating/adaptation could prove complicated across biometric technologies because of the multiple approaches currently used. The decision to update the template could come from the application, perhaps in response to an expiration date. The decision could come from the biometric service based on decreasing match scores (template “aging”), even though verification scores remain above the decision threshold. Template updating might be done by averaging new feature vectors into the old feature template, by replacing the old template completely with a new one from a verified sample, or by establishing a new model by mixing the original raw enrollment data with new submission.

F. Biometric Data Storage

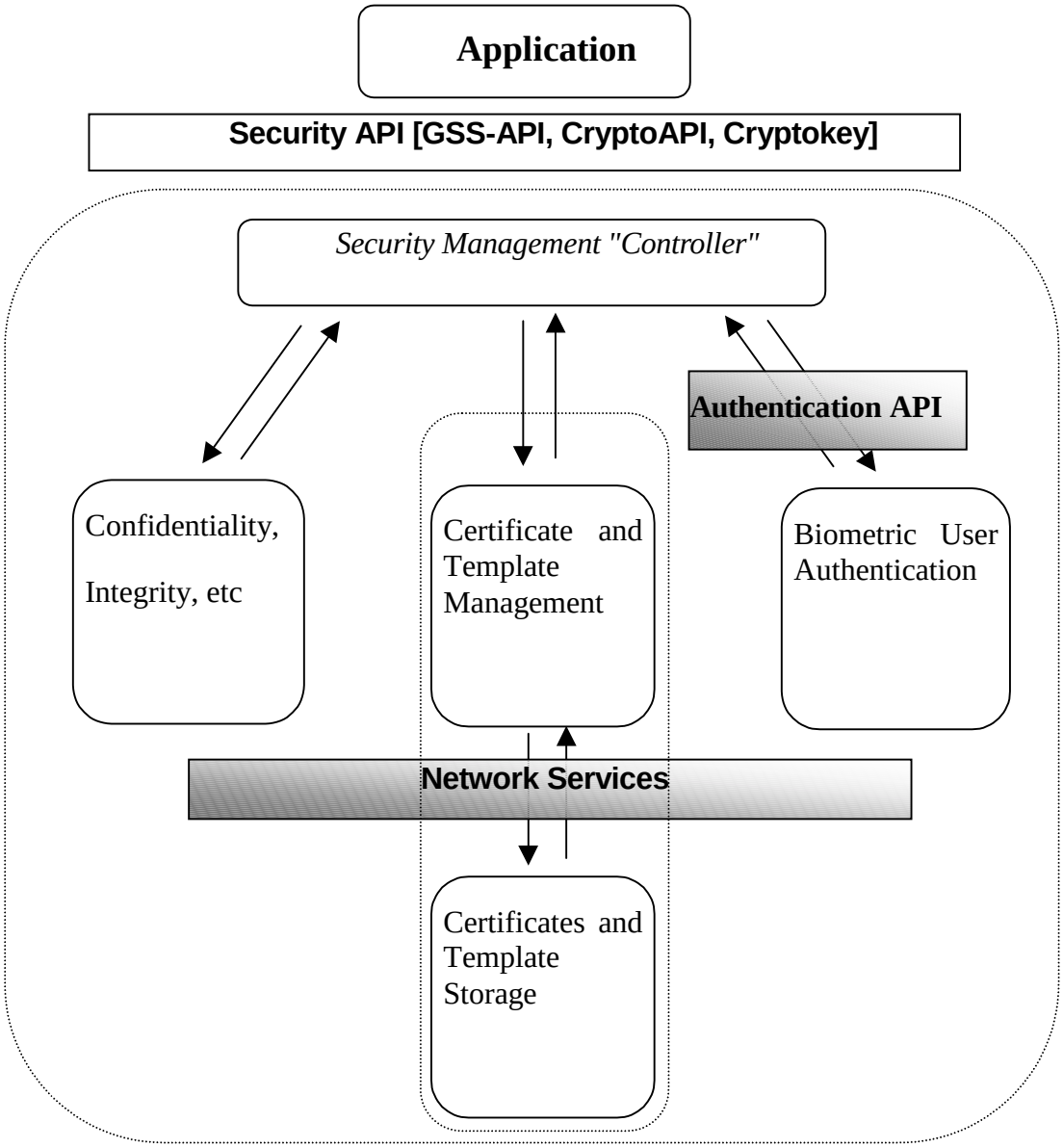
As previously described, the storage and protection of the biometric data, both raw data and templates, may well be one of the most important aspects of the API design. It appears current API designs allow biometric templates to pass across the API boundary. Ideally, the biometric template, similar to cryptographic private keys, should stay within some defined security perimeter to prevent compromise.

G. Cryptographic Support

One of the many considerations that drives an API design is the required cryptographic capability. The API may allow for encryption of transmitted and/or stored biometric data, as well as the digital signing of other types of data, such as the decision response, for the purpose of authenticating the source.

While these areas are of vital concern, the trade-off is in not growing an API too encompassing. Rather than "reinventing the wheel", existing cryptographic and operating system services should be utilized.

A few years back, a colleague of ours, Scott Reider, summarized an architecture where authentication and cryptographic services both fell under a security management controller. This fence provided the boundary for all information security services. We reproduce this figure, which depicts the parallel architecture of the authentication services with the other cryptographic services. This system design remains as pertinent today as it was a few years back.



H. GUI Architecture

Lastly, we briefly examine the graphic user interface. The trade-off in the API design concerns which level(s) is responsible for its management, control and "look-and-feel". Since the biometric device may need to interact with the user of the device, some biometric service middleware may require or allow various prompts, windows, feedback, images, etc or may not require any at all. Benefits exist at all levels for GUI control.

The application developer may desire to strongly integrate the biometric verification or capture with the application. In this case, the application should present the GUI, especially if the application is graphically oriented or multi-media. The biometric service middleware or device driver knows all the device-specific controls, so the GUI might be best developed at this level. Perhaps the API itself, while abstracting

different biometric technologies could also control the GUI, upon application calls for capture, enrollment or verification services.

V. Trade-Off Discussion

We have heard mention of four or five proposals/developments of biometric APIs. However, only two have been released publicly as of this writing. Table 1 provides a comparison of the four we have inspected.. As can be seen, each of the four have unique benefits and contribute to the functionality offered to application developers.

VI. Conclusion/ Recommendation

With the growth of network services and, especially, electronic commerce, the need for information security services grows. Reliance on commercial cryptographic products and solutions will be central in the computer and network industry. In the past few years, a similar development of Cryptographic APIs (CAPIs) has been occurring. Modularity and levels of abstraction (called "levels of cryptographic awareness") have been the cornerstone in these multiple efforts. User authentication is supported in most of the CAPIs, often as a prerequisite for use of private keys. Currently, this user authentication is based on PINs or passwords. Biometric authentication could well be integrated with existing CAPI services as a new auxiliary service module. Until such time, the biometric APIs being proposed should reflect compatible design principles, and track CAPI auxiliary service trends.

Much publicity has been given to biometric APIs since the initial briefings in December of 1997 at the U.S. Biometric Consortium and Commercial Biometric Developers Consortium meetings. Obviously, the final form of standards and specifications are difficult to predict from initial briefings, but hope that the final designs consider the broadest range of biometric technologies, while maintaining the greatest degree of flexibility and ease-of-use for the end-user.

Table 1: Synopsis of API functionality. HA-API and the AIS API are publicly available, though others have been proposed and developed.

Biometric Functionality	HA-API (1.04)	AIS C API (1.01)	TwoOther Proposals
Device Enumeration, Control and Configuration	✔	✔	-App password needed at API startup -Exclusive device control -Set Reader/Engine pairs -Config readers&engines
Data Capture - Acquire live Data	✔	-Check image quality, set quality thresholds -Get image format -Inquire formats	-Device online checking -Return handle to data, not data -Estimate quality
Processing and Feature Extraction	-Pointer to bio structure -Record is vendor specific, with length and header	-Pointer to bio structure	-Return handle to data, not data API manages data -Process options
Enroll, reenroll, and Adaptation	-Ability to use old enrollment raw data -Engine or app initiated adaptation	-Enrollment adds bio data to DB(s) -Check existing matches when enrolling -ReNew function	✔ -No specific adaptation support
Recognition - Match, Verify, Identify	-Engine initiated adaptation -Multiple scores -ONLY verify	-Identify & verify -Return match Unique IDs & score	-Identify & verify -Return match handles and score
Bio Data Storage	✕ - Not part of API	Data mgmt utilities -delete, add, retrieve	-Enumerate, get, delete, save, read & write data in API DB -Account manager accesses various NT DBs (SAM) -Convert from templates to binary blobs -associate bio data with user account
Cryptographic support	✕ -Not part of API	✕ -Not part of API	-Secure internal API storage of bio data
GUI Architecture	Engine (bio vendor) controls all GUI	API GUI calls -create, destroy, freeze, capture windows	-API supports enroll & verify dialog boxes -Device driver provide GUI controls

Best Practices in Testing and Reporting Performance of Biometric Devices

U,K. Biometric Working Group
Version1.0

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.

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.

Some Definitions

"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 "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.

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.

---

[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.

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

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.

"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.

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].

"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

---

[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.

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.

"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.

"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.

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.

"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

---

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".

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.

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.

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

---

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

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 understands these issues must be signed, then maintained in confidence by the researchers. A sample consent form is included as Figure 4.

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.

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

---

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.

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 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

---

[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.

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.

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.

---

[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.

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 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.

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.

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.

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.

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.

Multiple Tests

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.

---

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.

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.

Operational Evaluations

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

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

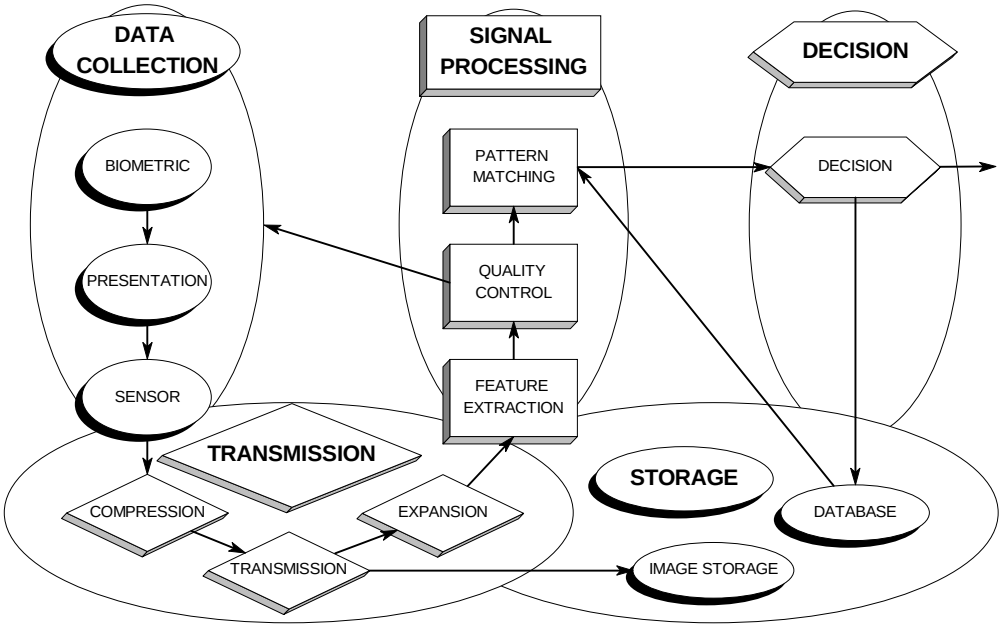


Figure 3 Diagram of General Biometric System

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 4 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.

