Automatic Signature Verification Using a Three-Axis Force-Sensitive Pen

HEWITT D. CRANE, FELLOW, IEEE, AND JOHN S. OSTREM, MEMBER, IEEE

Abstract—**A performance analysis of an automatic signature-verification system based upon the use of a three-axis force-sensitive pen to transduce the dynamics of the handwritten signature into electrical signals suitable for real-time computer processing is presented. False-rejection and impostor-acceptance (i.e., type I / type II error) curves are presented for a variety of operating conditions. For typical "real world" conditions, the equal-error rate (where type I error rate equals type II error rate) is about one percent. The performance analysis was based upon a data base of 5220 true signatures obtained from 58 subjects over a four-month period, and 648 attempted forgeries obtained from 12 forgers. The forgers were given copies of the true signers' signatures, told how the verification system operates and what it measures, allowed to watch video tapes with close-up views of the signatures to be forged as they were being written, and allowed to practice for a three-week period. Particular advantages of the signature-verification system reported here are high performance, low storage requirements (typically 200 to 300 bits per user), and low-cost implementation in a stand-alone microprocessor unit.**

I. INTRODUCTION

WITH THE development of widely dispersed networks of computer terminals and data banks, there has been a corresponding increase in computer crime and a growing need to protect sensitive information assets [1]. An important aspect of the problem is personal identification —that is, the ability to ensure that only authorized people have access to computer resources. Applications for personal identification range from high-security access control, to banking transactions using automated teller units, electronic funds transfer, and so forth.

The basic requirements for a personal verification system are that it be automatic (i.e., no human intervention in the decisionmaking process), real time, reliable, cost effective, and acceptable to potential users. Automatic personal verification systems based upon fingerprints, passwords, hand geometry, palmprints, voice, and the handwritten signature have been developed. These methods, among others, all vary in terms of complexity, cost, and performance. Signature verification has the important advantage that signatures have long been the primary form of legal attestation, and hence user acceptance is not the problem that it is with, for example, fingerprints, which carry the connotation of criminality. Other important advantages include low cost, low computational requirements, and high performance.

Automatic signature verification [2]–[5] requires a representation of the handwritten signature that is suitable for computer processing. There are basically two ways to obtain such a representation. One way is to scan the signature optically after it has been written; this technique is similar in principle to that used for optical character recognition (OCR). However, optical scanning devices are often bulky, expensive, and generally unsuitable for real-time applications of signature verification. A more attractive and useful approach is to have the writing device itself (i.e., the pen or the writing surface) generate electric signals representative of the signature during the writing process.

As discussed in Section II, our system uses a strain-gauge instrumented pen to transduce the forces and motions used to write the signature into electrical signals suitable for computer processing.[1] A major advantage of this approach is that the signature-verification process is based upon matching the dynamics of the signature rather than the final static image. Our experience indicates that this makes forgery much more difficult because dynamic information about a signature is more difficult for the forger to obtain and use. In this regard, a static image of a signature on a credit card or document is almost useless to the forger for counterfeit purposes; tracing a signature is one of the worst strategies for a forger because the dynamics of the forged signature are usually very different from that of the true signature.

In this paper, we focus on a signature-verification algorithm that is based upon extracting features from the signals generated by the instrumented pen. The advantages of this method are computational efficiency, minimal reference or template storage per user (typically 200 to 300 bits), and that it can easily be accommodated in a low-cost, stand-alone microprocessor system. More sophisticated signature-verification algorithms, based upon correlation techniques (which will not be discussed here), yield higher performance but with a corresponding increase in complexity, computational effort, and template storage requirements.

In Section II we discuss the essential features of the three-axis, force-sensitive pen used in the signature-verifi-

Manuscript received February 23, 1982; revised January 6, 1983. This work was supported by AFSC Rome Air Development Center under Contract F30602-79-C-0255.

H. D. Crane is with the Sensory Sciences Research Center, SRI International (formerly Stanford Research Institute), Menlo Park, CA 94025.

J. S. Ostrem was with SRI International. He is now with Communication Intelligence Corporation, Menlo Park, CA 94025.

[1]An alternative that has been considered and partially reduced to practice is based on a three-axis tablet instead of the three-axis pen described here.

0018-9472/83/0500-329$01.00

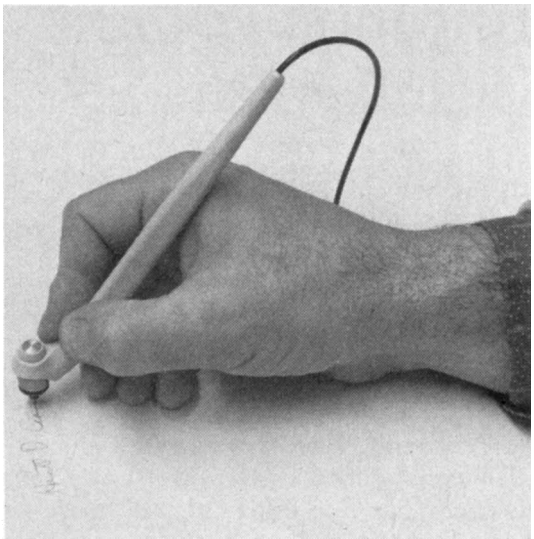

Fig. 1. Three-axis, force-sensitive pen.

