Evaluation of Automated Biometrics-Based Identification and Verification Systems

WEICHENG SHEN, MEMBER, IEEE, MARC SURETTE, MEMBER, IEEE, AND RAJIV KHANNA, MEMBER, IEEE

Recent advancements in computer technology have increased the use of automated biometric-based identification and verification systems. These systems are designed to detect the identity of an individual when it is unknown or to verify the individual's identity when it is provided. These systems typically contain a series of complex technologies that work together to provide the desired result. In turn, evaluating these systems is also a complex process. The authors provide a method that may be used to evaluate the performance of automated biometric-based systems. The method is derived from fundamental statistics and is applicable to a variety of systems. Examples are provided to demonstrate the practicality of the method.

Keywords*—*Binomial, biometrics, confidence, evaluation, face, fingerprints, identification, statistics, testing, verification.

I. INTRODUCTION

Recent advancements in computer hardware and software have enabled industry to develop affordable automated biometrics-based identification and verification systems. These systems are now used in a wide range of environments, such as law enforcement, social welfare, banking, and various security applications [1]–[3]. Many biometrics, including fingerprints, facial features, iris, retina, hand geometry, handwriting, and voice, have been used for the identification and verification of individuals. Each biometric has its own advantages and disadvantages, and choosing the best one for a specific application is influenced by both performance criteria and operating environment. When designing a biometrics-based system, it is very important to know how to measure the *accuracy* of a system. The accuracy is critical for determining whether the system meets requirements and, in practice, how the system will respond. Measuring the accuracy of these systems is a primary consideration and is necessary for the objective selection of such systems. In this paper, we provide a method for evaluating automated biometrics-based identification and verification systems. We discuss how to obtain an estimate of the accuracy of these systems as well as how to use that estimate to determine whether a system satisfies the needs of a particular application.

In our evaluation strategy, we first define system-performance metrics in terms of statistical error rates. These performance metrics are independent of the underlying biometrics or their features. Then we perform a test of the system by operating it under conditions that best approximate a normal operating environment, using a set of *known* biometric data as the test samples. In other words, we have prior knowledge about what the outcome should be, which is often referred to as the *ground truth*. Any inconsistency between the outcome of the system and the ground truth constitutes an error. We can then calculate the estimate of the matching errors produced by the underlying system. The matching errors are the parameters in our parameter estimation problem.

We classify automated biometrics-based systems into two major categories: one-to-one systems and one-to-many systems. A one-to-one system compares the biometric information (features) presented by an individual with biometric information (features) stored in a data base corresponding to that individual. The individual using the system asserts his identity, allowing the system to retrieve data from the data base corresponding to the individual. Then the one-to-one system decides whether a match can be declared. Such a system is often referred to as a verification system. In contrast, a one-to-many system compares the biometrics information presented by an individual with all the biometric information stored in a data base and decides whether a match can be declared. Such a system is often referred to as an identification system. One-to-many systems normally require more powerful match engines than one-to-one systems because of the great number of comparisons required when the biometric-information data base is very large.

The remainder of this paper is organized as follows. In Section II, we provide background and discuss the notation used. In Section III, we present the definition of a set of parameters characterizing the performance of automated

Manuscript received February 1, 1997; revised June 16, 1997.
W. Shen is with Pacer Infotec, Inc., McLean, VA 22102 USA.
M. Surette is with Silicon Graphics Inc. Chantilly, VA 20151 USA.
R. Khanna is with Mitretek Systems, McLean, VA 22102 USA.
Publisher Item Identifier S 0018-9219(97)06638-3.

biometrics-based identification and verification systems. In Section IV, we describe the approach for estimating the precision parameters as well as the hypotheses testing to decide if the system meets the requirements. In Section V, we present an example of how the approach is used for selecting an automated fingerprint identification systems (AFIS's). Section VI concludes the paper.

II. Background and Notation

We conduct the system performance parameter estimation test as an experiment of many independent trials. The test samples used consist of two sets: a *search set* and a *file set*. The search set is used to simulate a queue of incoming requests to the system, while the file set is used to simulate data stored in a data base. A *mate* is the biometric data in the data base that belongs to a member of the search set.

In a one-to-many system, each trial is initiated by submitting a search request to the automated identification system. A search request matches a submitted search subject against each file subject in a data base. The system will compare biometrics data of the search request with all biometrics data stored in the data base to determine if the search subject matches any file subjects. In a one-to-one system, each trial is initiated by submitting a verification request to the automated verification system. A verification request matches a submitted verification subject against one specified mate in a data base. In both scenarios, the system makes a match or no-match decision. A match decision means that the automated identification system has found at *least* one mate in the data base or that the automated verification system has matched the verification subject with the retrieved mate. In a one-to-many system, a no-match decision means that no mate of the search subject has been found in the data base. In a one-to-one system, a no match means that the search subject does not match the single retrieved record from the data base. If the search subject and file subject matched by the system indeed come from the same individual, we say that a *correct* match has been made. On the other hand, if the search subject and the file subject matched by the system come from different individuals, we say that a *false hit* (incorrect match) has been made. Each submission of a search request is considered a trial and each match or no-match decision made by the automated identification and verification system is considered the outcome of a trial. Repeating this process (trial) for each search request, we collect the outcomes of the experiment trials. The collection of such outcomes forms the basis for statistical performance analysis of the automated identification and verification system.

In general, various automated biometrics-based identification and verification systems have various parameters that can be adjusted to improve performance. These parameters may have different values for different application environments. It is to our advantage that the developers of the automated biometrics-based identification and verification systems are informed about the quality and characteristics of the biometrics used by their systems as well as how they affect the performance of their systems. Such information allows the developers to fine-tune the system to optimize performance. Therefore, we normally provide a small sample of the collected data as a development data set to developers for their information. We avoid the system's being trained to a particular data set by testing with the complementary set of collected data. The entire collected biometric data set therefore is partitioned into two main categories: the development data and the test data.

There are three data sets involved in the building, testing, and operating of an automated biometrics-based identification or verification system: the development data, the test data, and the production data. The development data are used by the vendors for developing a system. The test data are used to evaluate the system. The production data are encountered by the system during its normal operation lifetime. Both development data and test data shall be collected under the same conditions and shall be considered representative of production data. Indeed, we collect one "master" set of samples without differentiating them. Only after we have collected the necessary amount of samples for both development and test data do we partition the samples into one development set and one test set. Our objective is to estimate the system performance parameters on the production data based on the measurements using the test data.

For the purpose of testing, we will collect a set of biometric data from each individual multiple times, since the collected biometric data often vary with time. For example, if we want to test an automated facial-recognition system, we select a group of n individuals from whom we want to collect facial images. We collect multiple facial images from each individual and denote the entire collection of facial images as a set G_k. Each element of G_k, a facial image, is identified by a unique encounter identification (EID), normally a string of numerals or characters. An EID identifies one instance of encountering an individual by the system. The entire collection of facial images, G, is the union of n mutually exclusive subsets $G_k,\ k = 1, 2, 3, \cdots, n$, i.e., $G = \cup_{k=1}^{n} G_k$. Each subset G_k contains the facial images of one individual

$$G_k = \{\text{EID}_{k1}, \text{EID}_{k2}, \text{EID}_{k3}, \cdots, \text{EID}_{kn_k}\}$$

where k denotes the kth individual and n_k denotes the number of images collected from this individual. Each individual in the group is identified by a person identification (PID), normally a string of numerals or characters, e.g., $\text{PID}(G_k) = A27069588$. In other words, each individual could have multiple EID's but only one PID. For simplicity, we will denote $\text{PID}(G_k)$ by PID_k in the following discussion.

The collected facial images, each represented by an EID, can be partitioned into a search subject set S and a file subject set F, $G = S \cup F$, where each element of S and F is a unique EID and $S \cap F = \emptyset$. Let S_{PID} be the set of PID's that each identifies an individual with at least one EID in S, the search set. Similarly, let F_{PID} be the set of PID's such

that each identifies an individual with at least one EID in F, the file set. In other words, the elements of $S_{\rm PID}$ and $F_{\rm PID}$ are $\rm PID_k$, $k = 1, 2, 3, \ldots$ while the elements of S and F are EID's. Since different EID's of an individual (one PID) can belong to S or F (exclusively), the intersection of $S_{\rm PID}$ and $F_{\rm PID}$ is not necessarily empty. For the purpose of testing, we will partition G into S and F such that $S_{\rm PID} \cap F_{\rm PID} \neq \emptyset$. The intersection of $S_{\rm PID}$ and $F_{\rm PID}$ can be expressed as

$$W = S_{\rm PID} \cap F_{\rm PID}.$$

In short, W is the set of PID's in the search set that have at least one mate PID in the file set.

III. Performance of Automated Biometrics-Based Identification and Verification Systems

As discussed Section II, the two main categories of automated biometrics-based systems are identification and verification. According to their functionality, they are often referred to as one-to-many and one-to-one matching, respectively. We will establish the performance criteria for the automated systems that perform these two types of functions in this section.

In an identification system (one-to-many), when an individual is encountered, an EID is issued. The individual's biometric data (e.g., fingerprints or facial image) are used to search against the data base (file set). If no mate is found, then this is the individual's first encounter with the system. The individual's biometric data are recorded in the data base as a new file subject and a PID is issued to identify the individual. If a mate is found, then the individual's identity is discovered.

In a verification system (one-to-one), when an individual is encountered, an EID is normally issued. In the first encounter of the individual with the system, the biometric data (e.g., fingerprints or facial image) is *enrolled* into the data base and an EID is issued for the encounter. A PID is then issued for this individual. This individual may be enrolled once or multiple times. If enrolled multiple times, multiple EID's will be issued. In future encounters, the individual's biometric data is compared with his enrollment data for verification.

A. Performance Criteria for a One-to-Many Matching System

In a search process of a one-to-many matching system, the biometric information of a search individual is matched against the biometric information of each individual in the data base. The search result is a list of candidates ranked in descending order of their similarity to the search individual. Typically, the similarity is represented by a numerical value referred to as a score. A threshold value is established, and all candidates in the list above this threshold are considered matches. Two important measures for the performance of such a system are 1) the percentage of time the system declares no match when a mate of the search individual is in the file set and 2) the percentage of time the system declares a match to the wrong PID. These two measures are defined in this paper as the Type I and Type II errors, respectively.