When Bad Science Leads to Good Law: The Disturbing Irony of the *Daubert* Hearing in the Case of *U.S. V. Byron C. Mitchell*

James L. Wayman, Director
U.S. National Biometric Test Center

In my opinion, if a significant portion of one of your fingerprints is found at a crime scene, you had better be able to; 1) explain its presence; or 2) prove you were already in jail at the time the crime was committed. But I'm a scientist, not a fingerprint examiner, so I'm not paid for my opinions on these matters. Rather, I'm paid to apply the tools of science to test hypotheses such as, "No two individuals have any fingerprints, or portions of any fingerprints, in common". Proving or disproving this is really hard, because we scientists don't have access to all fingerprints from all the world's people. Consequently, we may have to use "statistical estimation". By using the word "statistical estimation", instead of the more realistic word, "mathematically-based guessing", we're hoping that most people will treat us with authority, like people used to treat physicians who actually made house calls, and not dispute these guesses. Certainly, statistical theory, when carefully and scientifically applied, can illuminate great areas of knowledge. But the forms and terminology can easily be misapplied to disguise crazy guesses and opinions. If you are a judge or serving on a jury, and I am an expert witness, I might be able to disguise my guesses with enough bogus "statistical estimation" techno-speak that you won't question them at all, even if they're absurd.

Before we can apply this erudite "statistical estimation" to fingerprinting, we must sharpen the hypothesis. In this case, exactly what do we mean by the words "fingerprint", "portion" and "in common"? "Galton ridges" are the line-like structures on the skin of the palm side of the finger past the distal (the last) joint. These structures may also include pores and will show signs of cracking, abrasion and scarring, depending upon how rough we have been on our hands recently and over the years. So the appearance of these structures is changing over time on all of us. Except on cadavers, scientists don't actually have these Galton ridges to compare and experiment with, only approximate images of these structures, called "fingerprints", perhaps acquired by rolling an inked finger on paper, or better yet, with an electronic scanner of limited resolution. So now our hypothesis is, "No two individuals can have any fingerprint images, or portions of any fingerprint images, in common at any single time".

We still haven't defined the words "portion" and "in common". The lack of a precise meaning for these terms, and the gross misuse of "statistical estimation" leading to absurd guesses about the likelihood of an error, are the central problems with the recent government testimony in the *Daubert* hearing in the U.S. v Byron C. Mitchell case. This hearing took place in September in U.S. District Court in Philadelphia. Putting aside, for the moment, the problem of defining "portion" and "in common", our hypothesis about fingerprints can easily be proved **false**: If the images are bad enough and the portions small enough from places outside the center of the fingerprint (perhaps only tiny segments of a couple blurry ridges), my images will be "in common" with almost anybody's. This extreme case can be established in our lab. Using good quality images of reasonable size and finger positioning, however, we have done tens of millions of computer comparisons with exceedingly few errors, all which could be resolved by human inspection. The scientific question addressed by the government in the *Daubert*

hearing for the Mitchell case should have been, "What is a reasonable estimation of the chance of an error when comparing fingerprint images of reasonable size, position and quality?". The answer, based on sound science, could have been, "Reasonably low". Unfortunately, the government's answer, disguised in the forms and terminology of "statistical estimation", was absurd.

Daubert v. Merrill Dow Pharmaceutical

The Daubert and Schuller families sued Merrill Dow Pharmaceuticals, claiming that the pre-natal use of a prescription drug had caused their children to be born with serious birth defects. The lower courts had ruled that scientific arguments presented by the families to show that the defects were caused by the drug did not meet the required criteria of "general acceptance" for expert evidence. The U.S. Supreme Court was asked to rule on the requirements for presentation of "scientific" evidence into a court of law. In their 1993 decision (509 U.S. 579), the court found that five conditions should be met for evidence to be admissible as "scientific":

1. The theory or technique has been or can be tested.
2. The theory or technique has been subjected to peer review or publication.
3. The existence and maintenance of standards controlling use of the technique.
4. General acceptance of the technique in the scientific community
5. A known potential rate of error.

Trial judges still retain some discretionary power over what scientific evidence does and does not get presented in a trial. Justice Blackmun, writing for the unanimous Court said, "…the trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable". Justice Rehnquist, although voting with the rest, dissented on this particular point, worrying that the court should not impose on judges "…either the obligation or the authority to become amateur scientists in order to perform that role". So now the above five requirements are the "law of the land" and must be met if evidence is to be introduced into any trial as "scientific".

U.S v. Byron C. Mitchell

In 1998, Byron Mitchell was arrested for robbery. The arrest was supported by the apparent match of his fingerprints with small portions of two fingerprints found on the getaway car. His public defenders argued that fingerprint comparison techniques did not meet the five criteria for admissibility established by the U.S. Supreme Court in the *Daubert* decision, particularly the fifth: that the potential rate of error is known. The Mitchell defense petitioned the court for a *Daubert* hearing to determine the admissibility of fingerprint match as "scientific" evidence. The government defense of fingerprinting was lead by the U.S. Department of Justice with assistance of government contractors. The hearing began in July, 1999.

The Government's "Statistical Estimation"

Mitchell's fingerprints had been matched by fingerprint experts. There are no data available on the error rates of these experts, but they are widely acknowledged to be very low. Arranging for a test of a suitable size to reveal even one error would be very expensive and time consuming, so the government proposed testing a computer fingerprint matching system instead. Because these systems do not seem to perform as

well as humans, substituting a computer for humans will lead to a higher error estimate, but such "conservative" estimates do make for good science.

To establish an estimate of the chance of an error by the computer system, the government concocted two tests. In the first test, 50,000 fingerprint images were compared to each other. That is, each of the images was compared to all other images, including itself. In computer fingerprint systems, a comparison of fingerprint image A to fingerprint image B leads to a different "score" than the comparison of the prints in reverse order (B to A). Consequently, these 50,000 data points lead to about 2 ½ billion comparisons. The comparison of images to themselves lead, of course, to extremely high scores, which researchers called the "perfect match" score. Because in life fingerprints are always changing, no real comparison of two <u>different</u> images of the same finger will ever yield such a high score. By adopting, as the definition of "in common", the score obtained by comparison of identical images, the government very strongly biased any results in the <u>government's</u> favor.

Now the government did something even worse: They looked at all the scores between different fingerprint images and declared them to follow a "bell curve". There are potentially an infinite number of curves that could fit the data, some better than others. There are simple tests available to show if the "bell curve", or any other curve, roughly fits the data. No such tests, which might have eliminated the "bell curve" assumption, were performed, however. Now, the government simply pulled out a college-level textbook on statistical estimation and, based on the "bell curve" assumption, found the probability of two different prints being "in common", as previously and unreasonably defined, to be one in 10^{97}. This number, 10^{97}, is extremely large. We have no word for this number in any language, as it is beyond human comprehension. In the entire history of mankind, there have been only about 10^{11} fingerprints. It is possible that in the entire future of all mankind there will never be 10^{97} fingerprints. Yet, the government is comfortable with predicting the fingerprints of the entire history and future of mankind from a sample of 50,000 images, which could have come from as few as 5,000 people. They have disguised this absurd guess by claiming reliance on "statistical estimation".

There was an additional logical problem that the government needed to address: The crime scene fingerprint images, called "latent prints", showed only a small portion of the finger. So to test the error rate for latent prints, the government researchers artificially cropped the size of the original 50,000 images, in effect changing the position of the finger in the images. The precise way in which this is done could have profound impact on the projected error rates, but the government doesn't reveal exactly their method. Further, the latent prints in the Mitchell, or any police, case would have been naturally "cropped" in a completely different way. The government's laboratory research gets quite sketchy at this point, but in court, the government claimed error rates between 1 in 10^{27} and 1 in 10^{97}**,** presumably using the same flawed methodology as in the first test. The government did not try to run any real crime scene prints against the same 50,000 database to determine comparison scores and establish error probabilities for latent prints in real cases.

In short, nothing in the government study or testimony gives us any indication of the likelihood that the crime scene fingerprints were falsely identified as belonging to the defendant Mitchell or, more broadly, that any latent fingerprints might be falsely

identified. In my opinion, the government failed completely to answer the fundamental question, "What is a reasonable estimation of the chance of an error when comparing fingerprint images of reasonable size, position and quality?" They could have done so simply by designing better experiments. If we had a good answer, we'd only have to establish that the crime scene prints in the Mitchell, or any, case were of reasonable size, position and quality to roughly estimate the possibility of error.

In fact, false fingerprint matches are not unknown and have been introduced as faulty evidence in criminal trials. See www.onin.com/fp for details of such occurrences in Illinois and Scotland.

Probability and Statistical Estimation in Legal Cases

There is a history in American juris prudence of human identification based on the gross misuse of statistical and probability theory. In the famous 1968 *People v. Collins* case, Malcolm and Janet Collins were convicted of robbery based on the testimony by a college math instructor that the chances of some other couple committing the crime was 1 in 1.2 million. The decision was reversed by the California Supreme Court on the grounds that the probability-based arguments were without foundation, and erroneous and misleading to the point of distracting the jury. Writing about the case in 1969, University of Houston Law Professor Alan D. Cullison, states "…it would be unsound for courts to reject expert probability testimony on the basis of the invalidity of probability theory itself…A more cogent basis for broadside objection to expert probability testimony is that the applications of probability theory to fact-finding problems in law cases have in the past been, crude, misleading and often just plain erroneous."

More recently, questionable use of probability theory in human identification has involved forensic DNA analysis. Referring to disagreements in National Research Council (NRC) studies of DNA analysis error rates, UC Berkeley Statistics Professor Peter Bickel (current chair of the NRC's Committee on Applied and Theoretical Statistics and a member of the National Academy of Science) writes in the *Proceedings of the National Academy of Science*,

> The existence of two reports (1992 and 1996), close in time, which disagree on aspects of methodology illustrates what scientists have always known but what the law sometimes wishes to ignore: that scientists can differ in their expert judgment of the accuracy of numbers predicted from data by model-based formulae. In this case, the focus of the disagreement is on the question of the extent to which models of population genetics can be applied in estimating the probability that the DNA of the suspect and DNA found on the victim match perfectly at each and every one of the preselected set of loci. This probability has to be computed under the assumption that the match occurred "by chance alone". That assumption is not enough to allow us to compute or rather estimate this probability. To finally arrive at a formula, further assumptions are made: treating the FBI and other databases effectively as random samples from the relevant population and, more significantly, that (certain statistical independence assumptions) are satisfied or are perturbed in a correctable way. Given that no laboratory error has been committed, there is, I believe, little disagreement between the committees or within the scientific community that the match probabilities referred to above are small, typically of order smaller than 1 in

1,000. But many scientists would not agree that the modeling assumptions made above can be verified to hold so precisely that the match probabilities can be ascertained to an order of 1 in a billion.

Prof. Bickel's arguments about error rate estimation in DNA analysis apply equally well to our discussion of fingerprinting. I am of that group that do not agree that the required assumptions about fingerprints hold so precisely that error rates on the order even of 1 in a billion can be ascertained, let alone 1 in 10^{97} .

Conclusions

In September, the U.S. Court of Appeals released their findings in the Daubert hearing of the U.S. v. Mitchell case, holding that fingerprinting meets the necessary criteria for admissibility as evidence. This is the correct decision. Fingerprinting is an established science, subjected to peer review and publication, with general acceptance and standards for its practice. Error rates are difficult to measure, precisely because they are so low. So I am pleased with the outcome. I'm saddened, however, that the government's case had to rest on such shoddy science. I'd certainly prefer to see good law resulting from good science. We must strive to do better.

The Federal Legislative Basis for Government Applications of Biometric Technologies

James L. Wayman, Director
U.S. National Biometric Test Center

Introduction

The electoral mood in the United States is that government must deliver more services to an increasing population in a more efficient, cost effective, and fraud-free manner, while limiting the size and scope of the governmental infrastructure. Encouraged or mandated by federal legislation, governmental agencies at all levels have turned to technology in an attempt to meet these competing requirements.

The direct delivery of government services to citizens inextricably requires human identification, both positive and negative: positive identification for efficiently preventing multiple persons from using a single identity; and negative identification to effectively prevent a single person from using multiple identities. Consequently, automated means of human identification (biometrics) are being rapidly introduced into governmental processes. Driver's licensing agencies are using fingerprinting and facial imaging to verify identities of those applying for or renewing driver's licenses. The Immigration and Naturalization Service is using voice recognition and hand geometry to speed up border crossings. Social service agencies are using fingerprinting to verify that benefits are being disbursed to enrolled persons while preventing multiple enrollments of the same person.