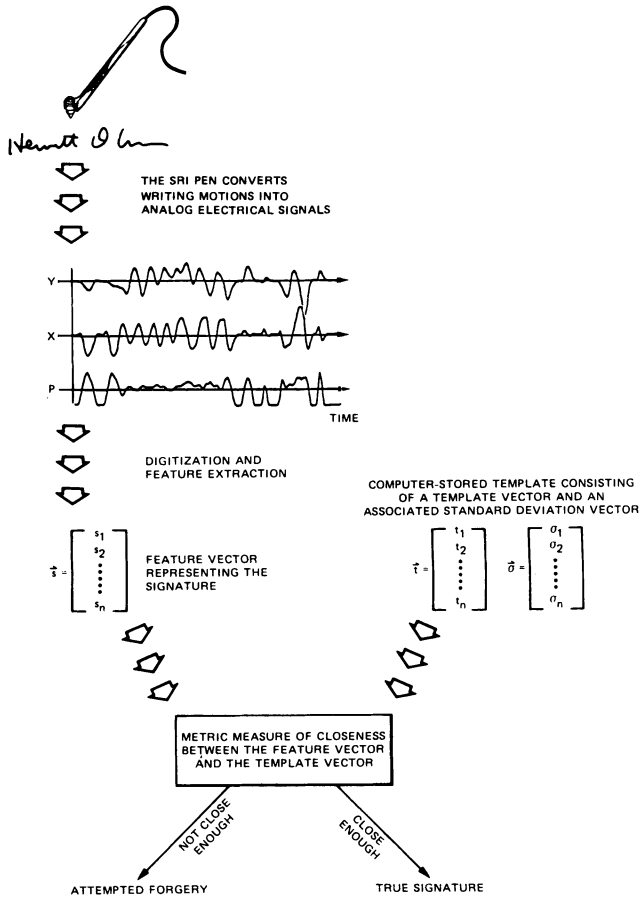

Fig. 2. Summary of the features approach to signature verification.

cation system. Section III summarizes the data base of true signatures and attempted forgeries that was collected for the purpose of analyzing system performance. In Section IV we present a general description of the signature-verification procedures. The specifics of feature extraction are discussed in Sections V and VII. The basic performance criteria for personal verification are the type I/type II error curves which are the subject of Section VI. In Section VIII we present the results of the performance analysis, including type I/type II error curves for a variety of operating conditions. Finally, Section IX summarizes the most important results of the performance analysis.

II. Three-Axis Force-Sensitive Pen

SRI International has developed a strain-gauge instrumented ballpoint pen, shown in Fig. 1, that generates three electrical signals that are representative of the instantaneous three-dimensional drag force at the writing tip. The pen is a simple and reliable device that writes like an ordinary ballpoint pen and requires no special writing surface or paper. The mechanical construction and associated electronics of the pen are described in [6].

When a signature is written, the pen generates a set of three analog signals that are a time history of the instantaneous three-dimensional force on the pen tip. A typical example of the signals generated for a signature is given at the top of Fig. 2. When the pen tip is vertical, the P-signal represents the downward force, and the X and Y force signals represent forces orthogonal in the plane of the paper. Note particularly that the pen signals are very clean and essentially noise free.

III. Data-Base Summary

The data base consists of 5220 signatures obtained from 58 subjects signing their own signatures over a four-month period, which we call the true-signer data, and 648 attempted forgeries obtained from 12 trained forgers. Each of the true-signer subjects attended one or two data-collection sessions per week during a four-month period. At each data-collection session, the subject signed three signatures while sitting down at a table and three signatures while standing at a counter. On the average, approximately 50 sitting signatures and 50 standing signatures were obtained from each subject. The purpose of collecting both sitting and standing data was to determine whether there was any significant difference in performance in the two configurations (i.e., whether the true signers wrote their signatures more consistently sitting down or standing up), as this would have important implications for how a practical signature-verification system should be designed.

The true-signer subjects were chosen at random from a large group of volunteers at SRI International. The only constraints imposed on subject selection were to obtain an approximately equal number of women and men, about 10 percent left-handers (based upon estimates of the percentage of left-handers in the general population), and a representative range of heights, weights, and ages. Thirty of the 58 subjects were women.

The 12 attempted-forgery subjects (henceforth called forgers) were given copies of the true-signers' signatures, instructed in how the signature-verification system works and what it measures, allowed to watch video tapes showing a close-up view of the true signers writing their signature, and allowed to practice as much as desired over a three week period. To provide motivation, cash prizes ($100, $50, and $25) were awarded for the best forgeries.

A more detailed description of the data-collection procedures is given in the Appendix.

IV. Signature Verification

When a person is to have his identity verified, he must identify himself to the system (e.g. by means of an identification number) and then write his signature. As shown in Fig. 2, the pen transduces the forces and motions used in writing the signature into a set of three analog signals that are a time record of the instantaneous force on the tip of the pen in three orthogonal directions: P, X, and Y. These analog signals are digitized at the rate of approximately 100 samples per second per channel and processed by a computer to extract a set of descriptors, called features, from the three signals. These features include various timing parameters such as the total time of the signature, the average force in each of the three directions and the corresponding energies, the number of pen-ups and pen-downs, and so on.

The set of features extracted from the discrete representations of the P, X, and Y signals forms a feature vector, as shown in Fig. 2. The computer then calls up the computer-stored template or reference feature vector corresponding to the person whom the writer claims to be. The template vector is constructed by averaging a set of known true signatures [2]. Associated with the template vector is a vector of standard deviations for the features (see Fig. 2), which provides a measure of how variable the true signer is from signature to signature for each feature. A measure of closeness (as discussed later, a weighted Euclidean distance metric) between the feature vector and the template vector is then computed. If the feature vector corresponding to the signature in question is "close enough" to the template vector (i.e, the distance between them is less than some preselected threshold), the signature is judged to be a true signature and the person's identity is verified. If the feature vector is not close enough, the signature is judged to be an attempted forgery. In a practical signature-verification system, the writer usually is given more than one chance to be verified; that is, if the first signature does not pass the above test, he or she will be allowed to write one or two more signatures to be tested for verification.