First, consider the definition of a miss. Assume that $\rm PID_{js}$ is the true identity of a given search subject and that this subject has at least one mate in the file set with $\rm PID_{jf}$. Let the EID of the search subject be $\rm EID_{is}$, where $\rm EID_{is} \in G_{js}$ and $\rm PID(G_{js}) = PID_{js}$. Note that the second index of G, PID, or EID indicates whether it is a search subject (s) or a file subject (f). For example, $\rm PID_{js}$ is a search subject whose mates are in $G_{\rm jf}$ and $\rm PID(G_{jf}) = PID_{jf}$. In this case, we have $\rm PID_{js} = PID_{jf}$. If the underlying automated biometric-based identification system (one-to-many) is not able to find any mates of $\rm EID_{is}$ in the file subject set $G_{\rm jf}$, a miss has occurred. It is expressed as

$$M_{1:M}(\rm EID_{is}) = \begin{cases} 1, & \text{if a miss occurs} \\ 0, & \text{otherwise.} \end{cases}$$

Then the conditional probability that $M_{1:m} = 1$ given $\rm PID_{js}$ is

$$\Pr(M_{1:m} = 1 | \rm PID_{js}) = \frac{1}{|G_{js}|} \sum_{\substack{\rm EID_{is} \in G_{js} \\ \rm PID(G_{js}) = PID_{js}}} M_{1:m}(\rm EID_{is})$$

where $|G_{\rm js}|$ is the total number of elements in $G_{\rm js}$. As a result, the total probability of $\Pr(M_{1:m} = 1)$ can be expressed as

$$\Pr(M_{1:m} = 1) = \sum_{\rm PID_{js} \in W} \Pr(M_{1:m} = 1 | \rm PID_{js}) \Pr(PID_{js}).$$

This is the definition of the Type I error rate $\rm T1_{1:m}$. It is important to realize that *a priori* probability $\rm Pr(PID_{js})$ refers to the probability of selecting $\rm PID_{js}$ in an operational setting, not the probability of selecting it from the search set S. Therefore, we assume that all the *a priori* probabilities are equal and have a value of $1/|W|$, where $|W|$ is the number of elements in the set W, i.e., $|W|$ is the number of PID's in the search subject set that have a mate PID in the file subject set. This assumption leads to the following expression for $\Pr(M_{1:m} = 1)$:

$$\Pr(M_{1:m} = 1) = \frac{1}{|W|} \sum_{\rm PID_{js} \in W} \Pr(M_{1:m} = 1 | \rm PID_{js}).$$

If one has more information regarding the *a priori* probabilities, then it can be used in the last equation. Last, the following definitions are given for $\rm T1_{1:m}$ error rate and reliability:

$$T1_{1:m} \equiv \Pr(M_{1:m} = 1)$$
$$R \equiv 1 - T1_{1:m}.$$

The $T1_{1:m}$ parameter is often referred to as $P_{\rm MD}$ (probability of missed detection). The parameter R is the reliability, or the probability that the system produces the correct result.

Now let us consider the definition of a false hit or false alarm. Assume that $\rm PID_{js}$ is the true identity of a given search subject and that this subject may or may not have mates in the file set. Let the EID of the search subject

be EID_{is}, where $\text{EID}_{\text{is}} \in G_{\text{js}}$ and $\text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}$. If the underlying automated biometric-based identification system matches the search subject (EID_{is}) with a file subject (EID_{mf}), which is not a mate of the search subject, then a false alarm has occurred. It is expressed as

$$\text{FA}_{1:m}(\text{EID}_{\text{is}}) = \begin{cases} 1, & \text{if a false hit occurs} \\ 0, & \text{otherwise.} \end{cases}$$

Then the conditional probability that $\text{FA}_{1:m} = 1$ given PID_{js} is

$$\Pr(\text{FA}_{1:m} = 1 | \text{PID}_{\text{js}}) = \frac{1}{|G_{\text{js}}|} \sum_{\substack{\text{EID}_{\text{is}} \in G_{\text{js}} \\ \text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}}} \text{FA}_{1:m}(\text{EID}_{\text{is}})$$

where $|G_{\text{js}}|$ is the total number of search subjects belonging to G_{js}. As a result, the total probability of $\text{FA}_{1:m} = 1$ can be expressed as

$$\begin{aligned} \Pr(\text{FA}_{1:m} = 1) &= \sum_{\text{PID}_{\text{js}} \in S_{\text{PID}}} \Pr(\text{FA}_{1:m} = 1 | \text{PID}_{\text{js}}) \Pr(\text{PID}_{\text{js}}) \\ &= \frac{1}{|S_{\text{PID}}|} \sum_{\text{PID}_{\text{js}} \in S_{\text{PID}}} \Pr(\text{FA}_{1:m} = 1 | \text{PID}_{\text{js}}) \end{aligned}$$

where S_{PID} is the set of PID's for the search subject set S and $|S_{\text{PID}}|$ is the total number of distinct subjects (PID's) in the search subject set. This is the definition of Type II error rate $\text{T2}_{1:m}$. The $\text{T2}_{1:m}$ parameter is often referred to as P_{FA} (probability of false alarm). Notice that we use S_{PID} in the last equation instead of W. This is because in the case of calculating misses, by definition, a mate must exist in the file set, whereas for false alarms, it is not required to have mates in the file set. Therefore, we can use all the unique PID's that exist in the search set.

B. Performance Criteria for a One-to-One Matching System

Consider the definition of a false reject as the case when an individual asserts his true identity and therefore retrieves the correct biometric data from the data base, but the system decides that the search data do not match the file data. More precisely, let the search subject be EID_{is}, where $\text{EID}_{\text{is}} \in G_{\text{js}}$ and $\text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}$, and, as in the treatment of the Type I error of the one-to-many matching system, PID_{js} and PID_{jf} denote the same individual in the search set and file set, respectively. In other words, $\text{PID}_{\text{js}} = \text{PID}_{\text{jf}}$. Now let EID_{mf}, where $\text{EID}_{\text{mf}} \in G_{\text{jf}}$ and $\text{PID}(G_{\text{jf}}) = \text{PID}_{\text{jf}}$, be the mate of EID_{is}. If the automated biometric-based verification system does not match EID_{is} with EID_{mf}, then a false reject has occurred. A false reject is a miss. In some cases, the automated system maintains multiple mates of the same PID (the same person). These multiple mates in the file set can be thought of as representing repeated trials during enrollment. If the search subject EID_{is} misses all the mates, then we say that a miss has occurred. Therefore, for a false reject (or a miss), we adopt a notation that is identical to the one-to-many case

$$M_{1:1}(\text{EID}_{\text{is}}) = \begin{cases} 1, & \text{if a miss occurs} \\ 0, & \text{otherwise.} \end{cases}$$

Similar to the one-to-many case, we can now determine the conditional probability that $M_{1:1} = 1$ given PID_{js}

$$\Pr(M_{1:1} = 1 | \text{PID}_{\text{js}}) = \frac{1}{|G_{\text{js}}|} \sum_{\substack{\text{EID}_{\text{is}} \in G_{\text{js}} \\ \text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}}} M_{1:1}(\text{EID}_{\text{is}}).$$

Again, following the development of $\text{T1}_{1:m}$, the total probability of $M_{1:1} = 1$ is expressed as

$$\begin{aligned} \Pr(M_{1:1} = 1) &= \sum_{\text{PID}_{\text{js}} \in W} \Pr(M_{1:1} = 1 | \text{PID}_{\text{js}}) \Pr(\text{PID}_{\text{js}}) \\ &= \frac{1}{|W|} \sum_{\text{PID}_{\text{js}} \in S_W} \Pr(M_{1:1} = 1 | \text{PID}_{\text{js}}) \end{aligned}$$

where $|W|$ is the number of elements in the set W, i.e., $|W|$ is the number of PID's in the search set that have a mate PID in the file set. This is the definition of $\text{T1}_{1:1}$ error rate. The $\text{T1}_{1:1}$ parameter is often referred to as P_{FR} (probability of false rejection).

Now consider the definition of a false acceptance. A false acceptance is the case when the system accepts the claim of an individual that he is of a given PID when in fact the individual is an impostor. More precisely, let the search subject be EID_{is}, $\text{EID}_{\text{is}} \in G_{\text{js}}$, and $\text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}$. Let PID_{kf} be the PID of a file subject and $\text{PID}_{\text{kf}} \neq \text{PID}_{\text{js}}$. Also assume that EID_{mf}, where $\text{EID}_{\text{mf}} \in G_{\text{kf}}$, is not a mate of EID_{is} ($\text{EID}_{\text{is}} \in G_{\text{js}}$). If the automated biometric-based verification system matches EID_{is} with EID_{mf}, then a false acceptance has occurred. Similar to the one-to-many case, we define the discrete random variable $\text{FA}_{1:1}$ as

$$\text{FA}_{1:1}(\text{EID}_{\text{is}}, \text{EID}_{\text{mf}}) = \begin{cases} 1, & \text{if a false acceptance occurs} \\ 0, & \text{otherwise.} \end{cases}$$

The conditional probability that $\text{FA}_{1:1} = 1$ given that PID_{js} fraudulently claims to be PID_{kf} is given by

$$\begin{aligned} \Pr(\text{FA}_{1:1} = 1 | \text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) &= \frac{1}{|GF_{\text{js}}| \cdot |G_{\text{kf}}|} \\ &\cdot \sum_{\substack{\text{EID}_{\text{is}} \in G_{\text{js}} \\ \text{PID}(G_{\text{js}}) = \text{PID}_{\text{js}}}} \sum_{\substack{\text{EID}_{\text{mf}} \in G_{\text{kf}} \\ \text{PID}(G_{\text{kf}}) = \text{PID}_{\text{kf}}}} \text{FA}_{1:1}(\text{EID}_{\text{is}}, \text{EID}_{\text{mf}}). \end{aligned}$$

The total probability that $\text{FA}_{1:1} = 1$ can thus be expressed as

$$\begin{aligned} T2_{1:1} &\equiv \sum_{\substack{all(\text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) \\ \text{PID}(G_{\text{js}}) \neq \text{PID}_{\text{kf}}}} \Pr(\text{FA}_{1:1} = 1 | \text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) \\ &\quad \cdot \Pr(\text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) \\ &\equiv \frac{1}{|F_{\text{PID}}| \cdot |S_{\text{PID}}| - |W|} \\ &\quad \cdot \sum_{\substack{all(\text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) \\ \text{PID}(G_{\text{js}}) \neq \text{PID}_{\text{kf}}}} \Pr(\text{FA}_{1:1} = 1 | \text{PID}_{\text{js}}, \text{PID}_{\text{kf}}) \end{aligned}$$

where $|F_{\text{PID}}|$ is the number of distinct PID's in the file set, $|S_{\text{PID}}|$ is the number of distinct PID's in the search set, and $|W|$ is the number of PID's in the search subject set that have at least one mate in the file set. $\text{T2}_{1:1}$ is the definition of Type II error rate in a one-to-one automated verification

system. The $T2_{1:1}$ parameter is often referred to as P_{FA} (probability of false acceptance). This is an unfortunate clash with the one-to-one notation; however, it is usually obvious for the context of the discussion.

