NAT'L INST. OF STAND & TECH R.I.C.
A11103 911323

REFERENCE

NIST
PUBLICATIONS

NISTIR 4990

OCR Error Rate Versus Rejection Rate for Isolated Handprint Characters

Jon Geist
R. Allen Wilkinson

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

QC
100
.U56
4990
1992

NIST

Note [illegible]
should [illegible] change

NISTIR Workform - NTIS

090 QC100 .U56 no.4990 1992

100 1‡a Geist, Jon

245 10‡a OCR Error Rate Versus Rejection Rate for Isolated Handprint Characters
/ ¦c Jon Geist, R. Allen Wilkinson .

260 0 Gaithersburg, Md. :¦bU.S. Dept. of Commerce, National Institute of Standards and Technology ;¦aSpringfield, Va. :¦bOrder from NTIS, ¦c1992.

300 12p. :‡bill. ;‡c28cm

440 0 NISTIR ;¦v 4990

500

504 Includes bibliographical references (p.12).

500 "Dec. 1992"

700 10 Wilkinson, Allan R. Allen.

740 01

NISTIR 4990

OCR Error Rate Versus Rejection Rate for Isolated Handprint Characters

Jon Geist
R. Allen Wilkinson

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

December 1992

U.S. DEPARTMENT OF COMMERCE
Barbara Hackman Franklin, Secretary

TECHNOLOGY ADMINISTRATION
Robert M. White, Under Secretary for Technology

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

OCR Error Rate Versus Rejection Rate for Isolated Handprint Characters

Jon Geist and R. Allen Wilkinson

Abstract

Over twenty-five organizations participating in the First Census OCR Systems Conference submitted confidence data as well as character classification data for the digit test in that Conference. A three parameter function of the rejection rate r is fit to the error rate versus rejection rate data derived from this data, and found to fit it very well over the range from $r = 0$ to $r = 0.15$. The probability distribution underlying the model $e(r)$ curve is derived and shown to correspond to an inherently inefficient rejection process. With only a few exceptions that seem to be insignificant, all of the organizations submitting data to the Conference for scoring seem to employ this same rejection process with a remarkable uniformity of efficiency with respect to the maximum efficiency allowed for this process. Two measures of rejection efficiency are derived, and a practical definition of ideal OCR performance in the classification of segmented characters is proposed. Perfect rejection is shown to be achievable, but only at the cost of reduced classification accuracy in most practical situations. Human classification of a subset of the digit test suggests that there is considerable room for improvement in machine OCR before performance at the level of the proposed ideal is achieved.

1 Introduction

Over 40 different OCR systems using different preprocessing, feature extraction, and classification algorithms were represented in the First Census OCR Systems Conference.[1] The Conference provided three tests, one with 58,646 segmented digits, a second with 11941 segmented upper case letters, and a third with 12,000 segmented lower case letters. Over 115 test results representing different systems and tests were submitted to NIST for scoring as part of the Conference.

Most of the test results submitted for scoring were accompanied by confidence files, and most of the rest by rejection files. Rejection files contain integers from the set $\{0, 1\}$, one integer per test-character image. A 1 indicates that the hypothetical classification should be scored as a *reject* rather than as *correct* or *incorrect*, and a 0 indicates that the classification should be scored as *correct* if identical to the correct classification, and *incorrect* otherwise. Each rejection file defines one point $e(r)$ on the error rate e