Positive identification does not require biometrics. I can prove who I am by supplying other forms of identification, such as a birth certificate, driver's license or utility bill. Or, I can prove who I am through a shared secret, such as a password, PIN, or my mother's maiden name. Although biometric methods are not foolproof (errors can and routinely do occur), identification can generally be done more rapidly and with less human intervention than with the use of documents. Biometric methods can compete in speed with password and PIN entry and may be more convenient. Further, unlike documents and secret knowledge, biometric measures are not transferable.

Negative identification can only be done with biometrics. No document or password can establish that I do not have multiple identities, so in government applications where negative identification is required (social service and driver's licensing applications) there is no reasonable alternative to biometric identification.

Some people, however, are concerned with the potential impact that government use of these technologies might have on personal freedoms. Perhaps it is the very personal nature of biometric identification, in contrast to the more impersonal nature of the alternatives, that raises concerns over its use by the government. Perhaps it is the fear that government will begin the "real time" tracking of the movements of individuals[1]. Those that propose, develop and implement biometric authentication technologies for government applications are sympathetic to these concerns. Vigilance with regard to the protection of personal liberty is always necessary and appropriate in a free society.

---

1 We would argue that biometric-based tracking will never be as effective as current methods of tracking though subpoenaed credit card and telephone records.

All current and proposed government implementations are "tactical", not "strategic" in scope, meaning designed to address specific problems in a limited way. Under considerably pressure from victim's groups, Congress has created some non-biometric residency tracking systems for some classes of people, such as sexual predators[2], criminal aliens[3], and "dead-beat" parents[4]. These systems do not perform "real-time" tracking of the activities of individuals, only the tracking of residency for the purposes of alerting neighborhoods or local authorities. There is no current or proposed national database of biometric identifiers of the general, non-criminal population

This paper will summarize the current federal legislation driving the use of biometric authentication in several key government sectors. Forensic and generally non-automatic forms of human identification for solving crimes and apprehending criminals are beyond the definition of biometric authentication as "automatic identification of individual humans based on behavioral and physiological characteristics" and will not be discussed in this paper

Drivers Licensing

Under the U.S. Constitution, the federal government has the right to "regulate Commerce …among the several States"[5]. Using this power, the Federal Highway Administration establishes licensing requirements for inter-state commercial truck drivers. Licensing of the rest of us is a power reserved to the States. Therefore, use of biometrics can be specified at the federal level only with respect to commercial drivers. At the State level, the American Association of Motor Vehicle Administrators (AAMVA) recommends voluntary "best-practices" and reciprocity policies.

Commercial Licenses

Whenever there is a fatal accident involving a commercial truck driver, the press and public rightfully show strong concern, demanding stricter licensing requirements, more careful tracking of driver violations, and more certain identification of drivers[6]. Congress has been interested in the biometric identification of commercial drivers since 1988. In fact, the first piece of federal legislation referring directly to biometrics was the 1988 "Truck and Bus Safety and Regulatory Reform Act"[7] (TBSRRA). Section 9105 of this act required the Secretary of Transportation to develop "minimum uniform standards for the biometric identification of commercial drivers". The stated goal of the legislation was to facilitate enforcement of the "one-driver, one-license, one record" provision of the

---

2 Public Law 104-236

3 P. L. 104-208, Division C, Title 1, Sec. 327

4 P.L.104-193, Title II, Subtitle B

5 Article I, Section 8, Clause 3

6 See, for instance, Colbert I.King, "Who's at the Wheel of the Motor Vehicles Bureau", The Washington Post, November 8, 1997, page A25. Also, "Driver who caused 101 crash had long list of convictions", San Jose Mercury News, October 12, 1995

7 Public Law 100-690

1986 "Commercial Motor Vehicle Safety Act"[8] (CMVSA). This latter act created the Commercial Drivers License Information System (CDLIS), allowing states to track driving violations of interstate truck drivers and to verify that such drivers were not holding multiple licenses. In passing the 1988 TBSRRA, Congress was responding to concern from the American Trucking Association that, without biometric identification, CDLIS might not be able to prevent commercial drivers from evading legal action for multiple violations by carrying multiple driver's licenses. It appears that the intent of Congress was a system for negative identification. The legislation is not clear regarding the inclusion of a positive identification function, linking drivers to their licenses through biometric identification.

Although the TBSRRA allocated $1.5 million for a pilot demonstration project, no standards for biometric identification were adopted. The Federal Highway Administration (FHWA) concluded that "more time is needed so that the technology has an opportunity to develop"[9]. In 1995, San Jose State University (SJSU) received a contract to revisit the earlier work and advise FHWA on the continued development of these standards. In December of last year, the final report of SJSU to FHWA was delivered. This report is available for downloading at www.engr.sjsu.edu/biometrics/fhwa.html.

After consultations with FHWA and AAMVA, three criteria were established for the selection of the preferred biometric technology: 1) vendor support for both negative and positive identification functions; 2) previous use in similar large-scale application for which an independent performance/cost audit is available; 3) available from multiple vendors supporting single standards. On this basis, fingerprinting was recommended as the standard biometric identifier. The report recommends accepting the Criminal Justice Information Service's (CJIS) standard for scanner image quality, the CJIS standard for image compression, and the ANSI/NIST standard "Data Format for the Interchange of Fingerprint Information".

We expect that the FHWA will accept the recommendations of the San Jose State University study and adopt these standards for the biometric identification of commercial drivers. The adoption of standards, however, is not the same as implementation of a system. Absence of conclusive data on the existence of a problem in identifying commercial drivers will probably lead the FHWA to accept the standards, but not to fund, or advocate for the creation of, a system of biometric identification of commercial drivers.

In June of 1998, Congress passed and the President signed the "Transportation Equity Act for the 21st Century"[10] which states[11] that each Commercial Drivers License issued

---

8 Public Law 99-570, Sections 12007 and 12009

9 "Minimum Uniform Standards for a Biometric Identification System to Ensure Identification of Operators of Commercial Motor Vehicles", Advanced Notice of Proposed Rule Making: Additional Information, Federal Register, Vol. 56, No.46, March 8, 1991, pg. 9925-9928

10 Public Law 107-178

11 Section 31302

after January 1, 2001, must "include unique identifiers (which may include biometric identifiers) to minimize fraud and duplication". The assignment of identification numbers would meet the requirements of this bill. We do not expect the permissive language regarding biometrics to cause any policy change in commercial driver's licensing at state or federal levels. This bill does effect the operation of CDLIS, the national tracking system developed for commercial drivers under the 1986 CMVSA.

Non-commercial Licenses

The American National Standards Institute (ANSI) is considering standards for State issued non-commercial drivers licenses and identification cards. These standards, known as ANSI B10.8, would be for physical characteristics, layout, security features, and data access, storage, and encryption for state and federally issued cards.

This committee has recommended establishment of fingerprinting, using both left and right forefingers, as the standard. These standards are advisory only, and have no binding power on the states who administer driver's licenses. It is not currently the policy of AAMVA, the organization of state drivers license administrators, that fingerprinting should be made a requirement for driver's licensing in any state.

Five states (CA, CO, GA, HI, TX) require fingerprinting in their driver's licensing programs. One state (WV) makes fingerprinting or facial imaging optional. Two states (AL, FL) have discontinued use of fingerprinting for driver's licensing during the past two years. In May of 1998, the Michigan legislature passed House Bill 4635 prohibiting the requirement of fingerprinting as a condition for the issuance of a driver's licenses in Michigan. Similar bills have been introduced in Washington state[12] and Alabama[13].

The current driver's license in all states carries physiologically-based identification information, such as height, weight, date of birth, hair color, eye color and photo image. These identifiers could be replaced with a less privacy intrusive measure, such as a fingerprint or eye scan.

Immigration

The federal agency with the most extensive use of biometrics is the Immigration and Naturalization Service (INS)[14]. This Department of Justice agency is ultimately headed by the Attorney General, who has been directed by Congress to establish several new programs and systems pertaining to border control and employment eligibility verification.

Border Control

It has long been the intent of Congress that the INS should increase the efficiency of border crossing systems through automation. Section 109 of Illegal Immigration

---

12 HB2730/ SB6399

13 HB123

14 A detailed overview of the biometric identification activities of the INS can be found in Brad Wing, "Overview of All INS Biometrics Projects", Proc. CTST'98, Vol. I, pg.543-552.

Reform and Immigrant Responsibility Act of 1996 (IIRAIRA)[15] states "the Attorney General, together with the Secretary of State, the Secretary of Agriculture, the Secretary of the Treasury, and appropriate representatives of the air transport industry, shall jointly undertake a study to develop a plan for making the transition to automated data collection at ports of entry." The INS has been instituting automatic data collection and immigration control systems based on biometrics for the last several years. These systems do, indeed, track the U.S. border crossings of all persons, as is the clear intent of Congress.

Over 65,000 frequent travelers to the United States (both U.S. and non-U.S. citizens) have voluntarily enrolled in the "Immigration and Naturalization Service Passenger Accelerated Service System" (INSPASS). The INSPASS system is currently in use at 7 airports with expansion to 2 additional airports planned for the near future. The system is also being used for pedestrian travelers at the Mexican border crossing in Hildalgo, Texas. Users of this positive identification system need not wait in lengthy lines to present their passport to immigration officials. Rather they present their INSPASS to a kiosk, which verifies the user's identity through hand geometry. The hand geometry records of all enrolled users are kept in a central database, but this database communicates with no other government systems. The user's immigration status is determined by the passport number, read off the INPASS card. Immigration records are searched and crossings recorded by this passport number, not by the hand geometry sample.

A similar voluntary system, the Automated Permit Port (APP), is used "after hours" at small ports of entry along the Canadian border. This positive identification system uses voice to verify the claimed identities of those crossing the border. The voice verification database communicates with no other government system.

Federal code[16] mandates the administration of the alien border crossing card, a "document of identity bearing that designation issued to an alien who is lawfully admitted for permanent residence, or to an alien who is a resident in foreign contiguous territory…for the purpose of crossing over the borders between the United States and foreign contiguous territory.." Section 104 of the IIRAIRA instructs the Immigration and Naturalization Service that the alien border crossing card shall "include a biometric identifier (such as the fingerprint of handprint of the alien) that is machine readable". The alien border crossing card is now issued by the Department of State (DOS) but made at the Immigration and Naturalization Service's (INS) card facility. The DOS is collecting fingerprints of aliens and taking photographs, both of which are included on the card.

Section 110 of IIRAIRA requires the INS to "collect a record of departure for every alien departing the United States and match the records of departure with the record of the alien's arrival in the United States (to) enable the Attorney General to identify, through on-line searching procedures, lawfully admitted non-immigrants who remain in the United States beyond the period authorized …" The challenge to the INS is to comply with this provision in a cost-effective manner, establishing a system which

---

15 Public Law 104-208, Division C, Title 1

16 8 USC 1101

does not unduly slow the departure of all persons from the U.S. The inherent limitations of biometric technology will require that the arrival/departure records be stored under record numbers, not biometric identifiers. However, biometric identifiers could be used to verify the identity of the holder of the record number, allowing the use of machine-readable cards similar to INSPASS or APP. Consequently, biometric technology is being seriously considered for this application, as well[17].

Employment Eligibility

The IIRAIRA also establishes "pilot programs for employment eligibility confirmation"[18]. The law calls on the Attorney General "to conduct 3 pilot programs of employment eligibility confirmation."[19]: the basic pilot program, the citizen attestation pilot program, and the machine-readable-document pilot program. None of these programs call for biometric identification, or use other terms indicating that biometric identification is required.

