Form Design for High Accuracy Optical Character Recognition

Michael D. Garris, mdg@magi.ncsl.nist.gov
Darrin L. Dimmick, dld@magi.ncsl.nist.gov

National Institute of Standards and Technology,
Building 225, Room A216
Gaithersburg, Maryland 20899
Phone: (301)975-2928, FAX: (301)840-1357

Published in IEEE Transactions PAMI, June 1996.

ABSTRACT

Financial institutions, insurance companies, and government agencies are all aggressively pursuing the integration of automated forms processing into their everyday work flows. To use existing optical character recognition (OCR) technology, the forms that are currently hand-keyed will probably need to be redesigned. This paper presents some of the quantitative results generated by a comprehensive study of three versions of a redesigned tax form. Analyses show that using separately spaced bounding character boxes to represent fields provides superior machine readability over fields without character boxes, fields containing vertical ticks (combs), and fields with adjoining character boxes. It is also shown that character boxes containing two vertically stacked ovals cause writers much more difficulty to complete than do empty character boxes. The analyses also provide quantitative proof that writer idiosyncratic responses on forms are the major source of errors within the recognition system. These idiosyncracies (such as writers crossing out previously printed characters or writing over them) must be effectively handled in order improve recognition performance. This paper demonstrates how form design can help, and it provides empirical data to support some of the *rules-of-thumb* by measuring the impact specific changes to a form have on machine readability and on the writer.

1. INTRODUCTION

Numerous companies are emerging that sell a wide range of document processing and optical character recognition (OCR) products and integration services. However, having a computer system and software product in hand is not enough to ensure successful reduction in labor costs. Many factors must be examined. For example, the forms that are currently hand-keyed will probably need to be redesigned. What factors should be considered in designing new forms? Are there

certain ways a field can be represented that may both influence the writer and increase machine recognition accuracy? This paper provides some answers to these questions.

The Internal Revenue Service (IRS) is one government agency that is actively integrating document processing technology as part of its tax modernization effort. Realizing how important form redesign is to the success of OCR integration, IRS staff are studying various redesigned tax forms. This paper presents results obtained by studying three similar versions of one of these redesigned tax forms [1]. To do this, the National Institute of Standards and Technology's (NIST) public domain form-based handprint recognition system [2] was modified to process these types of forms. Images of the forms were run through multiple configurations of the recognition system and the recognized text was scored and analyzed.

Automated recognition of handprint has been the topic of much research, and in May of 1992, the First Census Optical Character Recognition Systems (COCR1) Conference sponsored by the Bureau of the Census was run by the NIST [3]. The Conference compared the recognition results from 26 different participating organizations from the private sector, academia, and government. Properly segmented images of individual handprinted characters were recognized and the results reported. It was demonstrated that zero-reject error rates as low as 3% could be achieved on large samples of digits. Unfortunately, few real applications can be reduced to recognizing only well segmented and isolated characters. The processing of field information entered onto forms requires complex and intelligent processing to get to the point of classifying isolated character images. Steps including form registration, form removal, field isolation, and field segmentation must be conducted prior to classifying the characters in each field. Each one of these steps adds complexity and the potential for error. The analyses presented in this paper demonstrate that an automated forms processing system can achieve COCR1 levels of performance through proper form design combined with the detection of writer idiosyncratic responses.

Section 2 describes the composition of the redesigned tax forms. The components of the recognition system are discussed in Section 3. Section 4 gives an overview of how the recognition system results were scored, and observations on overall recognition system performance are pre-

sented. The results of categorizing various writer idiosyncratic responses are discussed in Section 5, and simulated performance results are reported when fields containing these idiosyncrasies are detected and rejected by the recognition system. Conclusions are drawn in Section 6.