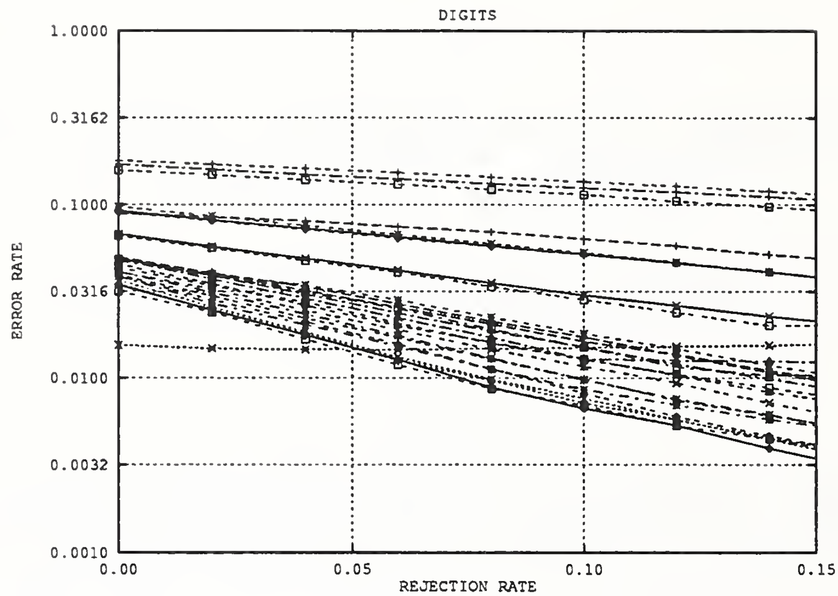

Figure 1: Error rate versus rejection rate for all systems providing confidence data with their classifications for the digit test.

versus rejection rate r curve, so many rejection files per hypothesis file are needed to show the detailed shape of the curve.

Confidence files contain fixed point numbers on the range from 0.0 to 1.0 inclusive, one confidence per test character image. The ordering of the confidence data indicates the order in which the hypothetical classifications should be rejected as unclassifiable when generating error rate versus rejection rate data for the given test and system. Only one confidence file per hypothesis file is needed to show the full detail of the $e(r)$ curve.

Figure 1 (2) shows all of the error rate versus rejection rate $e(r)$ data calculated over the range $0 \leq r \leq 0.15$ for all of the systems that submitted confidence (rejection) files to the Conference. Figure 1 suggests at least two questions: 1) Is there any significance to the fact that all of the curves in that figure seem to have similar shapes with a strong negative correlation between $e(0)$ and $d \ln e(0)/dr$, and 2) how close does the lower envelope of the curves in that figure come to the ideal OCR system performance?

To answer the first question, we derive the relation between the function $e(r)$ and its underlying probability distribution $q(r)$. We then show that the $e(r)$ data calculated from the test results submitted with confidence files is well described over a significant range of r by a simple three parameter equation, and that the probability distribution $q(r)$ associated with this equation represents an inherently inefficient rejection process compared to the perfect rejection process.

We also show that we do not have techniques that allow us to answer the second question. However, comparison with human classification of a subset of the digit test

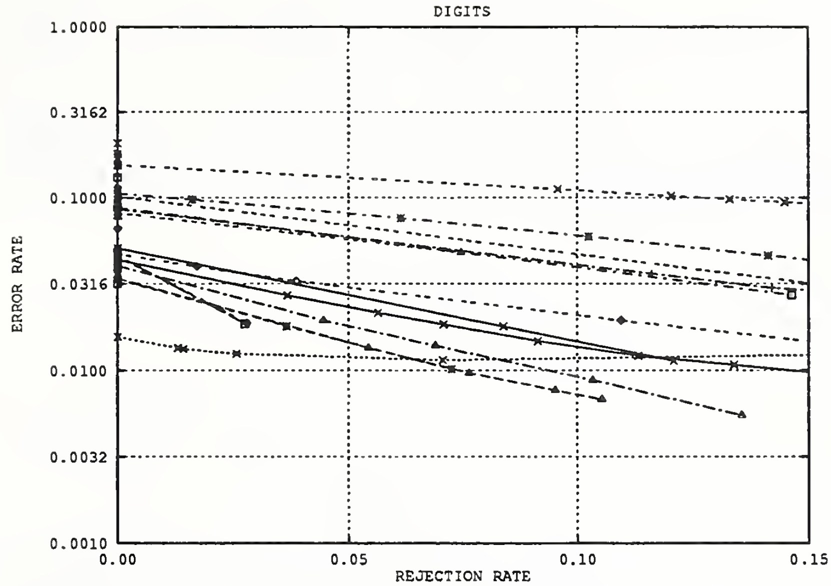

Figure 2: Error rate versus rejection rate for all systems providing rejection data with their classifications for the digit test.

suggests that there is considerable room for improvement in both $e(0)$ and in $e'(r)$ beyond the lower envelope of the $e(r)$ curve in Fig. 1.