If this features approach to signature verification is to be successful, the feature vector must contain as much information as possible that is useful for discriminating between true signatures and forgeries. The basic purpose for collecting a data base of true signatures and attempted forgeries is to provide data that can be analyzed to select a set of features that provides maximum discriminating power between the true signatures and the attempted forgeries. The process of optimizing the feature's technique consists of selecting the "best" set of features and an appropriate threshold for the distance metric measure of closeness.

The problem of selecting a "best" set of features has two aspects, which we call feature extraction and feature selection. In general, there is no way to make an *a priori* determination of what the best features will be for a particular situation so the usual approach is to extract a relatively large number of features that are expected to have useful discriminating power. However this generally results in substantial redundancy. The objective of feature selection is to obtain a reduced set of features that contains essentially all the discriminating power of the original feature set [7].

TABLE I
The 44 Original Features

Feature Number			Feature
X	Y	P	
1	11	21	SCALED MEAN
2	12	22	STANDARD DEVIATION
3	13	23	MINIMUM
4	14	24	MAXIMUM
5	15	25	AVERAGE ABSOLUTE
6	16	26	AVERAGE POSITIVE
7	17	27	NUMBER OF POSITIVE SAMPLES
8	18	28	AVERAGE NEGATIVE
9	19	29	NUMBER OF NEGATIVE SAMPLES
10	20	30	NUMBER OF O-CROSSINGS
31	32	33	MAXIMUM MINUS SCALED MEAN
34	35	36	MAXIMUM MINUS MINIMUM
37	38	39	SCALED MEAN MINUS MINIMUM
40			TOTAL TIME
41			NUMBER OF SEGMENTS -1
42			TIME UP
43			NUMBER OF SEGMENTS
44			TIME DOWN

V. Feature Extraction and Selection

Based on our knowledge of the characteristics of the P, X, and Y signals derived from the pen, and our experience with previous true-signature and forgery data bases, a set of 44 features was selected as the starting point in the current data-base analysis. These features are summarized in Table I.

The objective of feature selection is the following: given this relatively large set of 44 features, find the subset of features that yields the lowest type I/type II error rates [7]. Feature selection also improves computational efficiency by excluding or combining the features that contain redundant information.

In general, most feature-selection techniques follow the same basic procedure. The starting point is a large set of features that the analyst believes to be useful for discriminating between samples (in our case, between true signatures and attempted forgeries). The discriminating power of each of the features, or combinations of features, is determined by performing statistical tests on a training set of data that is believed to represent the population of interest adequately. The combination of features that yields the best performance (by some criteria) and that contains the minimum number of features is the "best" feature set.

Many procedures and algorithms for performing computerized feature selection have been devised. Some of these are based on univariate F-ratio evaluations [8], Fisher's discriminate analysis [7], information measures such as divergence [9], and a host of ad hoc procedures. For the current work we tried a number of these techniques. Although some of them performed reasonably well, we were not entirely satisfied with the results. The standard feature-selection techniques are all based on a number of assumptions about the underlying probability structure of the feature set. The exact assumptions differ somewhat from technique to technique, but in general it is assumed that the set of features is distributed as a multivariate Gaussian density, that the covariance matrices are equal, and the like. Our signature-verification features do not appear to satisfy these conditions, and the result is that the feature-selection techniques mentioned above do not operate in an optimum fashion; that is, there is no guarantee that the feature set obtained is the one that minimizes the type I/type II error rate. Because of the somewhat unsatisfactory performance of these classical feature-selection techniques, we devised an approach that uses as its basic criterion the direct minimization of the type I/type II error rate.[2] This approach, which resulted in improved feature selection and lower type I/type II error rates, can be summarized as follows. Assume that the original feature set contains N features: 1) calculate type I/type II error curves[3] for all $N - 1$ subsets of the N features and determine which subset yields the least error rate. Typically the criterion is to minimize the equal error rate (where type I error equals type II error), but this can be modified depending upon the application; 2) calculate the type I/type II error curves for all $N - 2$ subsets of the best $N - 1$ subset; 3) calculate the type I/type II error curves for all $N - 3$ subsets of the best $N - 2$ subset, and so on [10]. What typically occurs in this process is illustrated in Fig. 3. As useless and/or redundant features are removed, the error rate decreases until it reaches a minimum. Once this minimum is reached, the exclusion of more features results in reduced performance. The feature set that yields the minimum error rate is selected as the best feature set. This procedure is an approximation to the more complex process of calculating the type I/type II error curves for all possible subsets of the N features, which is computationally prohibitive—a set of only 20 features would require more than one million type I/type II calculations.

Compared to classical techniques for feature selection, this method has the following advantages.

- It requires no assumptions about the underlying probability distribution of the feature set.
- The calculations involved are relatively simple and intuitively reasonable.
- It selects a "best" feature set by choosing the subset that yields the least probability of error (subject to the qualification mentioned above that not all possible combinations of feature subsets are tested by our restricted search algorithm). This is not true of classical feature-selection procedures in general whose results can be said to minimize the probability of error only under a very restrictive set of assumptions, which experience has shown is not valid for the signature-verification features.
- Redundant features are excluded by the process of choosing the minimum point of the curve in Fig. 3.