All of the pilot programs are voluntary and are intended for persons or other entities which conduct hiring, recruitment or referral for employment. "The Attorney General may not require any person or other entity to participate in a pilot program"[20]. Further, the Act specifically does not establish a national identification card, stating, "Nothing in this subtitle shall be construed to authorize, directly or indirectly, the issuance or use of national identification cards or the establishment of a national identification card."[21]

The first pilot program is "the basic pilot program" which will be conducted "in, at a minimum, 5 of the 7 States with the highest estimated population of aliens who are not lawfully present in the United States". This goal of this program is to verify that "the person with the identity claimed by the individual is authorized to work in the United States, and (to determine if) the individual is claiming the identity of another person"[22]. If a potential employee does not attest to being a U.S. citizen, employers will be required to examine a specified identity document to determine if it reasonably appears to be genuine and to belong to the person presenting it. Employment eligibility will be established by the employer by calling a toll-free phone number.

The second pilot program is "the citizen attestation pilot program" to be conducted "in at least 5 States (or, if fewer, all of the States)" that have a driver's license that "contains a photograph, … security features, and … issued through application and

---

17 Section 112 of IIRAIRA also requires the "nationwide fingerprinting of apprehended aliens" as part of the "IDENT" program already in place. This inherently forensic program will be "expanded to apply to illegal or criminal aliens apprehended nationwide".

18 P.L.104-208, Title IV, Subtitle A

19 P.L.104-208, Section 401.

20 Ibid, Section 402

21 Ibid, Section 404(h)(2)

22 Ibid, Section 403(a)(2)(B)

issuance procedures, which make such document sufficiently resistant to counterfeiting, tampering, and fraudulent use..."[23]. Persons requesting employment can present this driver's license or attest through written means that they are legally eligible for employment within the United States.

The third pilot program, the "machine-readable-document pilot program" will be conducted in "at least 5 States (or, if fewer, all of the States) whose driver's license "contains a machine-readable social security number"[24]. This system will allow an employer to determine automatically via a data link whether a person presenting the driver's license is eligible for employment.

Employers participating in a pilot program will not "be civilly or criminally liable under any law for any action taken in good faith reliance on information provided through the confirmation system"[25].

In short, Congress has reacted to the demands of workers to limit employment of unauthorized aliens, and the demands of employers to limit liability for their inadvertent hiring. The proposed pilot programs do not call for biometric identification and, further, specifically do not to create a national identity card.

.Welfare

The Personal Responsibility and Work Opportunity Reconciliation Act of 1995 (Welfare Reform Act) does not mention biometric identification, but calls for the states to use "the most recent technology available that the State agency considers appropriate and cost effective and which may include personal identification numbers, photographic identification on electronic benefit transfer cards, and other measures to protect against fraud and abuse[26]". In response to this requirement, 34 states and the District of Colombia have instituted electronic benefit transfer (EBT) cards to deter food stamp "trafficking" (the illegal sale of food stamps for cash), estimated to be a $815 million per year problem nation-wide[27]. With the client use of EBT cards, food stamp redemptions can be electronically tracked, with all transactions recorded by store and client number. This system does not include the use of biometric identification, apart from a pilot project in Texas. In 1996, the Food and Nutrition Service (FNS) of the Department of Agriculture (USDA) and the Texas Department of Social Services jointly financed this pilot project for fingerprint identification of food stamp applicants in the San Antonio area. This project was for negative identification only, to establish that single individuals were not receiving multiple food stamp benefits. Although the cost/benefit analysis of the project was inconclusive, it will be tentatively extended on a county-by-county basis with additional assessments to determine government savings through fraud-deterrence. There

---

[23] Ibid, Section 403(b)(2)(A)

[24] Ibid, Section 403(c)(2)

[25] Ibid, Section 403(d)

[26] P.L. 104-193, Section 825.

[27] Estimate for 1993, from T.F. Macaluso, "The Extent of Trafficking in the Food Stamp Program", USDA, Food and Consumer Service, Office of Analysis and Evaluation, August , 1995

seems to be no current willingness to extend the tracking system to include biometrics at "points of sale", requiring that food stamp presenters prove through biometric means that they are the authorized holders of the stamps.

To date, eight states[28] have established operational biometric systems for the delivery of human services[29], including general welfare payments. Three more states (NC, PA, FL) have systems pending. All states are using finger imaging, although facial imaging, retinal scanning and hand geometry have also been used in some states. New York, New Jersey and Connecticut do exchange fingerprint images to detect interstate fraud, but this inter-operability is not directly mandated by the federal government. The other States using biometric identifiers in their human services programs are not networked and do not interchange data.

States using fingerprinting have adopted the two index fingers as the standard, and are generally compliant with the ANSI/NIST standards for scanner image quality and compression. However, the States with fingerprinting systems have procured these from a variety of vendors, each using proprietary feature extraction and matching algorithms, so merging of all the current systems to produce a national system would not be immediately feasible.

Section 408 of the Welfare Reform Act limits the life-time welfare eligibility of families to 60 months, total, in "any State program funded under this part attributable to funds provided by the Federal Government". Yet, the federal government has not specified or funded a mechanism for enforcing this provision. It has been suggested within the social service community that a national system of biometric identification might be required to fully comply with the legislation, but such a system would require additional Congressional action. Biometric characteristics change and no biometric technology (automatic by our definition) has ever been shown to be useful over the life-time of an individual. No biometric system for the enforcement of this section of the Welfare Reform Act is currently under development.

Airport Security

In August of 1996, President Clinton created the "White House Commission on Aviation Safety and Security", headed by Vice President Al Gore. The purpose of this "Gore Commission" was to " advise the President on matters involving aviation safety and security … (and) develop and recommend to the President a strategy designed to improve aviation safety and security, both domestically and internationally." [30]

Among the recommendation of the Commission were:

---

[28] AZ, CA, CT, IL, MA, NJ, NY, TX

[29] Extensive listing of these states and their programs can be found in D. Mintie, "Overview of Biometric Applications', *Proc. CardTech/Securtech'99, M*ay 1999, pp.103-113. See also the Connecticut Department of Social Services web site at www.dss.state.ct.us/digital.htm.

[30] Executive Order 13015 of August 22, 1996

"Access to airport controlled areas must be secured and the physical security of aircraft must be ensured"[31].

A bag-passenger match system should be implemented[32].

Federal Aviation Reauthorization Act of 1996[33] spoke directly to both recommendations, calling for the Federal Aviation Administration (FAA) to issue a report recommending enhancements to the screening and inspection of air cargo[34] and giving the "Sense of the Senate" that airports and air carriers should implement domestic bag matching[35].

Employee Access

The FAA has, for many years, required systems to be in place for controlling access to secure areas, even for airport employees. Federal Aviation Regulation 107.14 specifies that all major airports have security systems in place to control access to secure areas, to restrict access to particular portions of the airport for even those with access, to immediately deny access when authorization changes, and to have the capability of limiting access by time and date. [36]. Biometric devices have logically been employed in this application. In fact, San Francisco International Airport has been using hand geometry to control access by airport employees to secured areas for several years, as have individual airlines, such as United.

Federal Aviation Regulation 107.13 requires that access by ground vehicles to and movement of persons within air operations areas be controlled as well. Again, biometric identification seems to be a logical technology in these applications. This year, the Federal Aviation Administration , in cooperation with the Chicago Port Authority and the American Trucking Association, began a pilot project for the fingerprint identification of commercial drivers entering the air cargo area of O'Hare Airport.

Passenger-Baggage Matching

The Gore Commission report recommends that the FAA:

> ***"Begin implementation of full bag-passenger match.*** Matching bags to passengers ensures that the baggage of anyone who does not board the plane is removed. Full bag match ensures that no unaccompanied bag remains on board a flight. Manual and automated systems to conduct full bag match have been employed in international aviation for several years, but need additional work to ensure they can be phased into domestic airline operations. The Commission recommends implementing full bag match at selected airports, including at least

---

[31] Ibid, recommendation 3.11

[32] Ibid, recommendation 3.24

[33] P.L. 104-264

[34] Ibid, Section 313

[35] Ibid, Section 311

[36] Ibid, Section 311(b)

one hub, within sixty days to determine the best means of implementing the process system-wide."

As noted, the Federal Aviation Reauthorization Act of 1996 supports this recommendation and states that "if a bag match pilot program is carried out…. The (FAA) shall submit to Congress a report on the safety, effectiveness, and operational effectiveness of the pilot program. The report shall also assess the extent to which implementation of baggage match requirements (coupled with the best available technologies and methodologies, such as passenger profiling) enhance domestic aviation security."[37]

To our knowledge, the FAA has not yet begun the pilot program. Again, biometric technology, allowing the automatic recognition of a passenger as she/he boards the airplane, coupled with a method for baggage identification, such as RF tagging, could allow this function to be performed in a fully automated manner, with no inconvenience or delay to the traveling passengers.

Conclusions

State and federal governments, in response to numerous Congressional mandates, and ultimately, the demands of the electorate, are using or have proposed the use of biometric identification in a variety of limited systems with limited goals. This paper has considered many of these systems. None of these systems involves tracking the day-to-day movements of citizens. "Non-real-time" systems of residency tracking of sexual predators, criminal aliens, convicted felons, and "dead-beat" parents, have been mandated by Congress, but none of these systems involve the use of biometric identification as we use the term. There are no large-government databases of biometric identifiers, other than the criminal fingerprint databases maintained by the FBI, and there is no interest within the U.S. government of creating such databases.

The government's interest in biometric technologies is motivated by the desire to improve the delivery of services to citizens by increasing efficiency and convenience, while decreasing costs and fraud. The implementation of biometric technologies can represent a reasonable solution to difficult problems.

---

[37] Ibid, Section 311(a)

Biometric Identification Technologies in Election Processes—Summary Report

James L. Wayman, Director
U.S. National Biometric Test Center

Biometric technologies, allowing the automatic identification of people using voice patterns, eye scans, handwriting style, faces, hands or fingerprints, have been suggested for use in the election process for eliminating fraud. Fingerprinting, hand shape and eye scanning have been used in the United States in driver licensing and social service programs. Fingerprinting systems are being introduced into the election process in several countries, such as the Philippines, Jamaica, Argentina and Cambodia. What are the prospects for introducing these technologies into our voting systems?

We will look at the possible voting applications in this paper and conclude that biometric technologies could be effectively used, even on a voluntary basis, to detect and deter voting fraud. However, this use would require fundamental changes in the way we register voters and would necessitate the creation of government-controlled databases of physical and behavioral characteristics of at least some voters. Although such databases are inherently "fuzzy" and far less threatening to personal privacy than phone books or driver's licenses, changes in voter registration procedures to enable biometric data collection could be seen as contrary to the intent of the National Voters Rights Act of 1993 and would likely require enabling federal legislation.

What is Biometric Identification?

Biometric technologies use physical characteristics, such as voice tone or hand shape, to identify people automatically. Behaviors, such as handwriting style, can also be used by computers in this way. The term "identify" is used here quite loosely. There is actually nothing in your voice, hand shape or any biometric measure to tell the computer your name, age or citizenship, or to establish your eligibility to vote. External documents (passport, birth certificate, naturalization papers) or your good word establishing these facts must be supplied at the time you initially present yourself to the biometric system for "enrollment". At this initial session, your biometric characteristic, such as an eye scan, is recorded and linked to this externally-supplied personal information. At future sessions, the computer links you to the previously supplied information using the same physical characteristic. Even if the biometric system works perfectly, the personal data in the computer, such as your voting eligibility, is only as reliable as the original "source" documentation supplied.

Once the computer knows your claimed identity, it can usually recognize you whenever you present the required biometric characteristic. No biometric identification system, however, works perfectly. Problems are generally caused by changes in the physical characteristic. Even fingerprints change as cuts, cracks and dryness in the skin come and go. It is far more likely that the computer will not recognize your enrollment characteristic than link you to the characteristic of someone else, but both types of errors do occur. To minimize the possibility that you will be linked to another record, "positive identification" systems ask you to identify yourself. Your biometric characteristic is then compared to the characteristic stored at the time you enrolled. Biometric measures are always "fuzzy" to some extent, changing over time and circumstance of collection. If the

submitted and stored biometric measures are "close enough", it is assumed that you are indeed the person enrolled under the identity you claimed. If the presented and enrolled characteristics are not "close enough", you will generally be allowed to try again. If multiple attempts are allowed, the number of users "falsely rejected" can be under 1%, although there are always some people chronically unable to use any system who must be given alternate means of identification. The possibility that an impostor will be judged "close enough", even given multiple attempts, is usually less than one in ten. The threat of being caught in 9 out of 10 attempts is enough to deter most impostors, particularly if penalties for fraud are involved.