The discussion in this paper is confined to the digit test of the First Census OCR Systems Conference, but the results were similar for the upper and lower case letter tests, with a single qualification: $e(0) \leq 0.05, 0.10$, or 0.20 for the digit, upper case, and lower case tests, respectively, for roughly half of the results submitted for scoring. The $e(r)$ curves for all three tests are plotted over the range $0 \leq r \leq 0.50$ for all of the systems submitting results in the Conference report. [1]

2 Error rate versus rejection rate

Let A be a subset of the ASCII character set, let T be a set of segmented character images, let H be a function whose domain is T and whose range is A, and let R be a set of subsets of T including T, such that for each non-empty set that is a member of R there is one and only one set in R that has one less member.

H is a set of hypothetical classifications of T, and R is a complete rejection set for H. The rejection rate r is defined for each subset of T in R as the ratio of the number of members of that subset to the number of members of T. The classifications in H that correspond to the images in each rejection subset of R are rejected rather than scored *correct* or *incorrect* to generate each $e(r)$ point. The classifications that are not rejected are said to be accepted.

For any given T, the range of the variable r is a discrete set, but for simplicity, we treat it as a continuum in the following analysis. Let $q(r)$ be the fraction of the classifications rejected as the rejection rate is changed from r to $r + dr$. Thus, $q(r)$ is the probability as a function of rejection rate r that a rejected classification is actually an incorrect classification. In this case, the error rate $e(r)$, which is defined as the ratio of accepted (unrejected) classifications that are incorrect to the total number of accepted classifications, is given by

$$e(r) = \frac{e(0) - f(r)}{1 - r}, \tag{1}$$

where

$$f(r) = \int_0^r q(s)ds \tag{2}$$

is the fraction of the rejected classifications as a function of r that are actually incorrect, and is equal to r for perfect rejection. Equations 1 and 2 may be combined to give the slope of the error rate,

$$e'(r) = \frac{e(0) - f(r) - f'(r)(1 - r)}{(1 - r)^2} = \frac{e(r) - q(r)}{1 - r}. \tag{3}$$

If $e'(r)$ is zero in eq. 3 , then

$$q(r) = e(r) = e_c, \tag{4}$$

where e_c is a constant. This means that the probability of rejecting an incorrect classification is equal to the fraction of incorrect classifications remaining in the unrejected sample. In this case, the rejection mechanism just rejects classifications at random.

If $q(r)$ is equal to a constant q_c for $r_1 < r < r_2$, then

$$e(r) = \frac{e(0) - f(r_1) - q_c(r - r_1)}{1 - r} \tag{5}$$

and

$$e'(r) = \frac{e(0) - f(r_1) - q_c(1 - r_1)}{(1 - r)^2} \tag{6}$$

over the same subrange. Equation 6 can be written in terms of r_2 instead of r_1, but due to the integral definition of $f(r)$ in eq. 2 only one of the two end points of the interval over which $q(r)$ is constant is needed to express the derivative in this case.

A perfect rejection mechanism is characterized by

$$q(r) = h(r), \tag{7}$$

where $h(r) = 1$ for $0 \leq r \leq e(0)$, and $h(r) = 0$, for $r > e(0)$, in which case,

$$e(r) = h(r)\frac{e(0) - r}{1 - r}. \tag{8}$$

3 Ideal OCR system performance

Wilkinson and Geist [2] point out that it is not necessarily possible even in theory for an OCR system to correctly classify every test image in a real sample of segmented hand-printed characters without errors due to reader/writer (WR) ambiguity. The best performance that can be postulated for an ideal OCR system presented with WR ambiguous characters is 1) that it classify every WR unambiguous character image correctly and assign it a confidence of 1.0, and 2) that it classify every WR ambiguous image as the most probable character and assign it a confidence equal to the WR probability that the classification is correct over the appropriate set of writers and readers. This requires that the system ambiguity be identical to the WR ambiguity for each image in the test set. Conditions 1) and 2) constitute a practical definition of an ideal OCR system with respect to the task of classifying segmented characters.