IV. Performance Estimation and Confidence

We use Type I and Type II errors as the basic performance parameters to characterize automated biometrics-based identification and verification systems. The primary task of this section is to describe a strategy that produces reasonable estimates for these.

To estimate the Type I and Type II errors of an automated identification and verification system, we provide a set of test data for the system to process and collect the matching (comparison) results. We compare the matching results with the ground truth to produce the Type I and Type II error estimates. There are a number of issues to be addressed for this estimation process. First, what kind of test data should we use for parameter estimation? Second, how accurate and confident is the estimation for a given test data set? Third, how large should the test data set be in order to get a good estimate?

A. Collection of Test Data

The quality of the test data and the conditions under which the test data are collected will influence the outcome of the parameter estimation of an automated identification and verification system. Poor quality test data may not produce results that reflect the true performance of the system. Similarly, a very high quality data set may not reflect the true performance of an automated system.

To be able to produce the parameter estimation that characterizes the system performance under normal operating conditions, one needs to use, not surprisingly, a test data set collected under normal operating conditions. In other words, the test data set shall be collected under the same conditions as the normal operation. Using such test data, it is expected that the automated system would behave as if it were operating in a normal operating environment. A test data set with very poor quality could provide an indication of how the system might behave in the worst case scenario. A test data set with very high quality could provide an indication on how the system might behave in the best case scenario. Although this may be useful, we presently are addressing the issue of how to obtain an estimate of "realistic" system performance.

As an example, we briefly describe the considerations of collecting facial images for evaluating an automated facial-recognition system (AFRS). Assume that the AFRS will be used to recognize people passing through certain types of corridors. The few environmental factors that one may take into consideration when collecting the facial images include:

1) lighting conditions—light intensity, light source angle, and background light;
2) camera angles—azimuth angle and elevation angle;
3) weather—indoor/outdoor, dry, rain, or snow;
4) time—morning, noon, afternoon, evening, or night;
5) movement of the subject—static, fast moving, or slow moving;
6) surroundings—crowded, empty, single subject, or multiple subjects.

These factors at the chosen test data collection site must be similar to the actual operating environment in which the AFRS will operate. Otherwise, the test result will not reflect the true system performance.

Other considerations are technical factors, which include:

1) spatial resolution—the number of pixels representing a fixed size of area, such as 500 dots per inch;
2) gray-level resolution—the number of gray levels in the image (e.g., 256 is 8 bits per pixel);
3) image format—when the images are collected, do we compress them? If so, do we compress them using lossless compression or lossy compression?;
4) number of images collected from each individual—how many images from each individual do we need to collect? This consideration depends upon the applications of the AFRS to be tested.

These considerations help us to collect a facial-image set that can produce realistic performance estimation for the underlying AFRS.

B. Estimation Confidence

In this section, we describe the statistical tools used in our automated identification and verification system evaluation methodology. What is a parameter estimation? Parameter estimation is making a "best" guess of the value of a parameter based on the collection of outcomes of an experiment and recognizing the degree of confidence to be placed in the estimate. In an experiment, a sequence of identical and independent trials is repeated, each of which produces an outcome. If the outcome of each trial in the experiment depends on neither the outcomes of any of its predecessors nor any of its successors, then these are independent trials.

The experiment of independent trials is of particular interest to us for parameter estimation. It is intuitively clear that the outcome of a single trial can hardly represent any meaningful estimate of the underlying parameters. For example, by tossing a fair coin once and observing an outcome of "heads," it would be naive to claim that the coin is biased. However, a collection of outcomes of independent trials in an experiment can establish the basis upon which parameter estimation, a guess of a parameter, can be made. For example, tossing a biased coin 1000 times and observing 950 outcomes of "heads" would lead one reasonably to believe that the coin is very likely biased toward "heads."

As the number of independent trials in the experiment increases, one would reasonably expect that the collection

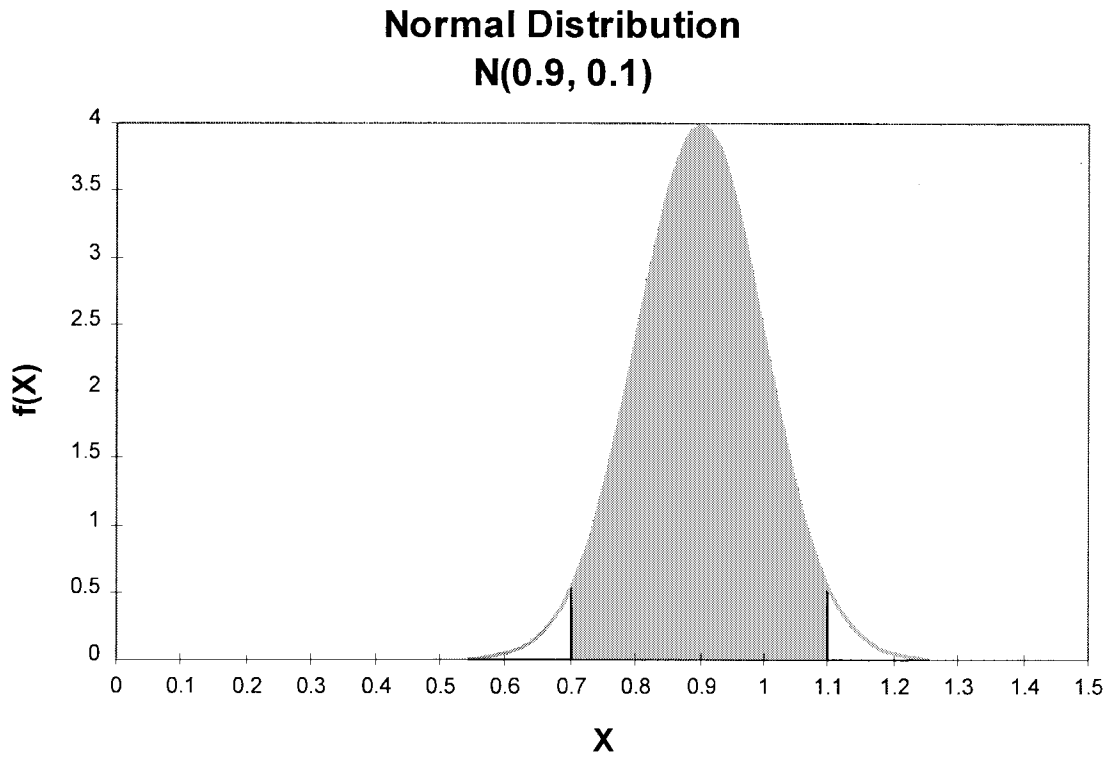


Fig. 1. A 95% confidence interval of the estimated parameter $p = 0.90$.

of the outcomes of the experiment might convey more meaningful information. Consequently, one would expect that conducting many independent trials would produce a reasonable assessment of the performance of an automated identification and verification system. Such observations form the basis of establishing the estimation confidence intervals and estimation errors.

Consider the case of conducting an experiment of one-to-many identification searches. The parameters that we wish to estimate are the Type I and Type II error rates, as defined in Section III. We will base our estimate on the outcomes of running an automated matching system on a set of test data. In the following discussions, we denote the parameters' Type I and Type II errors by P_{MD} (probability of missed detection) and P_{FA} (probability of false alarm), respectively. Using the outcomes of independent trials, we can form the estimates of the P_{MD} and P_{FA}, denoted by $\hat{P}_{\mathrm{MD}}$ and $\hat{P}_{\mathrm{FA}}$, respectively. The estimates for P_{MD} and P_{FA} are two single numbers. Each represents a "best" guess of a true parameter, P_{MD} and P_{FA}, respectively. As one might expect, the true parameter will most likely be numerically different from the best guess of the true parameter, but it would likely be in the neighborhood of the best guess. In this case, we wish to establish an interval around the best guess within which the true parameter would most likely be. Such an interval is known as the *confidence interval* in statistics. It is also commonly known as the margin of errors. The statistical term "confidence interval" is defined as the probability that the true parameter is within the interval that surrounds the estimate of the parameters P_{MD} and P_{FA}. Fig. 1 shows a 95% confidence interval around the estimated parameter $p = 0.90$.

The estimation of automated biometric-based identification and verification systems can be formulated as a parameter estimation problem based on the outcomes of repeated Bernoulli experiments. A Bernoulli experiment is a random experiment that has only two classes of outcome: success or failure. A sequence of Bernoulli trials is produced by performing a Bernoulli experiment several independent times with the same success rate p from trial to trial. When conducting the test of an automated system, each submitted search request will receive either a correct or an incorrect decision from the system. Since we are interested in the frequency of erroneous decisions made by the matcher, each incorrect decision is considered as an outcome of "success." If a search request received an incorrect match decision, an outcome of "success" is recorded for the trial. Otherwise, an outcome of "failure" is recorded for the trial.