Positive identification using biometrics can be made totally voluntary. People not wishing to use the system can instead supply the source documents to human examiners each time they access the system.

Many biometric methods have been used in public systems for "positive identification": hand and finger geometry, iris and retinal scanning, voice and face recognition, and fingerprinting.

TABLE 1: IDENTIFICATION: "POSITIVE" AND "NEGATIVE"

POSITIVE	NEGATIVE
To prove I am someone known to the system	To prove I am not someone known to the system
Comparison of submitted sample to single claimed template	Comparison of submitted sample to all enrolled templates
Alternative identification methods exist	No alternative methods exist
Can be voluntary	Must be mandatory for all
Biometric measures linked to personal information (name, age, citizenship) only through external source documents.	Linkage to personal information not required.

There is a another way some biometric systems can be used: "negative identification". In these applications, found in driver licensing and social service eligibility systems where multiple enrollments are illegal, the user claims not to be enrolled. Apart from the "honor" system, where each person's word is accepted, there are no alternatives to biometrics for negative identification.

During enrollment, the system must compare the presented characteristic to all characteristics in the database to verify that no match exists. Because of the ongoing changes in everyone's body, errors can occur in the direction of failing to recognize an existing enrollment, perhaps at a rate of a few percent. But again, only the most determined fraudster, unconcerned about penalties, would take on a system weighted

against him/her with these odds. False matches of a submitted biometric measure to one connected to another person in the database are extremely rare and can always be resolved by the people operating the system.

Negative identification applications cannot be made voluntary. Each person wishing to establish an identity in the system must present the required biometric measure. If this were not so, fraudsters could establish multiple enrollments simply by declining to use the biometric system. On the other hand, negative identification can be accomplished perfectly well without linkage to any external information, such as name or age. This information is not directly necessary to prove you are not already known to the system, although it may be helpful if identification errors occur.

Only two types of biometric methods have ever been used in this way in a public application: electronic fingerprinting and retinal scanning. Table 2 connects the technologies to publicly demonstrated systems.

TABLE 2: TECHNOLOGIES DEMONSTRATED IN PUBLIC SYSTEMS

POSITIVE IDENTIFICATION	NEGATIVE IDENTIFICATION
Hand geometry	Fingerprinting
Finger geometry	Retinal scanning
Voice recognition	
Iris scanning	
Retinal scanning	
Facial imaging	
Fingerprinting	

Will Biometric Identification Compromise My Privacy?

Whenever biometric identification is discussed, people always want to know about the implications for personal privacy. If I use a biometric system, will the government, or some other group, be able to get personal information about me? Biometric measures themselves contain no personal information. My hand shape, fingerprints or eye scans do not reveal my name, age, race, gender, health or immigration status. Although voice patterns can give a good estimation of gender, no other biometric identification technology currently used reveals anything about me as a person. More common identification methods, such as a driver's license, reveal my name, address, age,

gender, vision impairment, height and even weight! Unlike driver's licenses, however, biometric measures cannot be stolen or counterfeited.

The real fear is that my biometric measures will link me to my personal data, or allow my movements to be tracked. After all, credit card and phone records can be used in court to establish a person's activities and movements. There are several important points to be made on this issue.

Only biometric mesurements which I have surrendered to a system through 'enrollment' will be known to that system. If I have never enrolled (given my fingerprint with supporting identificaiton documentation) in a fingerpringt system any use I make of a fingerprint system cannot be linked to 'me' (my identity).

Biometric measures cannot generally be taken without my knowledge, so I cannot be enrolled in any system without my participation. Exceptions are face and voice patterns, which can be taken without my knowledge. "Latent" fingerprints left on surfaces can be "lifted" by those trained in investigative techniques, but such prints are generally not of a quality suitable for enrollment purposes in electronic systems.

Phone books are public databases linking me to my phone number. These databases are even accessible on the Internet. "Reverse" phone books also exist allowing my name to be determined from my phone number. Even if I have an unlisted number, my number and all information on calls made from that number may be available to law enforcement agencies through the subpoena process. There are no public databases, however, containing biometric identifiers, and there are only a few limited-access government databases containing biometric measures. Eight States have electronic fingerprint records of social service recipients (AZ, CA, CT, IL, MA, NJ, NY, TX), five States (CA, CO, GA, HI, TX) maintain electronic fingerprints of all licensed drivers[1], nearly all States maintain copies of driver's license and social service recipient photos, the FBI and State governments maintain fingerprint databases on convicted felons and sex offenders, and the federal government maintains hand geometry records on those who have voluntarily requested border crossing cards. General access to this data is limited to the agencies that collected it, but like credit card and phone "toll records", this information can be released or searched by law enforcement groups acting under court order.

Unlike your legal name and your Social Security, credit card and phone numbers, your biometric measures are rather fuzzy and inexact, being somewhat different every time they are measured. Further, your biometric measures will be rather similar to the biometric measures of others. Consequently, even if you could gain access to a database containing biometric measures, they could not be "reversed" like a phone book to reveal names from identifying numbers. Two technologies, electronic fingerprinting and retinal scanning, have been objectively demonstrated to be exceptions to this reversibility rule, if data is carefully collected from cooperating users. So if you want to discover someone's identity, the best way is with a phone number, not a biometric identifier. If you want

---

[1] WV maintains a voluntary fingerprint database on drivers who wish to use biometric identification.

personal information about someone, start with their name; a biometric identifier will be of no help.

Biometric identifiers in databases of drivers, social service recipients or border crossers, are far less distinctive than the names, addresses and ID numbers also in these databases, and do not allow users to be tracked or monitored like credit card and phone numbers do. For this reason, databases of drivers and social service recipients are always indexed by name or identification number even if they contain a biometric record. Biometric identifiers are nearly impossible to steal or falsify than these other identifiers, allowing protection from identity theft or impersonation. In conclusion, adding a biometric identifier to current voter registration databases would not present any privacy risk to any voter, but could be used to *prevent or deter privacy loss* through identity theft.

TABLE 3: BIOMETRICS AND PRIVACY

1) Unlike more common forms of identification, my biometric measures contain no personal information about me and cannot be stolen or forged.
2) Some biometric measures (face images, voice signals, and "latent" fingerprints left on surfaces) can be taken without my knowledge, but can't be linked to me without a pre-existing database.
3) The federal government maintains a fingerprint database on convicted felons and some State governments maintain fingerprint and image databases on drivers and social service recipients.
4) My social security or credit card number, and sometimes even my legal name, can identify me out of the entire U.S. population. This capability has not been demonstrated using any single biometric measure.
5) Like phone and credit card information, biometric databases can be searched outside of their intended purpose by court order.
6) Unlike your credit card, phone or Social Security numbers, biometric characteristics change from one measurement to the next.
7) Searching for personal data based on biometric measures is not as reliable or efficient as using identifiers like your legal name or your Social Security number.

Election System Goals

Biometrics have been successfully used to increase the integrity of the driver's licensing and social service benefit distribution processes in many States. There is no question that it is technically possible to use biometrics to limit fraud in voting processes as well. The 14th, 15th, 19th , 24th, and 26th Amendments to the U.S. Constitution establish voting as the right of all citizens 18 years of age or older who have not been convicted of a disqualifying crime. The recognition of voting as a "right", however, separates it from the identified "privileges" of driving and receiving social service benefits

Further, by federal law we have adopted the potentially competing goals of limiting fraud and enhancing voter registration. The National Voter Registration Act (NVRA) of 1993 seeks: (1) "…to increase the number of eligible citizens who register to vote in elections for Federal office…; (2) (to enhance) the participation of eligible citizens as voters in elections for Federal office; (3) to protect the integrity of the

electoral process; and (4) to ensure that accurate and current voter registration rolls are maintained."

It is fair to say that these are the goals of the current Congress as well and should be our goals when suggesting changes to the voting process. Protecting the integrity of the electoral process should include making sure that only eligible voters register, and that only registered voters cast vote. It seems clear that personal identification, possibly involving biometrics, is a key element here. The challenge will be to protect the integrity of the process without burdening this right to vote in ways that may decrease registration by eligible voters.

Making Sure Only Eligible Voters Register

The 14th, 15th, 19th , 24th, and 26th Amendments to the U.S. Constitution identify eligible voters as all citizens 18 years of age and older who have not been convicted of a disqualifying crime. Implicit in these Amendments and the NVRA, and explicit in voting codes, is the additional requirement that each citizen is eligible to register only once. Establishing that you are a citizen at least 18 years old cannot be done directly by biometric identification. This requires trusted source documents, like a certified birth certificate or a passport. If these source documents were linked to a biometric record, which they are not, positive biometric identification could be used to establish the connection of the presenter to the presented source documents. Driver's licenses in the States of Texas and Georgia display encoded fingerprints and could be used to link a presenter to an identity through biometrics, but they are not proof of eligibility to vote and merely transfer the original identification burden to the driver's licensing system. In the absence of biometric data on passports or birth certificates, biometric identification cannot be used to establish my eligibility to vote.

Each citizen is allowed to register only once and in one jurisdiction: "one voter, one vote". In registering to vote, I declare my previous registration, if any, so that I can be removed from the voter roles of my previous jurisdiction. Negative biometric identification could be used to determine if I am previously registered in the current or other jurisdictions, preventing voter fraud through multiple registration of the same voter.

Under the 14th Amendment, citizens can lose eligibility to vote for conviction of some crimes. In registering to vote, I attest that I am not someone who has lost eligibility through conviction of a disqualifying crime. The National Association of Secretaries of State has recommended that a task force investigate the creation of a national clearinghouse of names of disqualified voters[2] to allow the cross-jurisdictional enforcement. This negative biometric identification could be done with fingerprinting because fingerprint records are available on those convicted of disqualifying crimes.

In considering biometric identification for preventing multiple registrations or for preventing registration of disqualified voters, recall that such "negative" identification must always be mandatory for all enrolling in the system. In other words, enforcement of "one-voter, one vote" and disqualification provisions using biometrics would require the mandatory biometric measurement of all registration applicants. In the case of preventing

---

[2] National Association of Secretaries of State Resolution, "Clearinghouse for unqualified voters", 15 July 1998

registration of those disqualified by criminal record, fingerprinting of all registering voters would be required. This would not only require specialized equipment[3], it would place a burden on the entire process for all registrants to deter the activities of a very few. Further, mandatory fingerprinting might be considered a deterrent to registration by those who mistakenly believe that fingerprint databases on minor traffic offenders exist through driver's licensing systems.

Burdening the entire process should be considered only if there is adequate documentation of a clear need. We know of no documented studies on a national basis showing massive fraud through multiple registrations or through the registration of criminally disqualified voters. In short, it is not clear that there is a currently justification for adding mandatory "negative" biometric identification to the voter registration process.

Making Sure Only Registered Voters Vote

Another identification problem faced in the voting process is the positive identification of voters at the polls. Can the poll workers be certain that people appearing at the polls are who they claim to be? The current solution to this problem in many jurisdictions is to have voters announce their name aloud; the concept being that poll workers or other voters present could challenge false claims of identity. Voters are also required to sign a roster. If a voter's identity is challenged later, the roster signature can be compared to that given at registration.

This process could be strengthened in a number of ways. Voters could be asked at the polls to supply additional information given at the time of registration, such as their address or birth date. Voters could be asked to present identification documents, such as a driver's license, birth certificate or utility bill. Voters could be asked to bring to the polls mailed election materials showing name and address. Or voters could supply biometric identification.