It is important to distinguish between the system probability $p_S(r)$ that a classification is correct and the WR probability $p_{WR}(r)$ that a classification is correct. The WR probability is an *a priori* probability defined by a set of writers and a set of readers, which establishes the upper bound for the system probability. On the other hand, the system probability that a classification is correct $p_S(r)$ is an *a posteriori* probability equal to $1 - q(r)$. For Conditions 1) and 2) in the preceding paragraph to hold, it is necessary that $p_S(r) = p_{WR}(r)$.

It is also important to understand that an ideal OCR system as defined above will not produce the perfect $e(r)$ curve of eqs. 7 and 8 unless there are no WR ambiguous characters in the set of test images. However, it is possible to trade off performance with respect to ideal OCR system behavior to improve rejection performance. In the extreme case, one can purposely reclassify images with low confidences incorrectly using a character that is not allowed. This assures that the probability of rejecting an incorrect classification is unity, and therefore produces the perfect rejection behavior of eq. 8 while simultaneously increasing the error rate over the range of r where this strategy is employed. The bottom line is that the overall system performance at any value of r is no better, and is probably worse, but the rejection process is perfect. On the other hand, this does not mean that a near perfect rejection curve is necessarily a symptom of non-ideal classification. Perfect rejection is possible with WR unambiguous images. This discussion shows that the analysis of $e(r)$ curves requires care to assure that false conclusions are not drawn.

Finally, the fact that Conditions 1) and 2) are given in terms of probabilities means that an OCR system satisfying them is ideal only in a statistical sense. It is possible for a non-ideal OCR system to out-perform the ideal system on any given test, but, by definition, this cannot happen for the ensemble average of tests over which the WR probabilities are defined.

4 Form of Conference $e(r)$ data

To answer the first question posed in Section 1 about Fig. 1, we attempted to fit all of the data in that figure to a simple model. A visual examination of the curves in that figure suggests that they might be well described by

$$e(r) = \frac{(e_0 - e_{min}) \exp(-r/r_0) + e_{min}}{1 - r}. \tag{9}$$

To test this conjecture, we fit the natural logarithms of the measured $e(r)$ data to the natural logarithm of eq. 9 over the range $0 \leq r \leq 0.15$, where $e_0 \geq 0$, $e_{min} \geq 0$, and $r_0 \geq 0$ were adjusted in the fit. Natural logarithms were used to minimize the variance of the relative differences between the model and calculated $e(r)$ values rather than the variance of the absolute differences.

The results of the fits are summarized in Table 1, which lists the values of e_0, e_{min}, and r_0 for each curve in Fig. 1. This table also lists the residual standard deviation σ of each fit, and two ratios R_1 and R_2 that will be described later.

Eight data points were used in each fit. Three parameters were estimated. This leaves five degrees of freedom in each fit. Because the fits were carried out on the natural logarithms of the data, the residual standard deviations of the fits are actually the standard deviations of the relative differences between the measured error rates and those predicted by eq. 9. Thus a residual standard deviation of 0.01 corresponds to a standard deviation of the relative errors of the fit of 1% over the range of the fit.

Equation 9 fits the data of Fig. 1 very well as should be expected from visual inspection of that figure; only two residual standard deviations are greater than 3%, and two thirds are less than 2%. In fact, most of the $e(r)$ curves for all of the tests and all of the systems are well described by eq. 9 over a subrange $0 \leq r \leq r_{s1}$, and by

$$e(r) = e_s, \tag{10}$$

over a subrange $r_{s1} \leq r \leq r_{s2}$, where e_s, r_{s1}, and r_{s2} are system dependent constants, $r_{s1} \leq r_{s2}$, and $r_{s2} \gg 0.15$. The results of fits of eq. 9 to the $e(r)$ data obtained for the