Our objective is to estimate the frequency of "true" successes p based on the proportion of the outcomes of the Bernoulli trials that are successes. Assume that n identical independent Bernoulli trials are performed. Let Y denote the random variable that represents the total number of successes in n trials. Then Y is a binomial random variable $b(n,p)$, where Y/n is the observed proportion of successes in n trials, which can be used as an estimate of p for the probability of successes in n trials. Let the estimate for the probability of successes be denoted by $\hat{p} = Y/n$, a maximum likelihood estimator (MLE) [4]. Consequently,

for each trial, the estimate of probability of failures is $1-\hat{p}$. The probability that $y_1 \le Y \le y_2$ is expressed as

$$\Pr\{y_1 \le Y \le y_2\} = \sum_{y=y_1}^{y_2} \binom{n}{y} p^y (1-p)^{n-y}. \quad (1)$$

When n is sufficiently large $(n \ge 30)$ and p is neither too large nor too small, the binomial distribution can be well approximated by a normal distribution [4]. To approximate a random variable Y of binomial distribution by a random variable of normal distribution, we observe that Y can be transformed to a standard normal random variable Z by

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \sim N(0,1). \quad (2)$$

In the case of normal distributed random variables, the probability density function (pdf) is symmetric about the mean. Thus, we can let set $Y - y_1 = y_2 - Y$. As a result, we can approximate the probability that $y_1 \le Y \le y_2$ as

$$\begin{aligned}
&\Pr\{y_1 \le Y \le y_2\} \\
&= \Pr\left\{ \frac{(y_1/n) - p}{\sqrt{p(1-p)/n}} \le \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \le \frac{(y_2/n) - p}{\sqrt{p(1-p)/n}} \right\} \\
&\approx \Pr\left\{ -z_{\alpha/2} \le \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \le z_{\alpha/2} \right\} \\
&= \Pr\Big\{ (Y/n) - z_{\alpha/2}\sqrt{p(1-p)/n} \le p \\
&\qquad \le (Y/n) + z_{\alpha/2}\sqrt{p(1-p)/n} \Big\} \\
&= 1 - \alpha \qquad (3)
\end{aligned}$$

where $1-\alpha$ is the confidence coefficient. In Fig. 1, the shaded area is $1-\alpha$. This expression indicates that we are $(1-\alpha) \times 100\%$ confident that the true value of p is somewhere in the confidence interval $\left[(y/n) - z_{\alpha/2}\sqrt{p(1-p)/n}, (y/n) + z_{\alpha/2}\sqrt{p(1-p)/n}\right]$ when $Y = y$. The true parameter p, however, is generally not known in advance. To determine the confidence interval for the estimate, we normally replace p by its estimate $\hat{p} = (Y/n)$, which results in the confidence interval

$$\Big[(y/n) - z_{\alpha/2}\sqrt{(y/n)(1-(y/n))/n}, \\ (y/n) + z_{\alpha/2}\sqrt{(y/n)(1-(y/n))/n}\Big]$$

when $Y = y$. One can observe that the width of the confidence interval is determined by the number of trials conducted in the experiment, n, if y/n is known. For a given confidence coefficient, a larger number of trials results in a narrower confidence interval. This observation outlines the important relationship between the size of the test data sample and our confidence in the result.

On the other hand, if n is sufficiently large and p is close to either zero or one, the binomial distribution can be well approximated by a Poisson distribution [4]. Similarly, to approximate a random variable Y of binomial distribution $b(n,p)$ by a random variable of Poisson distribution, we observe that [4]

$$\begin{aligned}
\Pr\{Y = y\} &= \binom{n}{y} p^y (1-p)^{n-y} \\
&= \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\
&= \frac{n!}{(y-1)!(n-y+1)!} \\
&\quad \cdot \frac{(n-y+1)}{y} p^{y-1} (1-p)^{n-y+1} \cdot \frac{p}{1-p} \\
&= \frac{(n-y+1)}{y} \Pr(Y = y-1) \frac{p}{1-p}. \qquad (4)
\end{aligned}$$

When p is small and n is large, the above expression can be written as

$$\Pr\{Y = y\} \approx \frac{np}{y} \Pr(Y = y-1). \quad (5)$$

On the other hand, a Poisson distribution can be expressed as [4]

$$\begin{aligned}
\Pr\{X = x\} &= \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \frac{e^{-\lambda}\lambda^{x-1}\lambda}{(x-1)!x} = \frac{\lambda}{x} \Pr(X = x-1). \quad (6)
\end{aligned}$$

It follows that a random variable of binomial distribution $b(n,p)$ can be approximated by a random variable of Poisson distribution of mean $=\lambda$, where $np = \lambda$. Note that the parameter to be estimated in this case is λ. Let λ_L and λ_U be the lower and upper bounds of the $100(1-\alpha)\%$ confidence interval for estimating λ, respectively. λ_L and λ_U can be obtained by solving the equations [5]

$$e^{-\lambda_U} \sum_{k=0}^{x} \frac{\lambda_U^k}{k!} = \alpha/2$$

and

$$e^{-\lambda_L} \sum_{k=0}^{x} \frac{\lambda_L^k}{k!} = \alpha/2. \quad (7)$$

Although (7) does not appear to have closed-form solutions, λ_L and λ_U can be obtained using an automated solving algorithm. A frequently used approximation solution for λ_L and λ_U is obtained using a normal approximation to the Poisson distribution, when λ is expected to be fairly large. In that case, the confidence interval for λ is

$$\begin{aligned}
[\lambda_L, \lambda_U] = \Big[&X + 0.5Z_{\alpha/2}^2 + Z_{\alpha/2}\sqrt{X + 0.25Z_{\alpha/2}^2}, \\
&X + 0.5Z_{\alpha/2}^2 - Z_{\alpha/2}\sqrt{X + 0.25Z_{\alpha/2}^2}\Big]. \quad (8)
\end{aligned}$$

As discussed previously, our objective is to estimate the values of P_{MD} and P_{FA}. To estimate P_{MD}, let the number of search subjects (each with a mate in the file subjects) be n, the probability that the search subject does not match its mate in the file be p, and the number of "successes" (a search subject does not match its mate in the file) be Y.

Denote the estimation of P_{MD} by $\hat{P}_{\mathrm{MD}} = Y/n$. It follows that

$$\begin{aligned}
&\Pr\{y_1 \le Y \le y_2\} \\
&\quad = \sum_{y=y_1}^{y_2} \binom{n}{y} p^y (1-p)^{n-y} \\
&\quad \approx \Pr\Big((Y/n) - z_{\alpha/2}\sqrt{p(1-p)/n} \\
&\qquad \le p \le (Y/n) + z_{\alpha/2}\sqrt{p(1-p)/n} \Big) \\
&\quad = \Pr((Y/n) - \varepsilon \le p \le (Y/n) + \varepsilon) \\
&\quad = 1 - \alpha
\end{aligned} \tag{9}$$

where $p \approx Y/n$, 2ε is the confidence interval width, and α is the level of confidence. The estimation confidence interval for P_{MD} is $[y/n - \epsilon, y/n + \varepsilon]$ if $Y = y$.

To estimate P_{FA}, let the number of search subjects (with or without a mate in the file subjects) be n, the number of file subjects be m, and the probability that a particular file subject is incorrectly matched to a particular search subject be p_0 (the estimation of p_0 is described in Section IV-C). As a result, the probability that a given search subject matched against m file subjects produces at least one false hit is

$$\begin{aligned}
\Pr(1 \le Y \le m) &= \sum\nolimits_{y=1}^{m} \binom{m}{y} p_0^y (1-p_0)^{m-y} \\
&= 1 - \binom{m}{0} p_0^0 (1-p_0)^m \\
&= 1 - (1-p_0)^m = p.
\end{aligned} \tag{10}$$

Note that in (10), p is indeed the definition of P_{FA}, which demonstrates that P_{FA} is a function of m and p_0. The *invariance property MLE's* [4] states that if $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau\left(\hat{\theta}\right)$. Consequently, the probability that exactly one search subject among n of them produces at least one false hit is

$$\Pr(Y = 1) = \binom{n}{1} p(1-p)^{n-1}. \tag{11}$$

It follows that the probability of exactly k search subjects' each having at least one false hit can be expressed as

$$\Pr(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{12}$$

We can then obtain the probability that the number of search subjects each having at least one false hit, Y, is in the interval $[y_1, y_2]$

$$\Pr\{y_1 \le Y \le y_2\} = \sum\nolimits_{k=y_1}^{y_2} \binom{n}{k} p^k (1-p)^{n-k} \tag{13}$$

where p was defined as $p = 1 - (1-p_0)^m$. Let the estimated P_{FA} be $\hat{P}_{\mathrm{FA}} = Y/n$. Again, using a normal approximation, we can rewrite it as

$$\begin{aligned}
&\Pr\{y_1 \le Y \le y_2\} \\
&\quad = \Pr\Big\{ (Y/n) - z_{\alpha/2}\sqrt{p(1-p)/n} \le p \le (Y/n) \\
&\qquad + z_{\alpha/2}\sqrt{p(1-p)/n} \Big\} \\
&\quad = \Pr\{(Y/n) - \varepsilon \le p \le (Y/n) + \varepsilon\} \\
&\quad = 1 - \alpha
\end{aligned} \tag{14}$$

where $1 - \alpha$ is the confidence coefficient. Since the value of $\sqrt{p(1-p)/n}$ is not known in advance, it is replaced by an approximation $\sqrt{(y/n)(1-y/n)/n}$ when computing the confidence interval. The estimation confidence interval for P_{FA} is $[(y/n) - \epsilon, (y/n) + \varepsilon]$ if $Y = y$ and $\varepsilon = z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n}$.

In (14), it is observed that both the number of file subjects, m, and that of search subjects, n, can influence the confidence interval width of the estimated P_{FA}.