This use of biometrics for positive identification could be done on a voluntary basis. Jurisdictions wishing to give voters this option would allow those requesting biometric identification to record a biometric measure when they register. This would require special equipment at the registration sites, as well as at the polling places. It would require the centralized storage of these measures by the jurisdiction. It would also require the transmission of the biometric measures between the jurisdiction and the polling places on election day.

Of all the methods we've listed here for strengthening the process of identifying voters at the polls, biometric identification would require the most additional equipment and cause the most changes to the current systems. However, it would also be the method hardest to defraud. We have, again, seen no documented evidence showing widespread, national problems with voter identification at the polls. If there is a need to strengthen the system in a particular jurisdiction, it seems sensible to start with other less secure and less costly methods of voter identification. Only after these methods prove to be insufficient, or there is a general demand by the voters to allow substitution of biometrics for these methods, could a practical case be made for biometric identification.

---

[3] There are "ink-less" fingerprinting systems available, but there is no evidence that such systems can be successfully used, except by forensic experts, to acquire fingerprints suitable for electronic systems.

Absentee, Nomination and Petition Applications

The identities of voters applying for absentee ballots, petitioning the government or nominating candidates are currently verified by comparing the signatures on these documents to signatures in the voter registration rolls. This labor-intensive process is often aided by electronic "election signature retrieval" systems. Handwritten signatures from voter registration documents are optically scanned into a computer system. Then, election officials can electronically recall these signatures to compare them to those on petitions, absentee ballots and nomination forms. The actual comparison of the signatures is done by human eye.

This process of comparing signatures could be automated. Computer programs for comparing written signatures currently exist in laboratories, but are not currently commercially available. These systems require no special hardware and are different from commercially marketed "dynamic signature verification" that require special pens and tablets. Even if quite crude, this form of biometric identification could successfully reduce the human workload by automatically accepting the signatures that are clearly legitimate or at least very good forgeries: the same signatures that would be easily accepted by human examiners. Only signatures that are not obvious matches would require a human decision. We believe that such automated signature matching could be profitably integrated into current electronic signature retrieval systems.

Other Applications

We can imagine more elaborate uses of biometrics for prevention of "chain balloting" or to allow completely anonymous voting. Chain balloting is a method for corrupting document-ballot elections. A campaign worker gives the complicit voter a pre-marked ballot before he/she enters the polls. At the polls, the voter conceals the pre-marked ballot and is given a blank ballot. The pre-marked ballot is cast and the blank ballot surreptitiously returned to the campaign worker after leaving the polls. The campaign worker marks the ballot for the next voter. In 1992, about half of the States using a document ballot had procedures in place to prevent chain balloting.

Biometric identification could be used to print a biometric identifier on the ballot stub when the ballot is issued. The biometric measure on the stub could be compared to one taken from the voter when the vote is cast. The stub would be given to the voter so that no biometric record of the voter would remain at the polls after the voter has left. This application would require the mandatory biometric measurement of all voters.

In theory, completely anonymous voting could be accomplished by registering volunteering voters under a biometric identifier. Eligibility at registration would be ascertained using current methods and registration records would include the voter's name. Only the voter's biometric identifier, however, would be sent to the polls. At the polls, voters would present the biometric identifier in lieu of announcing a name. This extreme application would significantly alter the current system of publicly releasing the names of those who have voted.

Internet Applications

In 1999, the State of California created an "Internet Voting Task Force" to study the possibility of casting votes over the Internet. The task force found that one of the obstacles to Internet voting would be the identification of the person casting the vote.

The problem of identification of Internet voters is one of both positive and negative identification. Negative identification would be required if we wished to prevent multiple registrations of the same person. Positive identification would be required to identify the person casting the vote as the registered voter.

As discussed, negative identification must be mandatory for all voters. In the case of Internet voting, multiple Internet registrations could be prevented by the mandatory biometric identification of all Internet voters at registration. This would not require mandatory identification of non-Internet voters if we were willing to allow for the possibility of fraud through both Internet and paper registration of the same voter under different identities. Internet registration with the submission of a biometric identifier could not be securely done over the Internet, but would require "in person" registration and the collection of the biometric identifier by trained and trusted persons This identifier would be placed in a database under the control of the jurisdiction. Upon verification that the registering voter is not already in the database, a voter ID number, code or PIN could be issued. Biometric identification and specialized hardware at the time of voting would not be required for negative identification.

Positive identification by Internet voters using biometrics would require that biometric measures be previously registered "in person" with the jurisdiction and would require standardized biometric collection hardware and software on the computer used for voting. Positive biometric identification might be used on a voluntary basis to replace other types of PIN or password identification. An added problem is the occasional failure of all biometric techniques to recognize properly registered users. "Provisional" voting would have to be allowed in cases where the voter's submitted biometric measure did not seem to match the registered measure.

In short, biometric identification could be an important adjunct to Internet voting, but would not solve all identification problems inherent in Internet voting.

Conclusions

In this paper we have looked at specific applications of biometric technologies to the voting process. We can conclude that biometric identification could be effectively used, even on a voluntary basis, to detect and deter voting fraud. Biometric identification, however, is not a "silver bullet" capable of solving all problems of voter identification without any undesirable side effects. Use would require fundamental changes in the way we register voters and would necessitate the creation of government-controlled databases of physical and behavioral characteristics of at least some voters. Although such databases pose no threat to the privacy of voters, the process could be seen as an additional burden on the registration process. We would need to carefully consider the potential impact of such changes on the competing requirements of the National Voter's Rights Act of 1993 to both enhance voter participation and to protect the integrity of the electoral process.

Biometric Identification and the Financial Services Industry

James L. Wayman, Director
U.S. National Biometric Test Center

Biometric Identification

Biometric identification is the automatic identification, or identify verification, of an individual based on physiological or behavioral characteristics. These characteristics include voice prints, hand and finger shape, eye structures, habituated hand movements, facial features and fingerprints. Using a correctly chosen and designed biometric identification system, I may be able prove to reasonable certainty that I am, or am not, someone previously registered in the database of users. I want to emphasize that there are two possible functions: proving I am and proving I am not registered in the database. Many biometric technologies have been shown efficacious in the former application, few in the latter. Identitfication is always to reasonable, not absolute, certainty. The motivation for using biometric identification is to decrease costs and increase convenience (for both the users and system managers), while maintaining the required level of security.

Many successful systems are currently in use by government and industry to support a variety of applications. The Immigration and Naturalization Service uses hand geometry with the INSPASS card, and facial and voice recognition with the SENTRI border crossing program. Many companies use biometric devices for time and attendance recording. San Francisco International Airport uses hand geometry to control access to the tarmac. Disney World is using finger geometry with their season passes. At San Jose State University, we use hand geometry to control access to the computer center and facial recognition on the door of our biometrics laboratory. Eight states use fingerprinting with their social service programs. Five states use fingerprinting with their driver's licensing programs. Many localities use voice recognition in home incarceration programs. I emphasize that these are all current, existing applications.

The National Biometric Test Center

The Biometric Identification Research group at San Jose State University was started in 1995, receiving an initial contract from the Federal Highway Administration to advise the Secretary of Transportation on establishing "minimum uniform standards for biometric identification…of operators of commercial motor vehicles", as required by the 1988 Truck and Bus Safety and Regulatory Reform Act.

The U.S. federal government's Biometric Consortium had contemplated forming a biometric test center for several years. Sensing that biometric technology was potentially of great use by the government, but alarmed that vendor accuracy claims were not replicated in laboratory tests and that laboratory performance was not realized in field operations, the Biometric Consortium wanted to form an independent test agency for establishing a science of biometric device evaluation. After a multi-year competitive process, San Jose State University was named in 1997 as the "National Biometric Test Center". Our charge was to establish objective, statistically-based testing criteria, to collect and analyze performance data, and to advise the government on the use (or non-use) of this emerging technology.

Science is, by tradition, conservative and skeptical. It is usual within science to accept results only at a level that could have occurred by chance less than one time out of

twenty. For example, to prove that a coin has less than a 50/50 chance of coming up heads, you would need to flip 5 tails in a row the first time you tried. Applying this same statistical reasoning to biometric device testing, no errors out of 300 independent trials proves that the error rate is less than 1%. These 300 independent trials require 300 human volunteers, each giving two biometric samples separated in time by weeks or months. Unfortunately, this is extremely expensive.

The National Biometric Test Center, therefore, was created by a government consortium concerned over non-verifiable vendor claims and is rooted in a scientific tradition of conservatism and skepticism. We pride ourselves in being perhaps the most critical, questioning scientific voice in the field of biometrics.

Our work for the Federal Highway Administration centered on establishing methods for matching technologies to applications and a frame-work for standards development. Our work as the National Biometric Test Center has focused on advanced mathematical methods for system performance estimation and for lowering the cost of device testing by using field data. We have collected the world's largest test database of electronically scanned fingerprints and have completed extensive benchmark testing on the world's major fingerprint system vendors. We have begun testing the smaller fingerprint vendors, have analyzed data from a large hand geometry application, and have begun facial recognition device testing. Additionally, we have established an extensive library of reports on other independent tests.

Biometrics and Financial Services

Having established our credentials as a skeptical group, we can say without reservation that many biometric identification systems have been shown capable of increasing convenience, privacy and security, while decreasing costs and hassles, in a variety of applications. There may be many such potential applications within the financial services industry.

Some of these applications may be for internal or infrastructure protection, such as access control to restricted spaces or to networked computers used for electronic funds transfer. There is no reason that financial institutions could not implement such systems immediately, if they so chose. Of more interest perhaps at these hearings are the potential consumer applications, particularly those for identity and privacy protection. Certainly the first application that comes to mind is that of protecting ATM, credit cards and checks with a biometric identifier. Currently, ATM cards are protected with a pass code, but credit cards are protected only by the presence of a signature on the back panel. Checks rely on the protection of other identifying documents. The technology now exists to replace pass codes with biometric measures without a substantive decrease in security protection.

Identity verification with checks and credit cards is currently done by the human accepting them as payment. As we have been reminded by today's testimony of Shanin Leeming, in practice this affords no protection at all. For the financial services industry to provide for the biometric protection of credit cards and checks as a matter of consumer choice would be welcomed by the vast majority of customers. The experience of the Purdue Employees Federal Credit Union, also to be reported on today, supports this finding.

Practical Impediments to General Adoption for Consumer Appplications

We have argued that biometric identification as a consumer choice in the financial services industry would be a good thing. There are practical problems, however, to general adoption at the consumer lever, the major one being the lack of a single standard supported by the entire industry. For example, if my bank allows me to protect my various cards (or single function smart card) with an eye scan of some type, my ATM card will only be usable at other ATMs supporting this same device and not at ATMs supporting fingerprinting, voice printing, or no biometric identification at all. My checks and credit card will only be accepted at points of sale supporting that form of eye scanning. For the retailer, the diversity of biometric methods would require perhaps a dozen data collection devices at every cash register. Even if a single biometric measure such as fingerprinting were accepted as the standard, there are dozens of proprietary formats for data storage and analysis.

Why not simply pick the best biometric device, select that vendor as the standard and solve the standardization problem by creating an instant multi-billion dollar company? This is the computer industry paradigm. It has not occurred in the biometrics industry because there is no "best" device. Each technical approach has its own strengths and weaknesses. (For a more detailed explanation, see J. Wayman, "Testing and Evaluating Biometric Technologies: What the Customer Needs to Know", *Proceedings of CardTech/SecurTech'98,* pg. 329-348)

Biometrics and Privacy

When creating and implementing any new technology, we should always be vigilant regarding its impact on us as individuals and as a society. With any discussion of biometrics, the appropriate question always arises, "What about my privacy?" A productive discussion of this issue must be rooted in the actual capabilities of the technology, not on the capabilities imagined by overly zealous vendors or ratings-boosting radio talk show hosts. The primary concern seems to be "They will find me, track me or correlate my personal data". "They" is commonly thought to be some government agency or some hacker on the internet. Three technologies, face, voice and hand geometry, have been shown in independent testing to be incapable of singling out a person from a group exceeding a thousand. For most other technologies, such as iris scanning, vein patterns, and facial thermograms, we have no data supporting this capability.