SYSTEM	σ	e_0	e_{min}	r_0	R_1	R_2
AEG	0.0284	0.0347	0.0011	0.0525	0.6279	0.4889
ASOL	0.0319	0.0922	0.0000	0.2032	0.3971	0.2309
ATT_1	0.0293	0.0326	0.0018	0.0509	0.5915	0.3838
ATT_2	0.0159	0.0363	0.0013	0.0533	0.6942	0.5558
ATT_3	0.0721	0.0505	0.0077	0.0481	0.8828	0.5745
ATT_4	0.0199	0.0417	0.0012	0.0607	0.6540	0.5005
ERIM_1	0.0207	0.0391	0.0002	0.0597	0.6373	0.5336
ERIM_2	0.0151	0.0395	0.0009	0.0635	0.6208	0.4810
GTESS_1	0.0126	0.0667	0.0000	0.1044	0.6127	0.5082
GTESS_2	0.0068	0.0677	0.0030	0.1027	0.6030	0.5358
HUGHES_1	0.0288	0.0501	0.0000	0.0846	0.5697	0.4142
HUGHES_2	0.0298	0.0497	0.0000	0.0901	0.5274	0.4607
IBM	0.0144	0.0349	0.0016	0.0523	0.6233	0.5213
IFAX	0.0032	0.1703	0.0196	0.2062	0.6763	0.6724
KODAK_1	0.0415	0.0490	0.0008	0.0764	0.6109	0.4184
KODAK_2	0.0191	0.0413	0.0006	0.0708	0.5743	0.4117
NESTOR	0.0165	0.0452	0.0022	0.0650	0.7177	0.5268
NIST_2	0.0041	0.0918	0.0000	0.1469	0.5872	0.5552
NIST_3	0.0053	0.0973	0.0000	0.1386	0.6698	0.6357
NIST_4	0.0117	0.0501	0.0014	0.0782	0.6403	0.4941
NYNEX	0.0244	0.0441	0.0022	0.0674	0.6717	0.4708
OCRSYS	0.0042	0.0155	0.0134	0.0348	0.0474	0.0370
THINK_1	0.0093	0.0493	0.0017	0.0720	0.6928	0.5143
THINK_2	0.0195	0.0382	0.0022	0.0539	0.7413	0.5913
UPENN	0.0039	0.0905	0.0004	0.1484	0.5682	0.5436
VALEN_1	0.0078	0.1811	0.0000	0.2525	0.6534	0.5235
VALEN_2	0.0130	0.1595	0.0000	0.2228	0.6604	0.5166

Table 1: Parameters of fit of eq. 9 to data in Fig. 1 for $0 \leq r \leq 0.15$.

upper case and lower case letter tests, which can be found in Ref. [1], are very similar to those shown in Table 1. However, the single ratio shown in that reference is a less useful efficiency measure than the two ratios R_1 and R_2 that are discussed in the next section.

If $e(r)$ satisfies eq. 9, as do the $e(r)$ data shown in Fig. 1, then

$$e'(0) = -\frac{e_0(1 - r_0) - e_{min}}{r_0}. \tag{11}$$

Thus, if $e(r)$ satisfies eq. 9, then

$$\frac{d \ln e(0)}{dr} = \frac{e'(0)}{e(0)} = -\frac{1 - r_0 - e_{min}/e_0}{r_0}. \tag{12}$$

Since $r_0 \approx e(0) \ll 1$ and $e_{min} \ll e_0$ for the systems in Table 1, $d \ln e(0)/dr$ becomes more negative as $e(0)$ decreases. This produces the strong negative correlation between $e(0)$ and $d \ln e(0)/dr$ in Fig. 1.

5 Significance of shape of $e(r)$ function

Equation 10 corresponds to the case where the rejection process has degenerated to a random sampling of the unrejected classifications, as described in connection with eq. 4. On the other hand, according to eq. 3, eq. 9 corresponds to the case where the probability of rejecting a classification that is actually incorrect is given by

$$q(r) = \frac{e_0 - e_{min}}{r_0} \exp(-r/r_0), \tag{13}$$

which can be rewritten in terms of $e(r)$ as

$$q(r) = \frac{e(r)(1-r) - e_{min}}{r_0}, \tag{14}$$

and which is bounded above by

$$q(r) = e(r)/r_0. \tag{15}$$

The probability distribution of eq. 15 is an improvement by a factor of $1/r_0$ over the probability distribution for a completely random rejection process given in eq. 4, but it is still greatly inferior to the distribution for a perfect process. In fact, no probability distribution that is proportional to $e(r)$ can be efficient, because the very act of reducing $e(r)$ through the rejection process reduces the efficiency with which incorrect classifications are rejected.