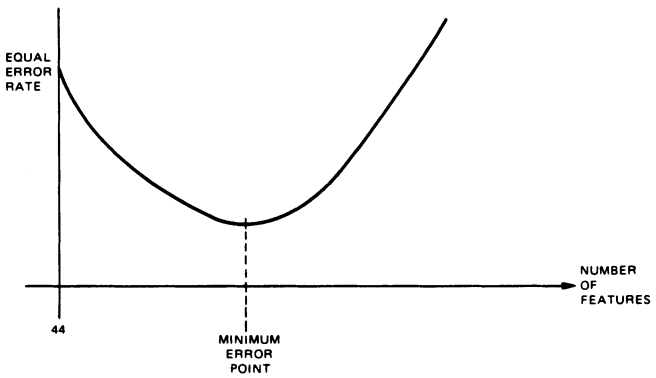

Fig. 3. Type I/type II equal-error rate versus number of features.

VI. Type I/Type II Error Curves

A type I error occurs when a true signature is rejected as an attempted forgery, and a type II error occurs when an attempted forgery is classified as a true signature. The type I error rate is often called the false-rejection rate and the type II error rate, the impostor-acceptance rate. To estimate the type I/type II error curves, we developed an analysis procedure that simulates how a real-world signature-verification system might operate. This program includes an enrollment phase in which templates are constructed from the first few (typically 10 or 12) signatures for each subject, and a verification phase in which subsequent true signatures and attempted forgeries are compared against the appropriate templates to determine the percentage of false rejections and impostor acceptances. The system allows up to three tries per verification trial. If the first signature fails to pass the verification criteria, a second signature is tested. If the second one also fails, a third signature is considered. If all three signatures for a particular verification trial fail, the subject is rejected as an impostor.

An adaptive template-updating procedure was used to continually modify templates based on successful verification attempts. Each time a verification trial was successful on the first try (i.e., the first signature satisfied the verification criteria), the feature vector for that signature was added to the template vector with a weighting of 1/8. Thus if a subject's signature varied slowly over time, the template automatically tracked the change. This template-up-

[2] Generally the classical feature-selection techniques cannot be related directly to type I/type II error rates except, as noted earlier, under a set of restrictive assumptions about the probability structure of the feature set (which are not satisfied by the signature-verification features).

[3] The procedure for calculating these curves is described in Section VI.

dating procedure was found to improve verification by reducing the percentage of true-signer rejections.

The basic criteria used to judge whether a particular test signature was a true signature or an attempted forgery is as follows. As in Fig. 2, let $\vec{s}$ be the feature vector representing the test signature. The components of $\vec{s}$ are the values of the set of "best" features determined by the method described earlier. Let $\vec{t}$ be the computer-stored template or reference vector and $\vec{\sigma}$ the associated standard deviation vector. The determinations of $\vec{t}$ or $\vec{\sigma}$ are based on an enrollment set of known true signatures (see [2] for explicit formulas for calculating $\vec{t}$ or $\vec{\sigma}$). Referring to Fig. 2, the measure of closeness between the test signature and the template is the weighted Euclidean distance metric

$$d(\vec{s}, \vec{t}) = \sqrt{\frac{1}{f} \sum_{i=1}^{f} \left(\frac{s_i - t_i}{\sigma_i} \right)^2},$$

where f is the number of features, s_i is the value of the ith component or feature in the feature vector $\vec{s}$, t_i is the ith component of the template vector, and σ_i is the standard deviation of the ith feature as computed from a set of enrollment signatures. (See [2] for a discussion of the reasons for selecting this Euclidean distance metric as the measure of closeness between the test signature and the template.) The smaller the calculated value of $d(\vec{s}, \vec{t})$, the greater the similarity between $\vec{s}$ and $\vec{t}$, and therefore between the test signature represented by $\vec{s}$ and the computer-stored template $\vec{t}$.

The decision rule for deciding whether a particular test signature satisfies the verification criteria is as follows.

- If $d(\vec{s}, \vec{t}) \leq d^{\text{thres}}$, the signature is judged to be a true signature.
- If $d(\vec{s}, \vec{t}) > d^{\text{thres}}$, the signature is judged to be an attempted forgery.

The quantity d^{thres} is a preassigned threshold value selected by using the type I/type II error curves to obtain the optimum trade-off between type I and type II errors for the particular application of interest. For example, for high-security applications, d^{thres} would likely be set to a relatively small value to minimize the impostor-acceptance rate, while for banking applications, where the concern is usually to minimize user inconvenience (i.e., minimize the type I error rate), a larger value for d^{thres} might be more suitable.

The procedure by which the type I/type II error curves are estimated from the data base is as follows. Let T^t represent the total number of verification trials in the true-signer data base and R^t the number of trials for which a true signer was falsely rejected. Note that R^t is a function of the decision threshold while T^t is not. In general, R^t decreases as d^{thres} increases, and increases as d^{thres} decreases. Recall that each verification trial allows up to three attempts, so a false rejection occurs only when all three signatures fail to satisfy the verification criteria. The type I error (false rejection rate for true signers) is estimated as

$$\text{type I error} = \hat{E}_{\text{I}} = \frac{R^t}{T^t}.$$

The caret is used to indicate that $\hat{E}_{\text{I}}$ is an estimate of the error rate. When $\hat{E}_{\text{I}}$ is plotted as a function of the decision threshold, d^{thres}, the type I error curve results. Similarly, the type II error is estimated to be

$$\text{type II error} = \hat{E}_{\text{II}} = \frac{R^f}{T^f},$$

where T^f is the total number of forger trials and R^f is the number of trials for which a forged signature passes the verification criteria (i.e., the number of impostor acceptances). R^f is also a function of the decision threshold, increasing with increasing d^{thres} and decreasing with decreasing d^{thres}. A plot of $\hat{E}_{\text{II}}$ versus d^{thres} yields the type II error curve. The justification for using the particular form of error-rate estimation given above is that $\hat{E}_{\text{I}}$ and $\hat{E}_{\text{II}}$ are the maximum-likelihood estimates (assuming independent trials) of the error rates for binomially distributed random variables.