Only two biometric technologies, fingerprinting and retinal scanning, have been shown in independent testing to be capable of singling a person out from a group exceeding a thousand people. The current design of the retinal scanning device supports only "cooperative" applications, those in which the user wants to be singled out. Fingerprinting, as used in the financial services industry, does not save data in a format compatible with large-scale searches. Only numbers derived from the fingerprint or retinal scan, not the image itself, are stored. Because of lack of standards regarding the method used to develop these numbers, they are useless to any other system, even to the FBI. So "they" can find you and track you, at best, only when you are using a single system.

What about the correlation of data? "If an unscrupulous person gets my biometric data, perhaps they can use it to assemble my health records, my driving record, my banking data." This common misperception seems to be modeled on vulnerabilities of

the social security number. Again, most biometric methods do not support large-scale search, so having my hand geometry template, for instance, will not help you to find any records indexed by it. Further, the lack of standards creates what my colleague, John Woodward, has called "biometric balkanization", meaning the inability of systems to communicate because of their diversity. Incidentally, my personal facial image, hand geometry and fingerprint template data are available on the National Biometric Test Center's web site. My social security number, credit card number, phone card number, ATM PIN, and mother's maiden name are not. It is the disclosure of this latter information, not of my biometric data, that presents a genuine threat to my privacy.

What about biometric identification as a positive tool for the active protection of personal information and identity? My California driver's license, which is seen rarely by traffic enforcement officials, but shown daily to bank tellers, airline ticket counter agents and grocery clerks, clearly displays, for the purpose of identification, my height, weight, date of birth, eye color, and hair color. Although possibly in violation of California law, I have no intention of correcting the error in my weight as noted on the license. All of this data, including also my name and address, could be replaced by a single biometric measure. Upon confirming through the biometric data that the card was mine, the clerk or traffic officer could simply note the license number, allowing contact with me through the issuing agency, if required. With regard to the positive protection of financial data, it is clear that the customer option of requiring biometric data for access to records would decrease the likelihood of unauthorized released. Biometric identification can serve as a positive tool for privacy protection.

Conclusions

Many successful applications of biometric technology currently exist. This technology has proven capable of decreasing costs and increasing convenience for both users and system administrators. Further, these systems are capable of increasing both privacy and identity security. There is no reason why these devices could not currently be used within the financial services community for internal applications and infrastructure protection, such as for access control to sensitive spaces and computers. The major impediment to universal implementation at the consumer level is the wide variety of competing, vendor-proprietary devices, all without general standardization. This cacophony of devices, however, further serves as a protection to privacy, preventing any one measurement to be used to access non-communicating systems. We believe that adoption of these devices, and the necessary standards, within the financial services industry will be driven by consumer demand for increasing privacy and identity protection.

Picture ID: Help or Hindrance? Do People Really Look at the Picture on a Picture ID?

Miss Shanin Leeming
Merritt Island, FL
Brevard Intracoastal Regional Science and Engineering Fair
Category: Behavioral and Social Science

My mom and I have often talked about the need to show an identification card at the airport and other locations. We discussed that no one seems to really look at the picture on the identification card that is presented. I wondered if that might just be an inaccurate observation on my part, or is it possible that people only pretend to be checking and verifying the identity of the person actually presenting the card for various reasons.

If people are not really checking the picture against the appearance of the person standing in front of them, then is it possible that we are putting too much faith in this whole system?

My project, then, is designed to test the validity of this picture identification system. I decided to set up ten test situations and dress my mother in ten ways which would alter her appearance. I chose to alter her appearance in varying ways, starting with gradual alterations. As my project progressed, I quickly realized that it did not matter how outrageous my alterations got. Because no one took any notice of her appearance at all. In fact, we were able to accomplish a number of varying tasks (including check cashing and legal document signing), with her being dressed in a clown suit and a man's attire.

Consequently, I strongly recommend that people should not put too much faith (if any at all) in the current system of picture identification. Recommendations are made to consider other forms of identification with more complex safeguards if we are to protect ourselves from fraud.

(Editor's. note: Clockwise from left, the pictures show Mom with notarization received in disguise, Mom with wine purchased while in clown suit, Mom having cashed check as man, Mom having used credit card while in wig and mortician's makeup)

Picking the Best Biometric for Your Applications

James L. Wayman
National Biometric Test Center

Lisa Alyea
NSA/CESG

1. Primer

Biometric Identification: The automatic identification or identity verification of (living) individuals based on behavioral and physiological characteristics.

There are two basic applications of biometric technology:

1) Positive identification: To demonstrate I am someone enrolled in the system.
2) Negative identification: To demonstrate I am not someone enrolled in the system.

For positive identification, the user will generally claim an identity by giving a name or an ID number, then submit a biometric measure. That measure is compared to the previously submitted measure to verify that the current user is the one enrolled under the claimed identity. The purpose of positive identification is to prevent multiple users from claiming a single identity. There are numerous non-biometric alternatives in such applications, such as ID cards, PINs and passwords. Consequently, use of biometrics for positive identification can be made voluntary and those not wising to use biometrics can verify identity in other ways. The U.S. Immigration and Naturalization Service's Passenger Accelerated Service System (INSPASS), in use at 11 airports, is an example of voluntary, positive-identification, biometric system. Those not wishing to use biometrics at ports of entry can verify their identity through their passport.

In negative identification, a user claims not to be previously enrolled in the system and submits a biometric measure, which is compared to all others in the database. If a match is not found, the user's claim of non-enrollment is verified. The purpose of negative identification is to prevent claims of multiple identities by a single user. There are no reliable non-biometric alternatives in such applications. The use of biometrics in negative identification applications must be mandatory. Biometric identification for driver's licensing in 5 states and welfare eligibility verification in 8 states are examples of mandatory, negative-identification, biometric systems.

Some biometric systems are used for both positive and negative identification. The State of Connecticut Social Service and Philippine Social Security System ID cards, for instance, require negative identification for issuance, but store fingerprint "templates" on the card for later positive identification applications.

In positive identification systems, a false match is called a "false acceptance" and a false non-match is called a "false rejection". In negative ID systems, the terminology is reversed. Regardless of whether a system is for positive or negative identification, false acceptances allow for fraud and false rejection are inconvenient, requiring "exception handling". The "false rejection" rate is immediately measurable from user demands for exception handling. Instances of "false acceptance" are almost never reported. The perceived rate, however, must be kept low enough to maintain deterrence.

A biometric sensor takes a one, two, or three-dimensional image from a user which is reduced by a computer in some proprietary way to a "template". A template is a collection numbers deemed to be adequately different between individuals and adequately stable over time for a single individual. Generally, the original image is

discarded and only the template is stored by the system. In almost all cases, the original image cannot be recreated from the template.

Nothing inherent in a biometric system can identify you by name, citizenship, age or race. If a system must know any of these items, they must be established through external means, such birth certificates and driver's licenses. Consequently, use of biometrics to establish "real" identities is only as reliable as the source documentation. For example, biometric systems cannot be used to establish that social service recipients are eligible for benefits beyond showing that they have not claimed multiple identities (negative ID) or have not falsely claimed the identity of a true beneficiary (positive ID).

Because a biometric system cannot know who you "really" are, use of biometrics to support anonymous transactions becomes are real possibility. For instance, a credit card could carry one of your biometric measures instead of your name. Further, as images cannot generally be reconstructed from templates (which are just a series of numbers), system administrators cannot generally obtain any information about users in any humanly recognizable form. Consequently, biometric identification technology is, at worst, neutral with regard to privacy.

Establishing the Business Case

All security systems require the expenditure of time, energy and money. Biometric systems are certainly no different in this regard. They are not free in any sense. In our experience, many failed biometric efforts do so, not because of deficiencies in the technology, but because the business case was not sufficient in the first place to justify the required expenditures. Fascination with the technology is not a sufficient business case. For positive identification applications, alternatives to biometrics exist that might be faster, cheaper and more seamlessly integrated into existing systems.

The most successful biometric implementations are those that replace existing systems deemed too expensive or problematic to the administrators, or too cumbersome to the users. As examples, we point to use of biometrics in INSPASS and for access to dormitory food service facilities at the University of Georgia. Other alternatives exist in these situations, but biometric identification has proved faster, cheaper and easier for all concerned.

Other successful implementations occur when the system management has carefully assessed the alternatives and is prepared to do the work necessary to make the systems effective. Examples are the access control applications at the San Jose State University Computer Center and for season pass holders at Walt Disney World.

Several points must be kept in mind in preparing the business case:

1) Alternatives to biometric identification exist in positive ID applications.

2) All security systems, even biometrics, require time, money and energy to set up and run. In addition to set-up and operational costs, system throughput rates must be carefully considered. Remember that enrollment sessions for all users is almost always required.

3) Not all people will be able to use any biometric system successfully every time. This implies that backup systems for "exception handling" will always be required.

4) Studies of user attitudes regularly show user acceptance of biometric technology to well exceed 90%. Nonetheless, there will always be a very few people who object to any new technology.

5) Choose your system integrator extremely carefully. Hardware/software integration will prove to be the hardest task. Biometric technologies are not "plug and play". Even ideal technologies will fail if the devices cannot talk to the database or open the gate. System integration may require changes in other pieces of hardware not considered at first glance to be part of the biometric technology.

6) Know the history and track record of the technology vendor. Commercial products and vendors are in a continual flux. The technology you invest in today may not have vendor support next year.

7) The addition of biometrics, or substitution for another component, will inevitably lead to a change in your business processes. Beyond the software/hardware integration is the most daunting problem of integrating the use of biometrics into the existing processes. If the finished business **system** is not more efficient than the alternatives, the use of biometrics will be seen as a mistake.

Assessing Your Application

The first task in picking a technology is to assess your application environment. The various technologies are strongly differentiated by their technical applicability to different environments. Beyond determining whether your application is positive or negative, it is necessary to establish the following:

1) Will your users be habituated or non-habituated? That is, after a period of time, will the average user be accessing the technology regularly or only sporadically? Some technologies require greater user involvement and cooperation than others.

2) Will your users be supervised or unsupervised? Systems vary in the level of required supervision and/or user prompting at enrollment and during operation. The level of training required by enrollment personnel also varies over the technologies.

3) Will deceptive users be cooperative or non-cooperative with the system? In negative identification applications, the fraudsters will attempt to foil identification at enrollment. Consequently, enrollment supervisors will require training in detection of fraudulent techniques. In positive identification applications, fraudsters will be generally cooperative with the system in an attempt to be positively identified.

4) Will your system be public or private? That is, will the enrolled users be employees or people otherwise under the management of the system administrator, or will they be members of a more general population?

5) Will your system be required to exchange data with systems operated by different management (open) or stand-alone (closed)? The new BioAPI standard will solve some of the interface problems, but not all. There are no existing standards for biometric templates, so systems from differing vendors will not generally be able to share templates or even images, even if based upon the same biometric characteristic.

6) Will the application be indoors in a "standard" environment or in a "non-standard" outdoor or otherwise harsh physical environment? Not only will system weather-proofing be a challenge, people in outdoor environments cover themselves in varying and unpredictable ways. To our knowledge, the only successful outdoor applications have been in very temperate environments.

7) Do you have data storage limitations anywhere in the system? Template sizes vary from 9 bytes to 6 kbytes depending upon both vendor and technology. Not all template sizes are suitable for mag stripe or even smart card storage, for instance. Further, some technologies require the storage of multiple templates for good performance.

8) How much system reporting will be required? Most systems log all activity by time, date and user ID, but the ability of the vendor software to generate reports differs widely. A good understanding of your audit trail requirements is necessary.

9) You must establish your throughput rate requirements for both enrollment and operation. Almost all systems require enrollment. Some requiring multiple enrollment images. The finger geometry system in use at Walt Disney World requires no special enrollment session, but rather automatically enrolls season pass holders upon first use.