2. REDESIGNED FORMS

The redesigned tax forms used in this study (570 in all) are double-sided and portrait-oriented with a page width of 215 cm and a page height of 279 cm (8.5 by 11 in). Unlike the tax forms currently in use, the instructional information on the redesigned forms is greatly reduced. There is typically a one-line heading for each field, the fields are demarcated within each region using blue drop-out ink, and the forms have a black registration mark in each corner of the page. Figure 1 contains a copy of the front of a redesigned tax form. This paper presents quantitative results obtained from the numeric fields on the forms. The results based on all the fields, including alphabetic text and mark-sense fields, are reported in [1].

****** Figure 1 about here ******

Three form versions were used in this study. The form shown in Figure 1 is of type *P1*. The only difference between the three versions is in the representation of money fields as illustrated in Figure 2. Money fields on P1 forms are demarcated as a single bounding box encompassing the entire field. Commas and decimal points are printed in blue drop-out ink with a vertical tick mark above each punctuation mark. The second form, referred to as type *P2*, has money fields demarcated by separately spaced boxes bounding each character position in the field. The last form, type *P3*, has money fields demarcated by separately spaced character boxes, and each box contains two vertically stacked ovals. The ovals are *intended* to guide the shape of characters as they are written so irregularities and variations are minimized. For the purpose of comparison, the structure of SSN fields is also illustrated in Figure 2.

****** Figure 2 about here ******

The forms, filled out by hand, were scanned front and back; the resulting images were scanned at 12 pixels per millimeter (300 pixels per inch) and digitized as binary (black and white).

Writers filling out the redesigned forms were instructed to enter one of two sets of contrived field values. These two sets of values were used by the scoring package as reference strings (ground truth) to measure recognition system performance.

3. RECOGNITION SYSTEM COMPONENTS

The NIST public domain form-based handprint recognition system was originally designed to process Handwriting Sample Forms as on the CD-ROM, *NIST Special Database 19* [4]. This software distribution of the public domain OCR system is available free of charge by writing the authors a letter of request.

The modified system performs form registration using a shape-based feature detector, a correlated run length algorithm (CURL) [5], to locate the registration marks in the corners of a form. By locating these marks, rotation, translation, and scale distortions within the image are measured, and an image of a blank form (transformed to conform to the distortion within the input image) is subtracted from the input image. A spatial template is used to isolate the handprint in the fields, and the fields are then processed based on their contextual type. Each character field is segmented into individual images, one character per image. The character images are size and slant normalized and feature vectors are derived using the Karhunen Loève (KL) transform [6]. The feature vectors are classified using a Probabilistic Neural Network (PNN) [7]. It has been shown that PNN outperforms other neural networks in terms of zero-reject accuracy [8]; therefore, it was selected over more popular neural network paradigms, such as Multi-Layer Perceptions [9].

Two different segmentation methods were used in the recognition system. The first method, referred to as the *blob segmentor*, is based on connected component labeling and was used on the P1 money fields. Each blob (a group of connected pixels) in the isolated field is assumed to be a separate character. This segmentor is prone to errors because, if two characters touch, a single blob will contain both characters, and if a character is comprised of several disjoint strokes, a blob will contain only part of the character (one of the strokes). To overcome the deficiencies of the blob segmentor, a second segmentation algorithm was developed. The P2 and P3 money fields on the rede-

signed forms have character positions demarcated with bounding boxes as illustrated in Figure 2. Assuming the writer stayed within the sides of the boxes, segmentation errors can be minimized by simply cutting along these boundaries. This segmentation scheme is referred to as the *cut segmentor.*

In [1], a comprehensive report gives a detailed description of each recognition system component, including the performance statistics from six different recognition system configurations. The results from only a very small portion of the comprehensive study are presented here.

4. RECOGNITION SYSTEM RESULTS

NIST has developed a recognition system testing methodology that has been implemented as the NIST recognition system scoring package [10]. The scoring package has been developed to measure the performance of character recognition systems and automated form processing systems. For this study, the recognition system was presented the images of redesigned forms, and the recognized hypothesis field values were stored. The scoring package reconciled the hypothesis strings with the reference strings the writers were instructed to enter on the form. If a hypothesis string was not identical to its reference string, errors were tallied accordingly, regardless of why they didn't match.