VII. Personalized Feature Selection

It is well-known, both theoretically and from practical experience, that the use of personalized feature sets generally yields better signature-verification performance (lower type I/type II error rates) than if the same set is used for all subjects, provided that enough training data are available to estimate the personalized sets with reasonable statistical confidence.[4] However, the use of personalized feature sets requires a more complex enrollment procedure, and it is not clear *a priori* that the improved performance is sufficient to justify its use for many practical applications.

If a standard feature set is used (i.e., if the same feature set is applied to all subjects), approximately 10 to 12 signatures are adequate for enrolling a subject. This seems quite practical and reasonable for a real-world signature-verification system. The requirements are quite different for personalized feature selection. Although the exact number of signatures needed cannot be determined without knowing the exact probability structure of the signature-verification features (although they are definitely non-Gaussian), a standard rule of thumb in such instances is that the number of independent training (enrollment) samples must be several times the number of features. Since we begin with 44 features, this implies that the number of enrollment signatures should be quite large, possibly greater than 100, although it might be possible, with less confidence, to make do with 40 or so. In any case, this means that personalized feature selection may require a relatively long enrollment procedure. However, a compromise may

[4] This is intuitively reasonable. Since all subjects write differently, we would expect their signatures to be best characterized by somewhat different feature sets. For example, the total time that it takes to write the signature is a good feature for subjects who are consistent in the timing of their signatures but a bad feature for those who are very inconsistent in total writing time.

be possible in which subject enrollment is based on a standard feature set, and as more signatures are collected through subsequent verifications, the feature set is gradually and automatically personalized. But this approach has its own disadvantage, that of requiring the system to store, at least temporarily, the large number of feature vectors required for the feature-selection process.

VIII. Performance Evaluation

In this section, we present type I/type II error curves based upon a standard set of "best" features derived from the original set of 44 features, and then give an example of the improvement in performance that results by using personalized feature selection.

A. Standard Feature Set

Type I/type II error curves are shown in this section, beginning with the so-called "trues versus trues" case. These curves are computed as follows. The known true signatures of a particular subject, say subject *ABC*, are compared against his own template. The percent of rejection as a function of the decision threshold is the type I error curve for *ABC*. The type II error curve for *ABC* is computed by comparing the true signatures of all the other true-signer subjects against the *ABC* template. This procedure is repeated for all subjects in the data base and the results are combined to obtain the overall type I/type II error curves.

Clearly the trues versus trues error rate is a kind of confusion rate, comparable to the situation in which one subject claims the identity of another subject but attempts to use his own signature for verification. This does not provide a very realistic measure of the system's true performance, and is included here only because this type of error-rate calculation is very common in the literature. Following the presentation of trues versus trues type I/type II error curves, we present the trues versus attempted forgeries type I/type II error curves. In this case the type I error curves are calculated in the same way as the above, but the type II error curves are computed using attempted-forgery data.

1) Trues Versus Trues: To select a "best" subset of the 44 features, we began by dividing the signature data into two sets, a training set and a testing set. Because we collected an equal number of sitting and standing signatures, a natural division was made on this basis. To begin with, we used the sitting signature data as the training set on which feature selection was performed in order to determine a best subset of the 44 original features (i.e., the minimum subset that yields the least error rate). The best subset consisted of features 1, 2, 3, 6, 11–14, 16, 20, 22, 25–30, 32, 33, 38, and 40–44. The standing data were then used to calculate the type I/type II error curves, the results of which are shown in Fig. 4. To compare results we will use the point at which the type I/type II errors are equal (i.e., the percent error where the curves intersect), which is called the equal-error rate. This equal-error rate, indicated

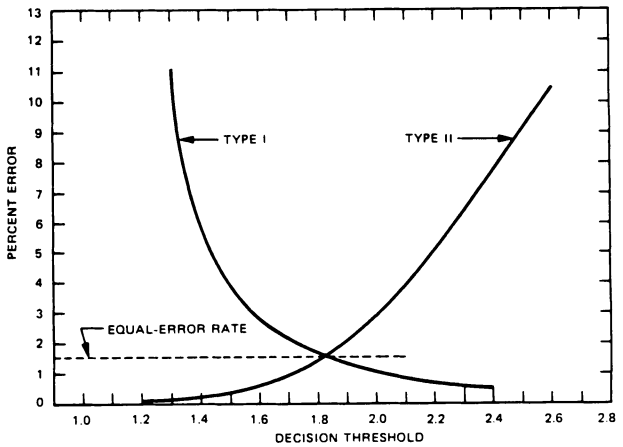

Fig. 4. Type I/type II error curves for trues versus trues, standing data.

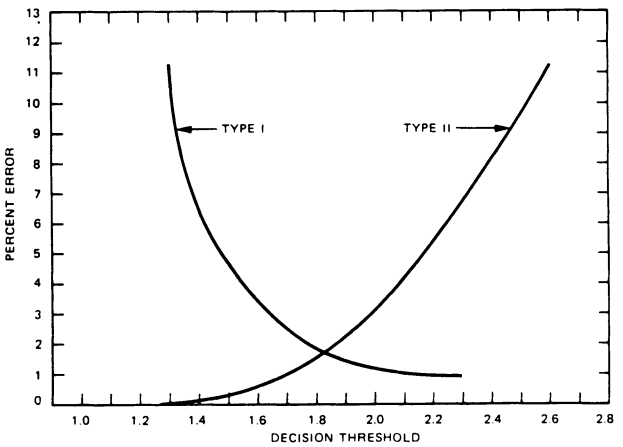

Fig. 5. Type I/type II error curves for trues versus trues, sitting data.

by the broken horizontal line in Fig. 4, is about 1.5 percent for the standing data. To cross validate the results, we reversed the roles of the sitting and standing data; the standing data were used for feature selection and the sitting data for error-rate calculation. The feature set selected using the standing data was identical to that derived using the sitting data. The type I/type II error curves for the sitting data are shown in Fig. 5. Comparison of Figs. 4 and 5 shows that the error curves are essentially identical so the cross validation yielded very consistent results, which gives us added confidence. It may also be concluded that there is essentially no difference in performance whether the subject is sitting or standing while writing his signature.