10) What number of errors per hour, day, etc. can be tolerated? "False rejection" errors will require "exception handling" and will greatly decrease the throughput of your system. "False acceptance" errors will erode the perceived integrity of your system. Errors can be decreased, often at the cost of throughput rate, through more careful enrollment and more quality-control feedback to the user. Systems vary considerably in the amount of automatic quality control applied to the acquired images and the nature of the image quality information given the users.

11) For how long will you expect your enrollment images to remain usable? "Template aging" effects all biometric systems as a person's physical and behavioral characteristics change over time. Some technologies, however, experience performance degradation more rapidly over time than others.

12) What is your budget? Systems can vary in price from tens of dollars per collection location to over US$25k.

Armed with the above assessment, it is often possible to find other users with experience in closely related applications.

The Technologies

Commercially available technologies include the following:

1) Facial recognition;
2) Fingerprinting;
3) Palm printing;
4) Hand geometry;
5) Finger geometry;
6) Iris scanning;
7) Retinal scanning;
8) Facial thermography;
9) Speaker verification;
10) Dynamic signature recognition;
11) Keystroke;

Of these, only fingerprinting and retinal scanning have been independently tested in negative identification pilot projects. The vendors of facial, palm and iris recognition systems claim that their products can be used in these applications as well, but such claims have not yet been confirmed in independent third-party pilots.

Independent, government-sponsored reports are available on the use of hand geometry, fingerprinting, speaker, signature, iris scanning, retinal scanning and facial imaging for positive identification. These reports always pertain to application environments of interest to the government. Many of these reports, however, are now outdated, as the pace of technology development exceeds the government's interest in evaluation. Private, third-party evaluations have also been done, but the results are always controlled by those paying for the test and are not made publicly available.

Although anecdotal, the experience of other users is usually the most valuable information. Care must be taken, however, to remember that system performance is highly application dependent. Consequently, experiences with a particular technology in one application may not readily transfer to other applications.

Some choices of technology may be obvious based upon hardware that is otherwise available, for instance, the use of facial recognition for systems already employing digital imaging or speaker verification for telephone applications. In most cases, however, the choice will be far from obvious and will ultimately be driven by price and the recommendation of other users.

Checklist

The following checklist summarizes the issues discussed above and may be helpful.

1. Have you investigated the alternatives to the biometric solution for your problem?
2. What legal/political issues could hinder your program (privacy, data access, etc.)?
3. Have you addressed the issues of ease of use of the biometric by both users and system administrators?

4. Do you need to establish enrollment template storage size(s)?
5. Will multiple templates per user be required?
6. What sort of computer resources do you envision will be needed to support your overall system?
7. Have there been any tests/evaluations of biometric systems similar to your particular application?
8. Have you addressed user enrollment, data collection, data capture, data transmission, data translation, signal processing, authentication policy, template storage, and user management features in your procurement document?
9. What kind of enrollment policy will you have? How long should enrollment take? Does the enrollment need to be supervised? Will the enrollment database need to have the capability to handle back-ups and perform simple recovery procedures?
10. What is the cost of the biometric solution in terms of hardware, software, personnel, training, and impacts on existing procedures?
11. What factors are most likely to increase costs of the system?
12. What are the likely costs likely for making the system mandatory to all as opposed to making it optional?
13. What are the benefits likely to be? a) In terms of money? b) In terms of "non-monetary" benefits?
14. Have you surveyed your user population as to the attitude towards using a biometric? A strongly negative response should indicate a reformulation of your plans.
15. Will your users be employees, customers, or both?
16. What is the degree of public acceptance/user perceived intrusiveness of the intended biometric?
17. Does the majority of your target user population contain physical characteristics that could pose either advantages or disadvantages for your chosen biometric system?
18. Will the biometric system in your particular application to be used for positive identification, negative identification, or both? If both functions are required, will they be required from the same biometric measure, or can two measures be used (e.g. fingerprint and voice, face and voice, etc.)?
19. Will the deceptive user be cooperative or non-cooperative in your application? What types of fraudulent user scenarios can you foresee?
20. Will your users be habituated, non-habituated, or a mixture of both? If both, what is your best estimate for percentage of users in each case? What will the vendor/integrator need to do to prepare the system for your particular mix of users?
21. Will the system be open or closed?
22. Will the system operate in a standard or non-standard environment? If non-standard, list the non-standard conditions.
23. Have you listed the available hardware for the application? Will interoperability of systems be an issue? What about backward compatibility? Is flexibility desired? Are upgrades possible with a minimal amount of fuss?
24. What sort of quality control and feedback will the vendor offer on the enrollment?
25. Does the biometric capture device have the capability to perform automatic self-diagnostic and calibration tasks, or will the system administrator have to attend to this periodically?

26. Will a human operator have the ability to intervene in the enrollment process in order to establish a better enrollment record?
27. Will the system automatically flag poor quality biometric input data? How much of the input do you expect to be flagged as poor quality data?
28. Will the system use more than one instance of captured biometric input data to create the enrollment template?
29. What throughput requirements do you have?
30. What is a tolerable error rate for both erroneously identifying a match when there is not one and not identifying a match that should actually be one?
31. Will the probability of a false match be low enough to deter fraud?
32. How many of false non-match errors can you tolerate? Will the user be given additional attempts to try and be recognized? What will you define as the tolerable rate of occurrence for false non-matches that require intervention by trained staff?
33. Did you define back-up methods for user authentication in the cases of equipment failure and biometric feature unavailability?
34. Is an appropriate contingency plan and disaster recovery policy important to the success of your program?
35. Have you defined the roles of a security officer, auditor/audit trail requirements, administrator, and user responsibilities for your application?
36. Does the system support a lockout threshold for excessive invalid access attempts?
37. Does the audit information need to include any or all of the following: the number of new biometric records accepted, the number of biometric records verified, the number of users the system was unable to enroll, the quality measurements for the captured biometric data, the amount of system down time, the kinds of system errors by type, and the average enrollment processing time on a daily, weekly, and monthly basis?
38. Have you investigated the possible usage of tamper deterrent and tamper indicative technologies for your system?
39. Must the system guarantee the integrity and security of the data it holds?
40. Have you done your homework on the potential vendors/integrators who have submitted for your proposal?

Biometric Authentication Standards Development

James L. Wayman, Director
U.S. National Biometric Test Center

Biometric authentication is "the automatic identification or identity verification of living, human individuals based on behavioral and physiological characteristics" and includes such technologies as speaker verification, automatic fingerprint identification systems (AFIS), eye scanning, and facial recognition. Each of these technologies originated at different times for different purposes and with different academic pedigrees. It has only been within the last decade that their commonality as automatic human identification methods has been recognized, so it should come as no surprise that common standards have been slow to develop.

But further, there has been little need for inter-operability among these systems. In fact, the non-interoperability within and across technologies has been touted as a privacy-preserving asset of biometric systems. Consequently, there has been no motivation for the tedious standards development process required to promote inter-operability. A notable exception is in AFIS, where law enforcement has long needed the capability of inter-jurisdictional fingerprint exchanges. Consequently, in the case of automatic fingerprint identification, some American National Standards Institute (ANSI), National Institute of Standards and Technology (NIST) and Criminal Justice Information Services (CJIS) recognized standards do exist.

The CJIS "Interim IAFIS Fingerprint Image Quality Specifications for Scanners," (CJIS-RS-0010v4, Appendix G) specifies requirements for signal-to-noise ratio, gray scale resolution and histogram, modulation transfer function and geometric distortion for fingerprint images scanned into an AFIS. This standard, commonly known as "Appendix G", was developed for computer-based "flat bed scanners" used for the digital imaging of ink fingerprint cards. It has been applied, with difficulty, as a standard for electronic fingerprint "live" scanning devices which digitally image the fingerprint directly without the use of intervening ink. "Appendix G" compliant scanners are considered to have the minimum quality necessary for capturing fingerprints for later human comparison to crime-scene "latent" prints. For want of any other standards, the AFIS community has adopted "Appendix G" as the standard for fingerprint capture devices in purely automatic search applications where human, "latent" print searches are not involved. Fingerprint devices not used for large-scale searches, such as those for computer or facilities access control, commonly do not meet "Appendix G".

Notably missing from "Appendix G" is a specification for image resolution and size. The resolution standard of 500 pixels per inch is actually included in the "Data Format for the Interchange of Fingerprint Information" standard, ANSI/NIST-CSL-1-1993. Oddly, neither standard specifies an image size. This latter standard specifies header and content data for exchanging full fingerprint images between jurisdictions. The standard was also expanded in 1995 to include facial ("mug shot") images and "scar, mark, and tattoo" information.

Images, in general, contain a lot of information. If a picture is to be "worth a thousand words", it will also require that much more storage space. Similarly, if images are to be transmitted, a good deal of bandwidth will be required. But images usually can be "compressed", allowing them to be stored or transmitted as smaller files, then expanded later to produce images nearly as good as the originals. A standard for

compression of general photographs—JPEG (Joint Photographic Expert Group) compression – has existed for many years. This is the standard compression method for facial images used for biometric identification. Unfortunately, JPEG does not work well with fingerprints, often causing unacceptable distortion in the expanded images.

Consequently, ANSI, NIST and the FBI have developed the "Wavelet Scalar Quantization (WSQ) Gray-scale Fingerprint Image Compression Standard" which allows 15-to-1 or greater compression of fingerprint image files. In other words, a 1 inch by 1 inch fingerprint image, containing 500 by 500 pixels, could be reduced in size from 250,000 bytes to about 17,000 bytes (17 kbytes). This 17 kbytes, while still much too large to be put on an ID card, is small enough for storage in a centralized AFIS.

During the past year, several groups within the community of AFIS users have become convinced that a new standard is necessary to allow storage of fingerprint data on ID cards. All biometric systems must extract "features" – that information considered most useful for matching – from the raw data images. Most AFIS extract as "features" the fingerprint ridge details known as "minutiae" (ridge splits and endings) from the images, then match different fingerprints by comparing the location of these minutiae. These locations can be stored in files of much smaller size than even the WSQ-compressed images. Consequently, it appears technically feasible to store fingerprint information on ID cards in the form of minutiae locations, if a standardized method for extraction and storage can be found. AAMVA, motivated by the need for cross-jurisdictional exchange of fingerprint data for driver identification, has taken the lead organizing a standardization committee and several very productive technical meetings have been held. I highly commend AAMVA for this effort. They have pulled together both industry and user interests and are moving rapidly toward a technical standard.

But enough on fingerprinting, what about standards for other biometric methods? Compression and transmission standards for facial imaging have already been mentioned. AAMVA also has a "Best Practices" document for the collection of facial images.

Voice systems based on telephone-collected speech have common digitization standards for converting sounds to numbers, but no storage or feature extraction standards are in place.

Hand geometry recognition is dominated by a single vendor, so their proprietary methods of data imaging, feature extraction, and storage are the de facto standard.

Commercially available iris recognition systems use a common feature extraction and storage system, but do not start with exactly the same iris images (The different vendors use different light wavelengths for imaging). No study on iris recognition data interchange between different systems has yet been published.

Beyond the issue of inter-operability standards is that of software protocol conventions. In the past, each vendor has used its own platform-specific software to support its own data collection, feature extraction, storage and matching. This has caused major headaches for biometric system integrators who try to use biometric devices as part of larger or more general access control or information retrieval systems. The integrators have been forced to learn and handle the idiosyncratic software of each biometric vendor. During the last 3 years, under the sponsorship of both the Department of Defense and NIST, common software standards have been emerging. This effort is currently known as the "Biometric Applications Programming Interface" (BioAPI) and version 1.0 has recently been announced. This standard will specify exactly how

information will be passed back and forth between the larger system and the biometric subsystems. It will allow system integrators to establish one set of software "function calls" to handle any biometric device within the system. The U.S. Army has announced that future Army procurements of biometric devices will require BioAPI compliance.

Clearly, we've seen only the beginning, not the end, of biometric standards development. The development of standards is clearly the next step in moving this technology forward to greater levels of use in practical application