****** Figure 3 about here ******

The table and graph in Figure 3 summarize one system configuration's recognition performance across all of the money fields on the forms. The first two columns of the table list character recognition error rates, and the third column lists field error rates. The measure used in the first column (character output error) is computed according to Equation (1). The measure in the second column (character decision error) is computed according to Equation (2). The first measure represents how the system performs overall, while the second measure represents how well the character classifier performs on those images that are segmented. The third column lists the percentage of fields incorrectly recognized. In this case, the system's hypothesized field string must match the reference string value exactly (character for character) to be considered correct.

$$\text{<character output error>} = 1 - \frac{\text{<\# segments recognized correctly>}}{\text{<\# characters in reference strings>}} \quad (1)$$

$$\text{<character decision error>} = 1 - \frac{\text{<\# segments accepted and recognized correctly>}}{\text{<\# segments>}} \quad (2)$$

The graph at the bottom of Figure 3 plots an error response curve based on rejection rates for each form version. The character classifier used in this study computes a confidence value associated with each classification decision it makes. By rejecting low confidence classifications, many of the errors made by the character classifier are detected and avoided. Rejecting classifications is designed to increase the accuracy of classifier decisions at the cost of decreasing the volume of automated system throughput.The resulting error rates are plotted on a log scale. The percentage of system error plotted in the error-reject graph is calculated according to Equation (2).

There was a consistently tight grouping of P1, P2, and P3 results across SSN fields (not shown). This grouping was expected, as the SSN fields are consistently represented across the three form versions. The results from the SSN fields served as a control group against which results on money fields were compared.

Unlike the SSN fields, the results for money fields in Figure 3 exhibit significant separations between P1, P2, and P3 results. Although not explicitly computed here, previous tests on the NIST recognition system show that changes approximately greater than 1.0% are statistically significant [3]. The changes in performance are primarily attributed to the differences in the way money fields are represented on the forms. This quantitatively supports the assertion that changing the design and layout of a form can directly influence character recognition system performance. The other system configurations, not presented in this paper, exhibited similar behavior.

The scoring package was also used to tally the number of characters deleted and inserted by the recognition system. These segmentation errors were accumulated over all the money fields

on the redesigned forms. The segmentation error rates were computed as $(D+I)/R$, where the number of deleted (D) and inserted (I) characters are added together and normalized by dividing the sum by the number of reference characters in all the fields (R). As a results, P1 money fields achieved a 14% segmentation error rate, while P2 and P3 fields had an 8% error rate. The segmentation errors for money fields are lower for P2 and P3 versions than they are for P1 versions. The difference is due to the blob segmentor being used on P1 money fields and the cut segmentor being used on P2 and P3 money fields. The inter-character demarcations in the P2 and P3 fields facilitated the use of the cut segmentor, which in turn produced fewer segmentation errors. Not only does form design influence the writer, but it can be used influence / guide the system as well.

5. ANALYSIS OF WRITER IDIOSYNCRACIES

One major challenge that recognition systems face is being able to handle a wide variety of writer idiosyncratic responses. At times these relatively unpredictable responses make a field unreadable by the computer. For example, if a writer leaves a field blank, enters the wrong information, or crosses out a previously written field value, the recognition system can do very little to compensate for these events apart from applying some type of external context. It is conceivable that certain types of these responses can be automatically detected (for example, blank fields), thereby reducing system errors and increasing recognition system performance. An analysis was conducted in which this type of detection was simulated.

The performance measures compiled across the test set of forms and reported in Figure 3 contain a combination of errors due to the writer along with other sources of system errors. An independent field study was conducted in which a select number of fields were manually verified to match their corresponding reference values. Any field not matching its associated reference value was removed from the performance analysis and later categorized as to why it was removed.