C. Sample Size of the Test Data

In the last section, we formulated the estimators for P_{MD} and P_{FA}, denoted as $\hat{P}_{\mathrm{MD}}$ and $\hat{P}_{\mathrm{FA}}$, as well as their respective confidence intervals. In this section, we formulate the sample sizes for estimating P_{MD} and P_{FA} with the specified confidence intervals.

For a binomial random variable $\hat{p} = Y/n$, the confidence interval for estimating p is shown in (15) (shown at the bottom of the page) if $Y = y$. The relationship between the test data sample size n and the confidence interval 2ε is given as

$$\varepsilon = z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n} \tag{16}$$

which can be rewritten as

$$n \approx \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 (y/n)(1-y/n). \tag{17}$$

n is the test data sample size needed to achieve the specified confidence interval $[(y/n) - \epsilon, (y/n) + \varepsilon]$. To solve for n, one needs to estimate the value of (Y/n) from the available test data sample prior to the initiation of this task. In other words, one might substitute (Y/n) *a priori*. Fig. 2. demonstrates the relationships between the size of the test data sample and the confidence intervals.

Recall that the estimation of P_{MD} can be approximated by

$$\begin{aligned}
&\Pr\{y_1 \le Y \le y_2\} \\
&\approx \Pr\Big((Y/n) - z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n} \le p \\
&\qquad < (Y/n) + z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n} \Big) \\
&= \Pr\{(Y/n) - \varepsilon \le p \le (Y/n) + \varepsilon\}.
\end{aligned}$$

$$\left[(y/n) - Z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n},\ (y/n) + Z_{\alpha/2}\sqrt{(y/n)(1-y/n)/n}\right] = [(y/n) - \varepsilon, (y/n) + \varepsilon] \tag{15}$$

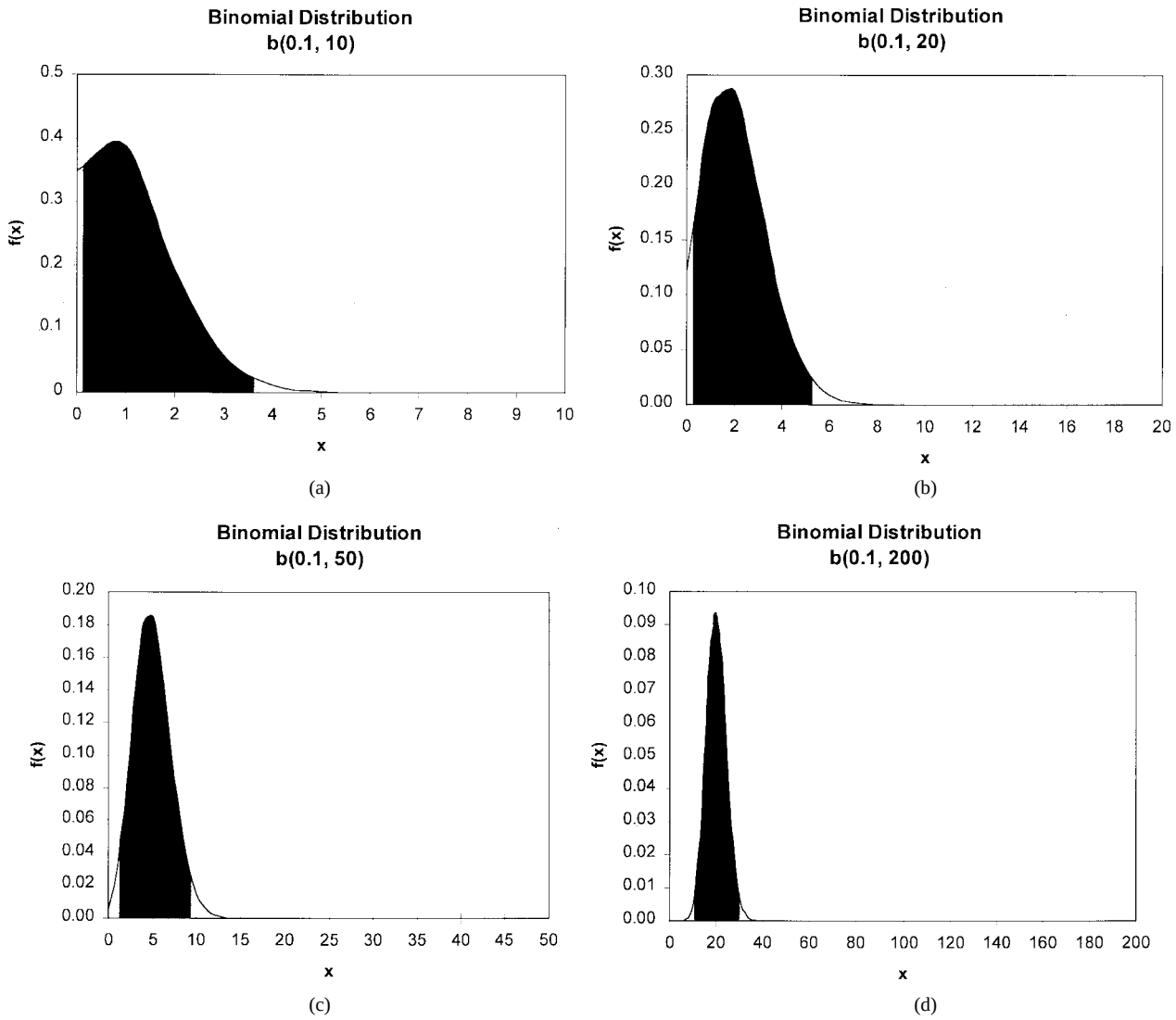


Fig. 2. The 95% confidence intervals of an estimated parameter from n trials. (a) $n = 10$, b(0.1, 10). (b) $n = 20$, b(0.1, 20). (c) $n = 50$, b(0.1, 50). (d) $n = 200$, b(0.1, 200).

Combining this probability formulation and the aforementioned sample size estimation formulation, we can obtain

$$n \approx \left(\frac{z_{\alpha/2}\sqrt{(y/n)(1-y/n)}}{\varepsilon} \right)^2 \tag{18}$$

where ε is the half-width of the confidence interval. Thus, the test data sample size needed to estimate P_{MD}, with $(1-\alpha) \times 100\%$ confidence with a 2ε confidence interval, is given by (18).

On the other hand, since $\hat{P}_{\mathrm{FA}}$ increases as m increases, as shown in (10) and (13), varying m results in different $\hat{P}_{\mathrm{FA}}$. It follows that when calculating $\hat{P}_{\mathrm{FA}}$, m must be set to the number of file records on which the user intended the system to operate. In many cases, it is often impractical to have a test data set as large as the production data set, and the estimation of $\hat{P}_{\mathrm{FA}}$ becomes very useful. This observation leads us to consider p_0 as the parameter to be directly estimated, since $\hat{P}_{\mathrm{FA}}$ is a function of p_0.

The users of an automated identification system normally specify the system accuracy requirements in terms of P_{MD} and P_{FA}. Therefore, we will express p_0 in terms of P_{FA} in the following analysis. Rewriting (10), we obtain $p_0 = 1 - \sqrt[m]{1-p}$, where p is P_{FA}.

In Section IV-B, we defined the number of search subjects (with or without a mate in the file subjects) as n, the number of file subjects as m, and the probability that a file subject is incorrectly matched to a search subject as p_0. If the automated biometrics-based identification system compares each search subject with each file subject in the data base, there are a total of $m \times n$ comparisons. Let us further assume that there are a total of x instances in which a search subject is incorrectly matched with a file subject. The probability that a search subject is incorrectly matched to a file subject, p_0, can be estimated from the frequency

$$\hat{p}_0 \approx \frac{x}{m \times n}. \tag{19}$$

Note that the number of incorrect matches produced by a system when comparing n search subjects and m file subjects, X, is a binomial random variable $b(p_0, nm)$, where $X/(nm)$ is the observed proportion of successes

in $n \times m$ trials, which can be used as an estimate of p_0. Each comparison of a search subject and a file subject is an identical independent Bernoulli trial. The probability that $x_1 \le X \le x_2$ can be expressed as

$$\begin{aligned}\Pr\{x_1 \le X \le x_2\} &= \sum_{x=x_1}^{x_2} \binom{nm}{x} \hat{p}_0^x (1-\hat{p}_0)^{nm-x} \\ &\approx \Pr\left\{ \left(\frac{X}{nm}\right) - z_{\alpha/2}\sqrt{\left(\frac{X}{nm}\right)\left(1-\frac{X}{nm}\right)/(nm)} \right. \\ &\qquad \le p_0 \le \left(\frac{X}{nm}\right) \\ &\qquad \left. + z_{\alpha/2}\sqrt{\left(\frac{X}{nm}\right)\left(1-\frac{X}{nm}\right)/(nm)} \right\}. \end{aligned} \tag{20}$$

The confidence interval is given as $[\hat{p}_0 - \varepsilon, \hat{p}_0 + \varepsilon]$, where $\hat{p}_0 = x/(m \times n)$. The number of comparisons needed to achieve the estimation confidence of $(1-\alpha)\times 100\%$ can be determined from $\varepsilon = z_{\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)/(nm)}$. It follows that the product of m and n is

$$nm = \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 (\hat{p}_0(1-\hat{p}_0)). \tag{21}$$

It is clear that the estimation confidence interval depends on the product of m and n. It can be further observed that if m is fixed due to the cost and availability of the file subjects, one might use n search subjects, each of which has no mate in the file subjects, to compensate for the limitation on m.

D. Hypothesis Testing

In previous sections, we formulated the estimators for P_{MD} and P_{FA}, their respective confidence intervals, and the sample sizes needed to achieve the specified estimation confidences. The next step in selecting an automated identification and verification system for our system integration effort is to determine whether the underlying system meets our system requirement. In fact, we wish to test the hypothesis that the performance of the automated system under consideration meets or exceeds the system design requirement.