The two ratios R_1 and R_2 in Table 1 address the efficiency of the rejection process. When $e(r)$ satisfies eq. 9, $e'(0)$ is given by eq. 11 and is bounded below by $e(0) - 1$ according to eq. 6. Thus

$$R_1 = \frac{e'(0)}{e(0) - 1} = \frac{e_0(1 - r_0) - e_{min}}{r_0[1 - e(0)]} \tag{16}$$

in Table 1 is a measure of the efficiency of a rejection process over the range of r (if any) for which it satisfies eq. 9. On the other hand, eq. 9 describes a very inefficient rejection process, so

$$R_2 = \frac{[e(0) - e(r_2)](1 - r_2)}{r_2[1 - e(0)]}, \tag{17}$$

where r_2 has a small value, is a measure of how efficient the early part of the rejection process is compared to the perfect process described by eq. 8. For Table 1, $r_2 = 0.02$. Since R_1 and R_2 measure efficiency over different ranges of r, they are not well correlated in Table 1.

The question that this section addressed was whether or not it is significant that all of the $e(r)$ curves in Fig. 1 appear to have the same shape. The answer is yes. All of the systems producing the $e(r)$ data in that figure seem to employ an inherently inefficient rejection process for which the probability of rejecting an incorrect classification decreases in proportion to the fraction of incorrect classifications remaining in the unrejected set of classifications. For all but three of these systems the proportionality constant ranges from 53% to 74% of the maximum value consistent with this type of rejection process. Two of the three are significantly less efficient, and the third is a little more efficient (88%), but has a relatively large (7%) residual standard deviation of the fit.

The use by 23 of 26 systems of what is essentially the same rejection process with a factor of 1.4 variation in its efficiency constitutes surprising uniformity in light of the fact that $e(0)$ ranges over a factor of more than 5.5 for the same systems, and the fact that these systems employ diverse preprocessing, feature extraction, and classification algorithms.

Both Figs. 1 and 2 have one curve that becomes flat for very small r. Both curves were obtained from the same system because both rejection files and confidence files were submitted with the hypotheses files for this system. This system had a significantly better value for $e(0)$ and a significantly worse value for for $d\ln(e(0))/dr$ than any other system. There is also a system in Fig. 2 whose $e(r)$ curve is defined by only two points, but which employs a rejection process that is significantly more efficient than any of the others shown in Figs. 1 and 2. However, 90% of the classifications that were rejected by this system to generate its second point ($e(0.03) = 0.0186$) in Fig. 2 had been classified incorrectly on purpose by submitting an illegal character as the hypothetical classification. So $e(0)$ was artificially increased to improve rejection. The rest of the $e(r)$ curves in Fig. 2 are not significantly different than those in Fig. 1. Thus, 34 out of 38 OCR systems show remarkable uniformity in the nature of their rejection process, and there does not appear to be anything significant from the point of view of rejection theory about the 4 outliers.

Thus the answer that the shape of the $e(r)$ curve signifies a very inefficient rejection process combined with the fact that there is a surprising uniformity among the $e(r)$ curves leads to a new question. Is the shape of the $e(r)$ curve determined in some fundamental way by the data? For instance, is it possible that the WR unambiguous images are distributed in image space in such a way that inadequacies in preprocessing, feature extraction, and classification generate system ambiguities whose rejection probabilities are given by eq. 13. If so, rejection efficiency will be improved by the same measures that improve forced decision accuracy. If not, special measures would apparently be required to substantially improve rejection efficiency.

6 Comparison with human performance

It is not clear that we have the means to determine the ideal $e(r)$ curve for any given test. Nevertheless, results of human classification are certainly a good start. One of the authors (JG) classified the first 10,000 images in the digit test under the same test conditions as the OCR systems represented in the Conference. The results were $e(0) = 0.0157$ and $e(0.0122) = 0.0035$.

The human value for $e(0)$ is very close to the lowest value, $e(0) = 0.0156$, obtained by any of the systems represented in the Conference, but this is misleading. All images that were perceived by the human classifier to be ambiguous were classified as question marks, which artificially increased $e(0)$ while producing perfect rejection for $0 \leq r \leq 0.0122$. Even a non-optimum strategy like random guessing would have reduced $e(0)$ by $0.1 \times 0.0122 = 0.0012$. Furthermore, many of the ambiguities existed between only two digits, so confining the guessing to the two most likely possibilities might have reduced $e(0)$ by as much as $0.5 \times 0.0122 = 0.0061$. Thus the human might have been able to obtain $0.0096 \leq e(0) \leq 0.0145$, while leaving $e(0.0122)$ unchanged. If the human were able to choose the more (or most) likely of the classifications when ambiguities existed, then even lower values for $e(0)$ would be possible.