2) Trues Versus Attempted Forgeries: For the trues versus attempted forgery[5] type I/type II error-curve calculations, we decided to use the same set of best features that had been used for the trues versus trues calculations. The reason for this is that the generality of the forgery data is uncertain because very little is known about the forger

[5]Each forger was allowed up to 18 tries to forge a signature, nine before and nine after viewing a video tape showing several close-up views of the true signer signing his own signature. Because we found that there is only a slight difference in the error rates for the two conditions, the type II error curves presented in this section are calculated using the combined set of forgery attempts.

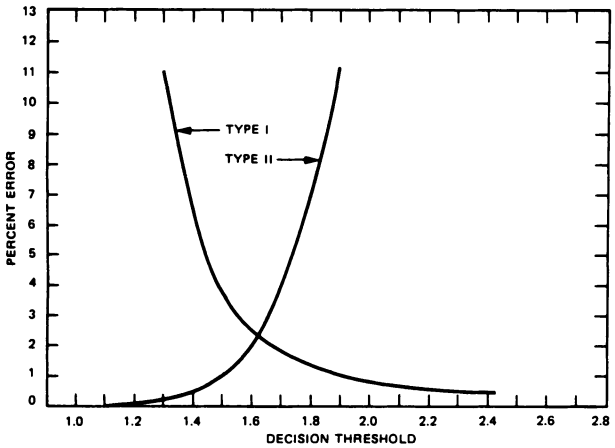

Fig. 6. Type I/type II error curves for trues versus attempted forgeries, standing data, no enrollment criteria.

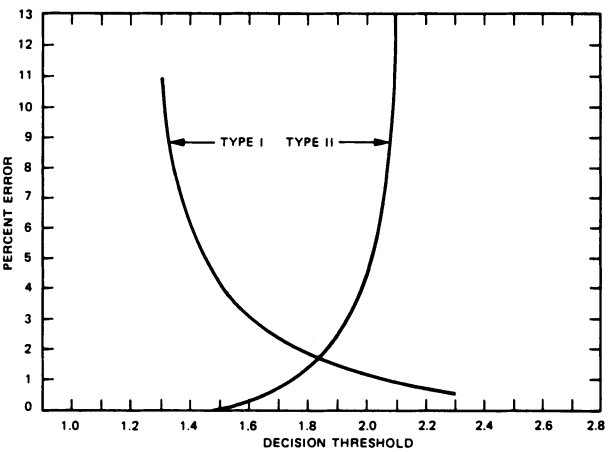

Fig. 8. Type I/type II error curves for trues versus attempted forgeries, standing data, with enrollment criteria.

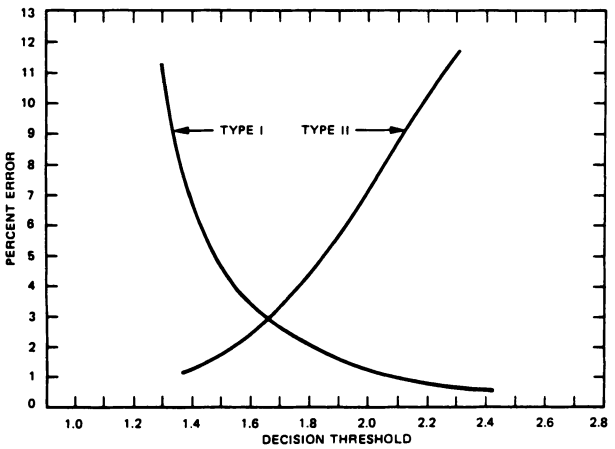

Fig. 7. Type I/type II error curves for trues versus attempted forgeries, sitting data, no enrollment criteria.

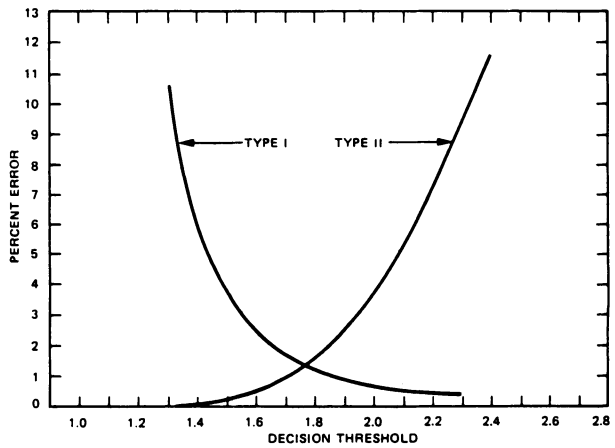

Fig. 9. Type I/type II error curves for trues versus attempted forgeries, sitting data, with enrollment criteria.

population. In any case, the use of the trues versus trues feature set is a conservative approach, and if anything, should lead to error-rate estimates that are too high.

The type I/type II error curves for the standing true-signature data versus the attempted-forgery data are shown in Fig. 6. The equal-error rate is approximately 2.25 percent which is somewhat worse than the 1.5 percent equal-error rate of the trues versus trues data. The type I/type II error curves for the sitting trues versus attempted forgeries are shown in Fig. 7. The equal-error rate is almost three percent.