Assume that we have determined that the number of search subjects (the number of independent trials) needed to achieve certain estimation confidence within a given confidence interval is n. Consider the performance parameter $P_{\mathrm{MD}} = \hat{\theta}$, which was to be estimated from the system test as $\hat{\theta} = 1/n\sum_{k=0}^{n} x_k$, where x_k is zero if no detection is missed at the kth trial and x_k is one if a detection is missed at the kth independent trial. In addition, assume that the system design requirement for P_{MD} is to be less than or equal to p_0, where p_0 is some prespecified requirement. Note that under both hypotheses, the performance parameter $P_{\mathrm{MD}} = \hat{\theta}$ is a random variable of binomial distribution. Our objective is to find out if the hypothesis $\hat{\theta} \le p_0$ is true, based on the output of a sequence of independent trials.

Denote the null hypothesis H_0 as $\hat{\theta} \le p_0$ and the alternative hypothesis H_1 as $\hat{\theta} > p_0$. Let $\mathbf{X} = [X_1, X_2, X_3, \cdots, X_n]$ be a random vector with pdf (probability mass function—pmf) p_θ, where each element x_i is an outcome of the independent trials of an experiment (either zero or one). The problem of hypothesis testing is formulated as testing the null hypothesis H_0: $\mathbf{X} \sim p_\theta, \theta \le p_0$ against the alternatives H_1: $\mathbf{X} \sim p_\theta, \theta > p_0$.

There are a number of ways to perform the hypotheses testing for the above formulation [6]. We choose the likelihood ratio test (LRT) for its practicality. A likelihood functional is a family of functions

$$p_\theta(x_1, x_2, x_3, \cdots, x_n) = \prod_{k=1}^{n} p_\theta(x_k) \tag{22}$$

for a given function form p_θ, where θ is an unknown parameter. To describe the LRT procedure for the hypotheses formulated above, we first define the likelihood ratio

$$\lambda(X) = \frac{\sup_{\theta\in\Theta_0} p_\theta(x_1, x_2, \cdots, x_n)}{\sup_{\theta\in\Theta} p_\theta(x_1, x_2, \cdots, x_n)} \tag{23}$$

where $\Theta_0 = [0, p_0]$ and $\Theta = [0, 1]$. This ratio varies from zero to one, $\lambda(\boldsymbol{X}) \in [0, 1]$. A larger $\lambda(\boldsymbol{X})$ implies that the likelihood of $\theta \in \Theta_0$ is higher, while a smaller $\lambda(\boldsymbol{X})$ implies that the likelihood of $\theta \in \Theta_0$ is lower. In other words, the likelihood of H_0 being true increases as the value of $\lambda(\boldsymbol{X})$ increases. An LRT of testing H_0 against H_1 is to reject H_0 if and only if $\lambda(\boldsymbol{X}) < r$, where r is some constant.

We are interested in finding a threshold r that can determine whether we should accept the null hypothesis H_0 based on the observed value $\hat{\theta}$, with a prespecified error rate α. This prespecified error rate α is the error when the null hypothesis H_0 is rejected, while $\theta \le p_0$ (the null hypothesis H_0 is true). It is often known as the size (or level) of the LRT and is expressed as

$$\alpha = \sup_{\theta\in\Theta_0} \Pr_\theta\{\mathbf{X} : \lambda(\mathbf{X}) < r\}. \tag{24}$$

One can determine the value of r from this expression. Note that if the elements of the random vector $\boldsymbol{X} = [X_1, X_2, X_3, \cdots, X_n]$ are the outcomes of Bernoulli trials, then $Y = \sum_{k=1}^{n} X_k$ is a random variable of binomial distribution, $Y \sim b(n, p_\theta)$. The size of the LRT, i.e., the probability of rejecting H_0 while H_0 is true, is given by

$$\alpha = \Pr\{\hat{\theta} > r'; p_0\} = \sum\nolimits_{y=r+1}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y}. \tag{25}$$

Let us consider a specific example to illustrate the procedure of determining r. Assume that the system "accuracy" performance requirement is $p_0 = 0.02$, there are 1000 Bernoulli trials, and the size of the hypotheses test is $\alpha = 0.05$. Assume that we have performed a system test and observed a sample of $X_1, X_2, X_3, X_4, \cdots, X_{1000} \sim iid\, b(1, p_\theta)$. Let $Y = \sum_{k=1}^{1000} X_k$ and $Y \sim b(1000, p_\theta)$. We will perform the level α LRT of H_0: $X \sim p_\theta, \theta \le p_0$ against the alternatives H_1: $X \sim p_\theta, \theta > p_0$.

The likelihood ratio for a binomial distribution is established as

$$\lambda(y) = \frac{\sup_{\hat{p}\le p_0} \binom{n}{y} \hat{p}^y (1-\hat{p})^{n-y}}{\sup_{0\le\hat{p}\le 1} \binom{n}{y} \hat{p}^y (1-\hat{p})^{n-y}} = \frac{\sup_{\hat{p}\le p_0} \hat{p}^y (1-\hat{p})^{n-y}}{\sup_{0\le\hat{p}\le 1} \hat{p}^y (1-\hat{p})^{n-y}}. \quad (26)$$

The value of $\hat{p}$ normally is unknown in advance, and it is often approximated by $\hat{p} = y/n$. However, it is easy to verify that the denominator of $\lambda(y)$ attains the maximum when $\hat{p} = y/n$

$$\sup_{0\le\hat{p}\le 1} \hat{p}^y (1-\hat{p})^{n-y} = \left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}. \quad (27)$$

This function is plotted in Fig. 3(a). Similarly, the numerator of $\lambda(y)$, $\sup_{\hat{p}\le p_0} \hat{p}^y (1-\hat{p})^{n-y}$, attains the maximum when $\hat{p} = y/n$ and $p_0 \ge y/n$. Hence, if $p_0 \ge y/n$, we can obtain

$$\sup_{\hat{p}\le p_0} \hat{p}^y (1-\hat{p})^{n-y} = \left(\frac{y}{n}\right)^y \left(1-\frac{y}{n}\right)^{n-y}. \quad (28)$$

When $p_0 < y/n$, the numerator attains its maximum at $\hat{p} = p_0$

$$\sup_{\hat{p}\le p_0} \hat{p}^y (1-\hat{p})^{n-y} = (p_0)^y (1-p_0)^{n-y}. \quad (29)$$

As a result, the likelihood ratio $\lambda(y)$ can be expressed as

$$\lambda(y) = \begin{cases} \frac{p_0^y (1-p_0)^{n-y}}{(\frac{y}{n})^y (1-\frac{y}{n})^{n-y}} & p_0 < \frac{y}{n}, \\ 1 & p_0 \ge \frac{y}{n}. \end{cases} \quad (30)$$

This likelihood ratio $\lambda(y)$ is plotted in Fig. 3(b). It is observed that $\lambda(y)$ is a nonincreasing function of y, which implies that the likelihood of H_0 being true increases as y decreases. To determine the "threshold" r, where $\alpha = \sup_{\theta\in\Theta_0} \Pr_\theta\{Y : \lambda(y) < r\}$, consider the probability that $\hat{\theta}(= y/n) > r$ given p_0

$$\alpha = \Pr\{\hat{\theta} > r'; p_0\} = \sum\nolimits_{y=nr'+1}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y}. \quad (31)$$

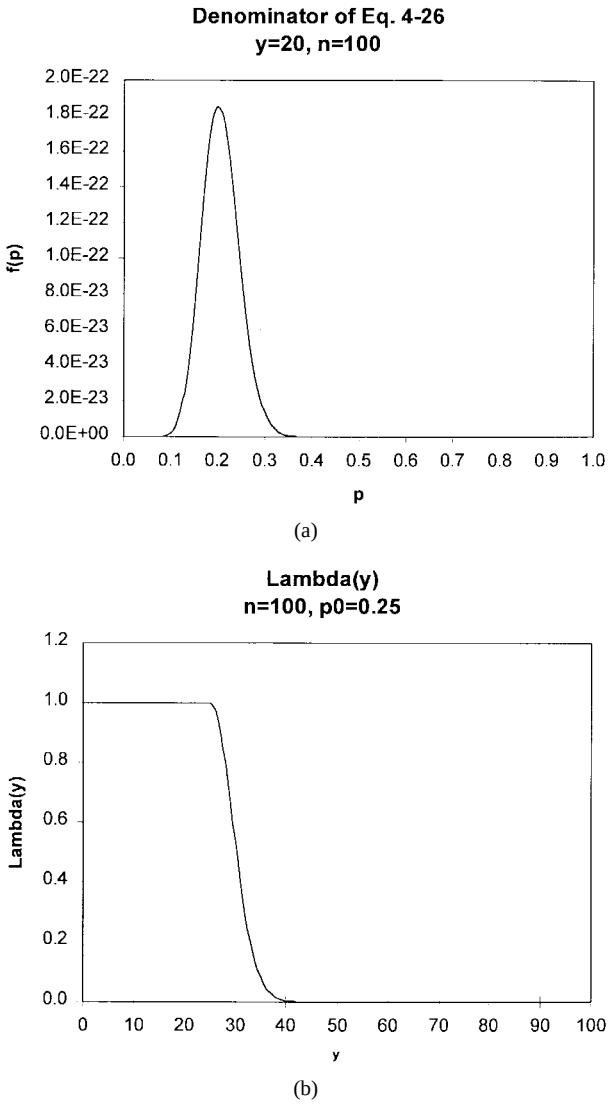
It follows that the "threshold," r', can be obtained from

$$0.05 = \sum\nolimits_{y=nr'+1}^{1000} \binom{1000}{y} 0.02^y (1-0.02)^{1000-y}. \quad (32)$$

From a binomial distribution table, we found that when $nr' = 27$

$$\sum\nolimits_{y=27+1}^{1000} \binom{1000}{y} 0.02^y (1-0.02)^{1000-y} = 0.051 \quad (33)$$

which is not the exact solution for $\alpha = 0.05$, since y is a discrete random variable. It follows that $r' = 27/1000 = 0.027$. Now recall that $\hat{\theta} = y/n$. We would reject the null hypothesis if the total number of missing detection, y, exceeds the threshold, i.e., $y > 27$. On the other hand, if the total number of missing detection is less than or equal to 27, i.e., $y \le 27$, then the null hypothesis would be accepted.