Moreover, the fact that the human value $e(0.0122) = 0.0035$ is well over a factor of four lower than the lowest value of $e(0.0122)$ in Figs. 1 and 2 strongly suggests that the lower envelope of the curves in those figures is still far from the performance of an ideal OCR system. The only caveats are that the human performance was obtained for a single human on a single test that is a subset of the test used for the OCR systems. Grother [3] has shown that it is unlikely that the human result would be significantly different for the complete digit test. It is also unlikely that the factor of four superiority of the human result is a statistical fluke that would change significantly over an ensemble of tests involving more writers and more human classifiers.

There is a fundamental problem with using a single human in an attempt to determine the ideal $e(r)$ curve for a set of real-world character images such as used in the Conference. Humans are not comfortable, and maybe not even capable, of generating confidences for their classifications. Humans with sufficient incentive are quite happy rejecting ambiguous characters images while classifying those that they find unambiguous, but they are not so comfortable assigning a single classification to an ambiguous image, much less a confidence. Even the plurality vote of a large number of human classifiers will suffer from this problem unless it happens that different humans usually find different character images ambiguous.

Our experience suggests that it might be possible to get humans to generate the data needed to calculate an $e(r)$ curve in a multipass process. On the first pass each human would hit the appropriate keyboard key to classify the subjectively unambiguous characters, and reject the rest by typing a question mark. The second pass would present only the rejected characters for classification. On this pass each human would hit two different keys to assign two different classes to any images that were subjectively ambiguous between only two characters, and so forth. We can even imagine letting the human classifiers hit the key corresponding to each character of an ambiguous character set a number of times proportional to his or her subjective estimate of the relative plausibility of the classification. Still, it is not clear that humans would be comfortable with this task when more than two-character classifications were attempted. Nevertheless, pooling the results of a number of human multipass classifications might give a good estimate the ideal $e(r)$ curve for a given set of test images, at least over a useful subrange of r.

7 Conclusion

We have derived the relation between $e(r)$ and its underlying probability distribution $q(r)$. We also showed that the $e(r)$ data submitted for the digit test of the First Census OCR Systems Conference are well described for $0 \leq r \leq 0.15$ by eq. 9, and that the corresponding probability distribution $q(r)$, given in eq. 13, describes an inherently inefficient rejection process compared to the perfect rejection process. We have introduced some measures of the efficiency of the rejection process for isolated character OCR, and have proposed a definition of ideal performance in the latter task. The definition is statistical in nature, but it is general enough to allow ideal performance to be better than human performance, since we have no reason to expect human performance to be ideal. We have also discussed the difficulties of determining ideal performance on any given test, and have compared the digit test results to human classification of a subset of that test. The results suggest that there is considerable room for improvement in machine OCR before it can challenge human performance for accuracy. Of course, that does not mean that it cannot already challenge human performance in applications where accuracy must be balanced with cost.

References

[1] R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. L. Wilson. The First Optical Character Recognition Systems Confernce. Technical Report NISTIR 4912, National Institute of Standards and Technology, August 1992.

[2] R. A. Wilkinson, M. D. Garris, and J. Geist. Machine-Assisted Huamn Classification of Segmented Characters For OCR Testing and Training. In *Proceedings: Character Recognition Technologies*, volume 1906. San Jose, SPIE, February 1993.

[3] P. J. Grother. Cross Validation Comparison of NIST OCR Databases. In *Proceedings: Character Recognition Technologies*, volume 1906. San Jose, SPIE, February 1993.