Data analysis showed that almost all the forgeries occurred for the two or three true signers who were the most inconsistent in writing their signatures.[6] A simple enrollment criterion based on the total variance of the template was subsequently tested. If the combined standard deviation was larger than some assigned threshold, the subject failed the enrollment criterion and was excluded. This resulted in the exclusion of three subjects out of 58 and yielded considerable improvement in signature-verification performance. Figs. 8 and 9 show the type I/type II error curves (for standing and sitting data, respectively) when this enrollment criterion is used. The equal-error rates are reduced to 1.75 percent for the trues versus forgeries (standing) and to 1.25 for the trues versus forgeries (sitting). By making the enrollment criteria even more stringent, to where six or seven of the most inconsistent subjects (out of 58) were excluded, we found equal-error rates on the order of 0.5 to 0.75 percent.

Tests were also made to measure the effect of allowing forgers to view video tapes of the true signers writing their signatures. The error rate was only slightly worse when the forgers were allowed to view the video tapes, which implies that the forger can learn something of the signature dynamics by closely observing the true signer, although the effect is not very significant.

B. Personalized Feature Sets

Here we present an example of the kind of improvement that may be expected if personalized feature sets are used. Fig. 10 shows the type I/type II error curves (the solid lines) for the worst subject in the data base, the subject for which the type I/type II equal-error rate was the greatest (over six percent).

[6]This behavior is typical of verification systems. Generally most of the errors are contributed by a very small percentage of system users—the "goats" of the system.

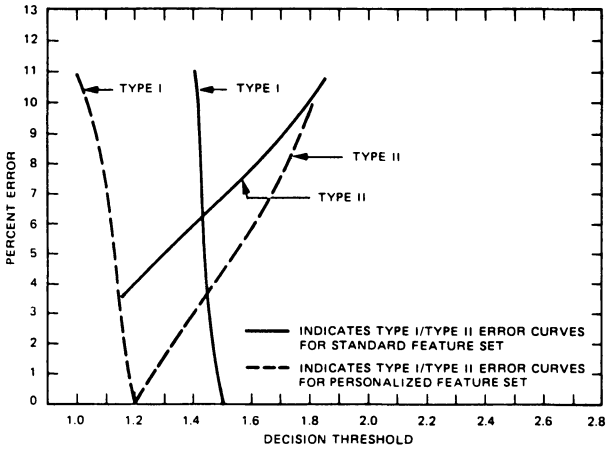

Fig. 10. Type I/type II error curves for a standard feature set and for a personalized feature set.

A personalized feature set for this subject, which was derived using the method described in Section V, consisted of features 1, 2, 6, 11, 13, 16, 26, 27, 32, 38, 40, and 44. The type I/type II error curves using this personalized feature set are given by the broken-line curves in Fig. 10. Note the substantial improvement compared to the type I/type II error curves for the standard feature set. In fact, for the personalized feature set there is no crossover at all of the type I/type II error curves, and so the equal-error rate is zero. However, this is based on only one subject's true signatures and the associated attempted forgeries, and it would not be appropriate without extensive further testing to conclude that personalized feature selection would yield a type I/type II error rate of zero. However, because of this result and our previous experience, we believe that personalized feature selection should reduce the equal-error rate by at least a factor of 2 (i.e., an equal-error rate on the order of 0.5 percent compared to the 1 percent reported in the preceding subsection).

IX. Summary

The estimated performance of the signature-verification system described in this paper is shown in the type I/type II (false-rejection/impostor-acceptance) error curves of Fig. 9. The equal-error rate (i.e., where type I error equals type II error) is slightly above one percent. Although up to three tries (signatures) were allowed for each verification attempt, the average number of signatures required per verification trial was actually only 1.1. The first 10 to 12 signatures for each subject were used to construct the template, or reference, except in those cases where the subject was so variable that he could not meet the enrollment criteria. Of the 58 subjects, three were rejected on the basis of the enrollment criteria. If the enrollment criteria are not used, the error rates are slightly worse (see Figs. 6 and 7). The type II (impostor-acceptance) error rates were estimated using attempted-forgery data obtained from forgery subjects. These forgers were given copies of the true-signers' signatures, instructed in how the signature-verification system operates and what it measures, allowed to watch video tapes showing a close-up view of the signatures being written by the true signers, and allowed to practice over a three week period. Essentially these subjects were given all the information that a hard-working "real-world" forger could be expected to obtain.

Signature verification based on the features technique, but with personalized feature selection, was also considered. This algorithm was tested using one of the problem subjects (the few subjects in the data base who caused essentially all the errors) and resulted in substantially improved performance compared to the features technique based on a standard feature set for all subjects (on which the results of Fig. 9 are based). Although we were unable, because of resource limitations, to process enough data to obtain a statistically confident estimate of the type I/type II error curves for personalized feature selection, the results of our limited testing with problem subjects, and our previous experience, have led us to believe that the equal-error rate is probably at least a factor of two better than for the features technique using a standard feature set for all subjects. The primary disadvantage of personalized feature selection is that it may require a relatively large number of enrollment signatures, which may not be justified for all applications.

Appendix

Data-Collection Procedures

1) Data-Collection Area: The data were collected in a partially enclosed area containing a table and a counter (a podium-like stand). At each data-collection session the subject wrote signatures both while sitting down at the table and while standing at the counter. The operator sat in front of a computer terminal immediately adjacent to the partially enclosed area. Although the area was partially enclosed, the subjects were not totally isolated from view nor acoustically shielded from the normal computer noise. In essence, the data-collection environment was essentially what might be expected for a personal identification system used for access control to a computer area.