(a)

(b)

Fig. 3. The likelihood ratio for a binomial distribution. (a) The denominator of (26). (b) The likelihood ratio (26).

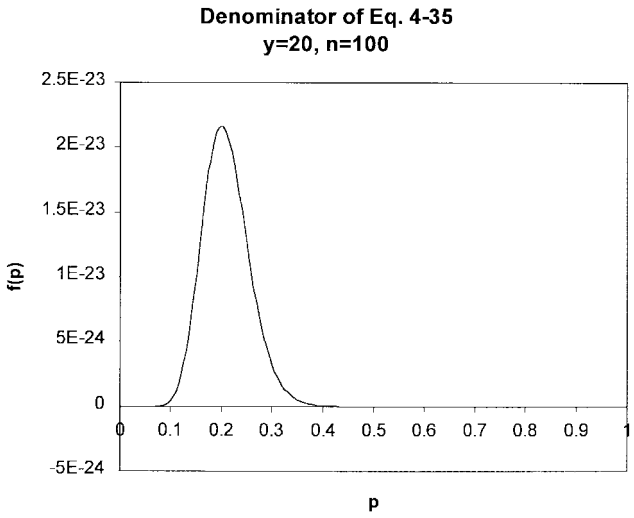
An alternative solution to the above hypotheses testing is to use an approximate pmf for the binomial mass function. When n is large and p_0 is small ($np_0 \le 10$), the binomial pmf normally can be approximated by a Poisson pmf with $\lambda = np_0$

$$\alpha \approx \Pr\{Y > r; p_0\} = \sum\nolimits_{y=r+1}^{n} \frac{\lambda^y e^{-\lambda}}{y!}. \quad (34)$$

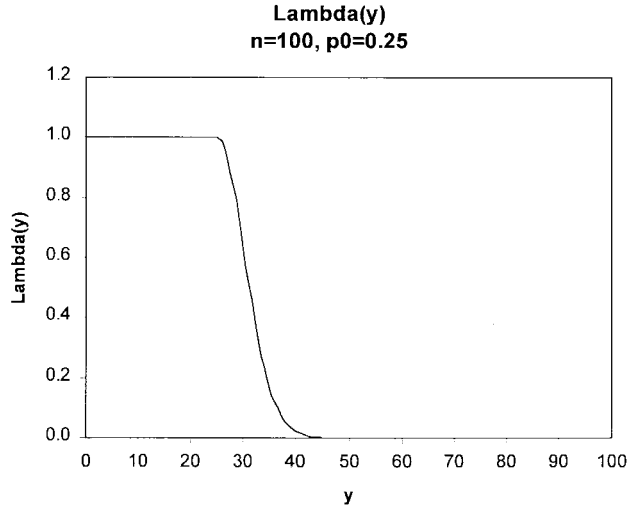
The likelihood ratio for a Poisson distribution is established as

$$\lambda(y) = \frac{\sup_{\hat{p}\le p_0} \frac{\lambda^y e^{-\lambda}}{y!}}{\sup_{0\le\hat{p}\le 1} \frac{\lambda^y e^{-\lambda}}{y!}} = \frac{\sup_{\hat{p}\le p_0} \lambda^y e^{-\lambda}}{\sup_{0\le\hat{p}\le 1} \lambda^y e^{-\lambda}} = \frac{\sup_{\hat{p}\le p_0} \hat{p}^y e^{-n\hat{p}}}{\sup_{0\le\hat{p}\le 1} \hat{p}^y e^{-n\hat{p}}}. \quad (35)$$

It is easy to verify that the denominator of $\lambda(y)$ attains the maximum when $\hat{p} = y/n$, and the numerator of



(a)



(b)

Fig. 4. The likelihood ratio for a Poisson distribution. (a) The denominator of (35). (b) The likelihood ratio in (35).

$\lambda(y)$ also attains the maximum when $\hat{p} = y/n$ if $p_0 \geq y/n$. The denominator of $\lambda(y)$ is plotted in Fig. 4(a). As demonstrated in Fig. 4(a), the numerator of $\lambda(y)$ attains its maximum at $\hat{p} = p_0$ if $p_0 < y/n$. As a result, the likelihood ratio $\lambda(y)$ for the Poisson distribution can be expressed as

$$\lambda(y) = \begin{cases} \frac{p_0^y e^{-np_0}}{(y/n)^y e^{-y}} & y > np_0, \\ 1 & y \leq np_0. \end{cases} \tag{36}$$

This likelihood ratio $\lambda(y)$ is plotted in Fig. 4(b). It is observed that $\lambda(y)$ is a nonincreasing function of y, which implies that the likelihood of H_0 being true increases as y decreases. To determine the threshold r, where $\alpha = \sup_{\theta \in \Theta_0} \Pr_\theta\{X : \lambda(y) < r\}$, consider the probability that $\hat{\theta} > r$ given p_0

$$\alpha = \Pr\left\{\hat{\theta} > r; p_0\right\} = \sum_{y=r'+1}^{n} \frac{(np_0)^y e^{-np_0}}{y!}. \tag{37}$$

Table 1 Relevant System Requirements for Selecting and AFIS

Number of Fingers Used	P_{MD}	P_{FA}	Database Size
1	5%	10%	10 000 000

It follows that the "threshold" can be obtained from

$$0.05 = \sum_{y=r'+1}^{1000} \frac{20^y e^{-20}}{y!}. \tag{38}$$

From a Poisson distribution table, we found that when $r' = 27$

$$\sum_{y=28}^{1000} \frac{20^y e^{-20}}{y!} = 0.052$$

which is not the exact solution for $\alpha = 0.05$, since y is a discrete random variable. Note that the threshold value $r' = 27$ is the same in both the direct binomial distribution and the indirect Poisson distribution approaches. If the number of missing detection is more than 27, then we reject the null hypothesis H_0 with no more than 5.2% error.

After we perform the Bernoulli trials (e.g., using an AFIS to perform a number of fingerprints matches), we count the total number of detections missed. Each missing detection is recorded when a search print (with a known mate in the data base) does not find its "true" mate after being searched against the entire data base. If the total number of missing detection exceeds 27, the "threshold" r', then we conclude that the underlying AFIS does not achieve the required accuracy of $P_{\text{MD}} \leq 0.02$.

Although the example that we used to illustrate the hypotheses testing is for P_{MD}, the same approach works for P_{FA}.

V. An Example: Selection of an AFIS

The evaluation methodology developed in previous sections is applicable to various automated biometrics identification and verification system "accuracy" evaluations. In this section, we provide a specific example of AFIS evaluation to illustrate the complete automated biometrics-based identification and verification system evaluation process.

This example illustrates the process of testing and selecting an AFIS for a hypothetical application (one-to-many searches). Some of the relevant system requirements are summarized in Table 1.

Also assume that these specified parameters are to be estimated with 95% confidence and ±1% margin of error for the AFIS under consideration. The objective of a test is to estimate the AFIS performance parameters of an AFIS and to determine if it can meet the specified requirements.

A. Sample Size of the Test Data

Before conducting the test of an AFIS, we first need to determine the size of the test data set. How many individuals' fingerprints shall there be in the search set (the number of search subjects)? How many individuals' fingerprints shall there be in the file set (the number of file

subjects)? To determine the number of search subjects for testing the AFIS under consideration, recall (18)

$$n = \left(\frac{z_{\alpha/2}\sqrt{(y/n)(1-y/n)}}{\varepsilon} \right)^2.$$

From the system requirements, we know that $\alpha/2 = 0.025$ and $\varepsilon = 0.01$. Using a normal distribution table, we obtain $z_{\alpha/2} = z_{0.025} = 1.96$. The system requirements of the AFIS evaluation, provided that $(y/n)(1-y/n) \approx 0.05 \times 0.95 = 0.0475$, are

$$n = \left(\frac{z_{\alpha/2}\sqrt{(y/n)(1-y/n)}}{\varepsilon} \right)^2 = \frac{1.96 \times 1.96 \times 0.0475}{0.01 \times 0.01} \approx 1825. \quad (39)$$

This calculation tells us that at least 1825 search subjects are needed to obtain an estimation of P_{MD} with 95% confidence and $\pm 1\%$ maximum error of the estimate. To perform 1825 searches, one needs to collect pairs of fingerprints from 1825 individuals. Note that different individuals normally have different fingerprint images, and the automated identification systems searches fingerprints from different subjects against the data base under normal operating conditions. If search prints are all collected from very few individuals, i.e., multiple search prints from each individual, it is likely that the system performance measured might be biased toward a certain small group of individuals. On the other hand, if the search prints are collected from 1825 different individuals, it is unlikely that these search prints all are similar, which allows one to measure the underlying system performance more "realistically." Each of these 1825 searches is to compare a search print with all the prints in the file print set, where each search print has a mate. In the fingerprint identification community, it means that 1825 mated fingerprints are to be collected for the evaluation purpose. These mated fingerprints are mainly used for estimation of the AFIS parameter P_{MD}.

To estimate the AFIS parameter P_{FA}, we collect some fingerprints of no mates for the search data set. As discussed in Section IV-C, we will not directly estimate P_{FA}. Instead, we estimate the probability that a search subject is incorrectly matched to a file subject, p_0, using (19).

First, let us determine what the value of $\hat{p}_0$ will be assuming that $\hat{P}_{\mathrm{FA}} \approx 0.1$ (from the system requirements) and $m = 10\,000\,000$ since $\hat{P}_{\mathrm{FA}}$ is not known in advance. m is the number of file subjects specified in the system requirements. Recall (10), $\hat{P}_{\mathrm{FA}} \approx 1 - (1-\hat{p}_0)^m$. We can solve it for $\hat{p}_0$

$$\hat{p}_0 \approx 1 - \sqrt[m]{1 - \hat{P}_{\mathrm{FA}}} \approx 10^{-8}.$$

Next, let m be the number of file subjects in the test data and n be the number of search subjects in the test data. From (21), we can solve nm' from

$$(nm) = \left(z_{\alpha/2}/\varepsilon\right)^2 (\hat{p}_0(1-\hat{p}_0)).$$

Letting $\alpha = 0.05 \left(z_{\alpha/2} = 1.96\right)$, $\varepsilon = 10^{-8}$, and $\hat{p}_0 = 10^{-8}$, we found that $(nm) \approx 3.84 \times 10^8$, which is the product of the number of file subjects and the number of search subjects in the test data. At this point, one can determine a reasonable value for n and m', respectively.