NIST-114
(REV. 9-92)
ADMAN 4.09

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

MANUSCRIPT REVIEW AND APPROVAL

(ERB USE ONLY)

ERB CONTROL NUMBER	DIVISION
W93-0212	875
PUBLICATION REPORT NUMBER	**CATEGORY CODE**
NISTIR 4990	290
PUBLICATION DATE	**NUMBER PRINTED PAGES**
DECEMBER 1992	15

INSTRUCTIONS: ATTACH ORIGINAL OF THIS FORM TO ONE (1) COPY OF MANUSCRIPT AND SEND TO: THE SECRETARY, APPROPRIATE EDITORIAL REVIEW BOARD.

TITLE AND SUBTITLE (CITE IN FULL)

OCR Error Versus Rejection Rate for Isolated Handprint

CONTRACT OR GRANT NUMBER

TYPE OF REPORT AND/OR PERIOD COVERED

AUTHOR(S) (LAST NAME, FIRST INITIAL, SECOND INITIAL)

Geist, Jon
Wilkinson, R. Allen

PERFORMING ORGANIZATION (CHECK (X) ONE BOX)

- [X] NIST/GAITHERSBURG
- [ ] NIST/BOULDER
- [ ] JILA/BOULDER

LABORATORY AND DIVISION NAMES (FIRST NIST AUTHOR ONLY)

Computer Systems Laboratory, Advanced Systems Division (875.12)

SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

RECOMMENDED FOR NIST PUBLICATION

- [ ] JOURNAL OF RESEARCH (NIST JRES)
- [ ] J. PHYS. & CHEM. REF. DATA (JPCRD)
- [ ] HANDBOOK (NIST HB)
- [ ] SPECIAL PUBLICATION (NIST SP)
- [ ] TECHNICAL NOTE (NIST TN)
- [ ] MONOGRAPH (NIST MN)
- [ ] NATL. STD. REF. DATA SERIES (NIST NSRDS)
- [ ] FEDERAL INF. PROCESS. STDS. (NIST FIPS)
- [ ] LIST OF PUBLICATIONS (NIST LP)
- [ ] NIST INTERAGENCY/INTERNAL REPORT (NISTIR)
- [ ] LETTER CIRCULAR
- [ ] BUILDING SCIENCE SERIES
- [ ] PRODUCT STANDARDS
- [ ] OTHER

RECOMMENDED FOR NON-NIST PUBLICATION (CITE FULLY) [X] U.S. [ ] FOREIGN

IS&T/SPIE 1993 Symposium on Electronic Imaging

PUBLISHING MEDIUM

- [X] PAPER
- [ ] CD-ROM
- [ ] DISKETTE (SPECIFY)
- [ ] OTHER (SPECIFY)

SUPPLEMENTARY NOTES

ABSTRACT (A 1500-CHARACTER OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, CITE IT HERE. SPELL OUT ACRONYMS ON FIRST REFERENCE.) (CONTINUE ON SEPARATE PAGE, IF NECESSARY.)

vendors, manufacturers (might be a better choice of words)

Over twenty-five systems participating in the First Census OCR Systems Conference submitted confidence data as well as character classification data for the digit test in that Conference. A three parameter function of the rejection rate r was fit to the error rate versus rejection rate data derived from this data, and found to fit it very well over the range from $r = 0$ to $r = 0.15$. The probability distribution underlying the model $e(r)$ curve was derived and shown to correspond to an inherently inefficient rejection process. With only a few exceptions that seem to be insignificant, all of the systems submitting data to the Conference for scoring seem to employ this same rejection process with a remarkable uniformity of efficiency with respect to the maximum efficiency allowed for this process. Human classification of a subset of the digit test suggests that there is considerable room for improvement in the performance of machine OCR before the theoretical ideal is achieved.

vendors

KEY WORDS (MAXIMUM 9 KEY WORDS; 28 CHARACTERS AND SPACES EACH; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES)

error rate; isolated character; hand print; OCR; Optical Character Recognition; rejection rate; segmented character.

AVAILABILITY

- [X] UNLIMITED
- [ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NTIS.
- [ ] ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GPO, WASHINGTON, D.C. 20402
- [X] ORDER FROM NTIS, SPRINGFIELD, VA 22161

NOTE TO AUTHOR(S) IF YOU DO NOT WISH THIS MANUSCRIPT ANNOUNCED BEFORE PUBLICATION, PLEASE CHECK HERE. [ ]

ELECTRONIC FORM