2) True-Signer Data Base: Upon entering the data-collection area, the subject was given a standard form on which to write his signatures. For purposes of documentation, the form was labelled with the subject's name, the date, and other pertinent information. The operator then informed the subject whether the data collection was to begin at the table or the counter. To prevent bias, we alternated the order of collection between sessions. Each data-collection session consisted of six signatures, three obtained while the subject was sitting at the table and three while standing at the counter. Normally each subject attended one or two data-collection sessions per week over a four-month period, although vacations and trips sometimes interfered with the schedule. It was necessary to obtain three signatures for both sitting and standing conditions during each data-collection session because we wanted the flexibility in the performance analysis to simulate a system in which up to three tries at verification were allowed.

During the first session, the subject was given brief instructions. He was told that the system measures forces

and dynamics, and that any unusual pauses in writing were likely to cause the signatures to be rejected. The subject was instructed to use his or her standard signature. A subject who typically used one or more signature variants (e.g. a full middle name one time and only an initial the next) was requested to use the most common version of the signature. The subject was instructed to inform the research assistant of any obvious mistakes such as leaving out a middle name or initial, or other gross signature variants. There were very few such mistakes, and those that occurred were excluded from the data base.

The signature data for each session were collected on a real-time, on-line basis. Each signature was automatically digitized and written out on disk, including a header record consisting of the subject's initials, the date and time, a response or index number, and other pertinent information. All data forms for all subjects and for all data-collection sessions, as well as hard-copy records of all program transactions, were saved. In all, sufficient records were maintained so that whenever questions arose about the data it was possible to reconstruct exactly what happened during the session in question.

For a signature-verification system operating in the "real world," the users must cooperate with the system or risk being denied access to a secure area, computer account, or the like. However, no such motivation exists for a data-collection effort of the type described here. Thus there is always the danger that subjects will grow careless after the initial novelty wears off, which can lead to unnaturally large variations in the way signatures are written and cause an artificially low estimate of system performance (compared to a real-world system in which users are continually motivated by the need for access). Hence, to better simulate real-world operating conditions, prizes were offered ($100, $50, and $25) for the signatures that were most consistent over the data-collection period. The intent here was to provide at least some motivation for the subjects to perform as they would in a real-world environment.

3) Forger Data Base: The procedure for collecting and storing attempted-forgery data is basically the same as that for the true-signer data collection described in the preceding subsection. This is to be expected because in the real world there is no *a priori* knowledge as to who is the true signer and who is the forger, so both must be treated the same (up to the point of verification). To provide motivation, prizes ($100, $50, and $25) were also offered for the "best" forgeries.

The one difference was the problem of selecting subjects for the attempted-forgery data base. This was difficult because the SRI signature-verification system is based on the dynamics of a signature (i.e., the forces and motions used to create a signature) rather than its final static image. Thus the requirements for being a successful forger are quite different than those for a "classical" forger whose purpose is to duplicate the static image of a signature, and, since dynamic signature verification is relatively new, there is no known population of forgers. For example, tracing a true signature is one of the worst strategies for forgery in a dynamic system because tracing usually results in dynamics that are very different from those of the true signer even though the final static signature image may appear essentially identical. Our approach, therefore, was to select motivated people who had good manual dexterity and the capability of understanding the basic concepts behind the verification system.

Rather than requiring the forgers to make a few attempts at all the different signatures in the true-signer data base, we decided that a more realistic simulation of how a real forger would operate would be to have each forger concentrate on three or four different signatures. They were given several samples of these signatures and a description of how the signature-verification system operates: that it measures signature dynamics, that timing and forces are generally important, and that some of the typical features on which the verification is based are the total time of the signature, average force in the three orthogonal directions and the respective energies, the number of pen-ups and pen-downs, and so on. Each forger was allowed 18 attempts to forge a particular signature. After the first nine attempts he was shown a video tape with a close-up view of the subject signing his signature. This was intended to simulate the condition in which a real forger surreptitiously observes a person writing his signature to learn as much as possible about the dynamics of the signature. Before the actual forgery attempts, the forgers were allowed unlimited practice within a three-week period. In essence, the forgers were provided with all the information that a dedicated real-world forger could be expected to obtain.

References

[1] D. B. Parker, *Crime by Computer*. New York: Scribner's, 1976.

[2] H. D. Crane, D. E. Wolf, and J. S. Ostrem, "The SRI pen system for automatic signature verification," in *Symp. Proc. NBS Trends and Applicat. 1977*, May 1977, pp. 32–39.

[3] J. Sternberg, "Automatic signature verification using handwriting pressure," *1975 Wescon Tech. Papers*, paper 31/4, Sept. 1975.

[4] N. M. Herbst and C. N. Lin, "Automatic verification of signatures by means of acceleration patterns," in *Proc. IEEE Comp. Soc. Conf. Pattern Recognition and Image Processing*, June 1977, pp. 331–336.

[5] C. N. Lin, N. M. Herbst, and N. J. Anthony, "Automatic signature verification: System description and field test results," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 35–38, Jan. 1979.

[6] H. D. Crane and R. E. Savoie, "An on-line data entry system for hand-printed characters," *Comput. Mag.*, vol. 10, no. 3, pp. 43–50, Mar. 1977.

[7] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973, p. 66. Under certain assumptions concerning the probability distributions of the feature set, it can be shown that the process of feature selection cannot reduce the type I/type II error rates. However, as a practical matter, an improvement in error-rate performance often results from feature selection.

[8] The *F*-ratio technique for feature selection is described in many textbooks. For example, see W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw-Hill, 1969; G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 6th ed. Iowa: Iowa State University Press, 1967; and D. E. Bailey, *Probability and Statistics*. New York: Wiley, 1971.

[9] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[10] T. M. Cover and J. M. van Campenhout, "On the possible orderings in the measurement selection problem," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-7, no. 9, pp. 657–661, Sept. 1977.