We show one way of selecting the values of n and m'. Since we have previously determined that 1825 search subjects (each with a mate) are needed for a significant test of evaluating $\hat{P}_{\mathrm{MD}}$, it is reasonable to set n to be greater than 1825. Let n be 3650, which includes 1825 search subjects without mates. It follows that

$$m \approx 3.84 \times 10^8 / 3650 \approx 105\,206.$$

Hence, we would use about 105 000 file subjects.

The test data sample for evaluating whether a proposed system satisfies the requirements specified in the beginning of this section will consist of a search subject data set and a file subject data set. The search subject set will contain 3650 subjects; 1825 of the search subjects will have mates, and the remaining 1825 will have no mates. Each mated search subject will have exactly one mate in the file subject data set. The file subject set will contain at least 105 000 subjects, including 1825 mates of the search subjects.

B. The Collection of the Fingerprints for the Test

As discussed, we need to prepare two sets of fingerprints for the test: the search set and the file set. The fingerprints without mates are normally available through the existing AFIS operations. We can collect the nonmate fingerprints of search subjects and file subjects from the existing data base (fingerprint repositories). The mated fingerprints normally have to be collected through a special collection effort since we have to have multiple copies of fingerprints from each individual. If we collect two impressions of the right index finger from each individual, we label them s and f, respectively. We can insert the f impression into the data base and use the s impression as a search print.

Once we have determined the test data sample size, we would select a site for fingerprint collection. The chosen fingerprint collection site shall be representative of the typical operating environment of the AFIS under consideration. Some of the important considerations in selecting such a site include, but are not limited to, the population whose fingerprints would be captured, the cleanness of their fingers, the weather conditions, and the timing condition under which the operators are operating (capturing fingerprints).

In general, since the collected fingerprint quality will have a significant impact on the AFIS performance, we select fingerprint collection site(s) with characteristics representative of the data that will be collected for the AFIS. These particular sites must represent the environment of the possible fingerprint collection sites for the AFIS. If the AFIS operating environment differs significantly from site to site, it is important to have multiple fingerprint collection sites so that the collected fingerprints represent the various collection conditions.

Once a sufficient amount of fingerprints has been collected, a fingerprint examiner would need to verify the mated pairs of fingerprints to ensure that no mismatch had happened when collecting them. It shall be noted that collection errors made during the fingerprint collection process would only be discovered in the collected fingerprints verification process. A number of collected mated fingerprints would be excluded after the verification process if it was decided that they were not paired correctly (not true mates), which reduces the total number of usable fingerprints for the AFIS test. Therefore, we would collect more fingerprints than the numbers calculated in the previous section. In our experience, an extra 30% shall cover such losses. For this example, we hypothetically collected mated fingerprints from a group of about 2400 individuals and nonmate fingerprints from another group of about 2400 individuals.

After a fingerprint expert examines these fingerprints, we will randomly select 1825 qualified mated search prints and 1825 qualified nonmate search prints to form the search set. We will combine the 1825 mates of the mated search prints with about 105 000 nonmate file prints to form the file print set. The search print set and the file print set are collectively referred to as the test data set or sample.

C. Conducting the Test

The objective of an AFIS test is to determine the P_{MD} and P_{FA} we can expect in a "real world" operating environment. Therefore, we do our best to ensure that the AFIS test results reflect the product's "true" performance. The following list summarizes the key steps to achieve that objective.

1) Provide the vendor with a set of development data, which includes a set of search subjects' fingerprints, a set of file subjects' fingerprints, and a ground truth table. The ground truth table describes which file subject is the mate of each search subject. Also, provide the vendor a specification of what the resulting format should be.
2) The vendor can use the development data to tune its proposed AFIS. The purpose of this step is to make sure that the vendor can operate the AFIS on the test data and produce valid matching results.
3) Deliver the file subject set of the test data to the vendor. Allow sufficient time for the vendor to load this set into the AFIS since it may take a long time to extract features, check for duplicates, and load large numbers of fingerprints into a data base.
4) Deliver the search subject set to the vendor. Since the search subject set is considerably smaller than the file subject set, it is expected that it takes a proportionally smaller amount of time to perform feature extraction and prepare them for search.
5) Match the fingerprint of each search subject against the fingerprints of file subjects in the data base. Record the matching results, i.e., the search subject ID, if it is known to have a mate in the file subject set. This record can be referred to as the "answer sheet."
6) Compare the ground truth table with the vendor-provided answer sheet to determine what discrepancies there are between the two. Use the comparison results to calculate the initial estimates of P_{MD} and P_{FA}. In general, the ground truth table and the answer sheet can be stored as tables in a relational data base, and appropriate relational data base operations would produce the desired results.
7) If there are any discrepancies between the answer sheet and the ground truth, a human fingerprint expert will manually compare the fingerprints in question to determine whether the error was due to the AFIS operation or to error in the original data.
8) After the human fingerprint expert verifies each fingerprint in question, we calculate the final estimates of P_{MD} and P_{FA}. The final estimates are then compared to the requirements to determine if the underlying AFIS meets the requirements.

D. Analysis

In this example, assume that we have found that 35 mated search prints did not find their mates. They either found no mate or found the incorrect "mates." Furthermore, we assume that 188 search prints found incorrect matches. It follows that the $\hat{P}_{\mathrm{MD}} = 35/1825 \approx 0.0192$. These values are compared to the thresholds that can be determined from (37).

First, consider the threshold for $\hat{P}_{\mathrm{MD}}$, r. Using (37) and the system requirements, we have

$$0.05 = \sum_{y=r+1}^{1825} \frac{91.25^y e^{-91.25}}{y!}$$

where $n = 1825$ and $p_0 = 0.05$. It follows that $r = 75$ is the appropriate threshold. Since the number of mated search prints that did not find their mates is 35, and $35 < 75$, we accept the null hypothesis that $P_{\mathrm{MD}} \leq 0.05$.

Next, consider the threshold for $\hat{P}_{\mathrm{FA}}$, r. Using (37) and the system requirements, we have

$$0.1 = \sum_{y=r'+1}^{3650} \frac{365^y e^{-365}}{y!}$$

where $n = 3650$ and $p_0 = 0.1$. (Note that the value of p_0 here is the system requirement P_{FA}.) It follows that $r' = 389$ is the appropriate threshold. Since the number of search prints that found incorrect matches is 188, we accept the null hypothesis that $P_{\mathrm{FA}} \leq 0.1$. This leads to the conclusion that the underlying AFIS demonstrated "considerable" evidence (95% confidence level) that it meets the system requirements for $\mathrm{P_{MD}}$ and $\mathrm{P_{FA}}$.

VI. Conclusion

We have presented a tutorial on an automated biometrics-based identification and verification systems evaluation methodology based on fundamental statistics. It provides

a practical tool that engineers and users of automated biometrics-based identification and verification systems can use to evaluate various systems to help determine which ones meet the user-defined system requirements.

In summary, the following list includes the major steps involved in this evaluation methodology.

1) Determine the size and type of biometric data to be used in the evaluation. Also, select the environment in which the biometric data will be collected.
2) Collect the biometric data set and validate it. From the collected biometric data set, construct a development data set and a test data set.
3) Provide the development data set to the potential automated biometric system vendors.
4) Perform the matching test runs on the automated systems using the test data set and record the matching results. Manually verify the matching errors produced by the automated system.
5) Use the matching results to produce the parameter estimates. Perform hypotheses testing using the parameter estimates.
6) Analyze the hypotheses testing results and make the decision about whether a particular automated system meets the user-specified system requirements.

ACKNOWLEDGMENT

The authors wish to thank Dr. S. Barash for his insightful comments and suggestions, which greatly improved the clarity of this paper.

REFERENCES

[1] *Proc. 8th Biometric Consortium Meeting*, San Jose, CA, June 1996.
[2] *Proc. 9th Biometric Consortium Meeting*, Crystal City, VA, Apr. 1997.
[3] *Proc. BiometriCon' 97*, Arlington, VA, Mar. 1997.
[4] G. Casella and R. L. Berger, *Statistical Inference*. Belmont, CA: Wadsworth & Brooks/Cole, 1990.
[5] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions*, 2nd ed. New York: Wiley, 1992.
[6] V. K. Rohatgi, *An Introduction to Probability Theory and Mathematical Statistics.* New York: Wiley, 1976.

Weicheng Shen (Member, IEEE), for a photograph and biography, see this issue, p. 1346.

Marc Surette (Member, IEEE) was born on February 6, 1964, in Brighton, MA. He received the B.S. degree (*summa cum laude*) in computer systems engineering from the University of Massachusetts, Amherst, in 1986 and the M.S. and Ph.D. degrees in electrical engineering from the University of Colorado, Boulder, in 1989 and 1991, respectively.

From 1992 to 1995, he was a Researcher in the Optical Sciences Division of the Naval Research Laboratory (NRL) in Washington, D.C., where he developed and evaluated various hyperspectral image-processing algorithms. Prior to his image-processing work, he performed a combination of experimental research and computer simulation in the area of high-power semiconductor optical amplifiers. A publication related to this research was awarded NRL's "Alan Berman Research Publication Award" in 1993. He then was with Electronic Data Systems, where he was responsible for evaluating various image-processing-based biometric identification and verification systems. These included automated fingerprint identification, hand geometry, voice recognition, and facial recognition. Since September 1996, he has been with Silicon Graphics Inc., where he currently is a Systems Engineer in the Government Systems Area Technology Center. His current interests include scaleable high-performance computing, image processing, and scientific visualization.

Dr. Surette is a member of the Optical Society of America, Tau Beta Pi, and Eta Kappa Nu.

Rajiv Khanna (Member, IEEE), for a photograph and biography, see this issue, p. 1346.