Two particular fields selected (a money field and an SSN field) were to be completed on every form, providing the maximum coverage across the set of redesigned forms. Those fields not correctly entered by the writers were logged and categorized. It was observed that writers occasion-

ally 1.) leave a field blank when it requires an entry, 2.) transcribe the wrong value onto the forms, 3.) cross out previously printed characters and write over them, 4.) print radically malformed characters that would challenge any character classifier, 5.) leave spurious marks in the field such as partial erasures, and 6.) provide punctuation marks in fields where the punctuation was already provided on the form. Even though these sources of error seem rather obvious, they must be effectively handled in order for recognition system performance to improve.

A breakdown of idiosyncratic responses for a selected money field is shown in Figure 4. The graph plots the percentage of fields determined to contain one or more of the categories listed above. The percentages are broken out by form version (P1, P2, and P3). The graph plots these percentages with the x-axis representing each type of idiosyncracy and the y-axis representing the corresponding percentage of fields removed. Notice the P3 money field contains a significantly higher amount of anomalous responses than the P1 and P2 versions. On the other hand, the breakdown of idiosyncratic responses for the SSN field (not shown here) were relatively uniform. Remember these SSN fields are represented consistently across the form versions, and the relative uniformity implies that the differences seen for money fields are significant.

****** Figure 4 about here ******

Simulated recognition system performance across a set of selected money fields is recorded in Figure 5. These performance measures were derived from those fields determined to be free of idiosyncratic responses. The recognition system performs best on the P3 then P2 versions of money fields, while it does not perform nearly as well on the P1 version. This observation again supports the assertion that fields represented by separately spaced bounding boxes for each character improve the accuracy of the recognition system. A large separation across form versions is seen in the graph. P1 versions of money fields produce an 11% character output error rate, P2 versions produce 6%, while P3 versions only produce 3%. The 3% performance is near the best isolated handprint digit recognition results reported in COCR1; however, this was achieved by rejecting many more P3 fields.

****** Figure 5 about here ******

The SSN field analysis achieved a character error rate of 9% after idiosyncracies were removed, which is substantially higher than the character error rates associated with P2 and P3 money fields. This leads to the assertion that the recognition accuracy of SSN fields can be greatly improved by adopting the separately spaced bounding character box field structure. As can be seen in Figure 2, the SSN fields are represented with boxes for each character, but the boxes share interior sides and are not separately spaced. Evidence shows this spacing is needed to properly influence the writer and improve recognition.

6. CONCLUSIONS

This paper has presented some of the quantitative results generated by a study of three versions of a redesigned tax form. A comprehensive report can be found in [1]. The NIST public domain OCR system was modified and used in conjunction with the NIST scoring package to generate performance measures at the field and character levels. Quantitative analyses showed that representing fields with separately spaced bounding character boxes provides superior machine readability over fields without character boxes, fields containing vertical ticks (combs), and fields with adjoining character boxes. It was also shown that character boxes containing two vertically stacked ovals cause writers much more difficulty to complete than do empty character boxes. Due to consistencies exhibited across system configurations and form versions within control groups of fields, one can expect a similar gain in system performance if all fields on a form, including alphabetic fields, are represented with separately spaced character boxes. Segmentation was also improved by taking advantage of inter-character demarcations provided by the form design. This demonstrates that not only does form design influence the writer, but it can be used to improve / guide the system as well. Analyses also conclude that writer idiosyncratic responses on forms are the primary source of errors within the recognition system. This study has demonstrated that by redesigning forms, these idiosyncrasies are reduced, and the remaining errors in an automated form processing system can be effectively reduced to classification errors.

REFERENCES

[1]. M. D. Garris and D. L. Dimmick, "Evaluating form designs for optical character recognition," Technical Report NISTIR 5364, National Institute of Standards and Technology, Feb. 1994.

[2]. M. D. Garris, J. L. Blue, G. T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C. L. Wilson, "NIST form-based handprint recognition system," Technical Report NISTIR 5469, National Institute of Standards and Technology, July 1994

[3]. 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 Census Optical Character Recognition System Conference," Technical Report NISTIR 4912, National Institute of Standards and Technology, July 1992.

[4]. P. J. Grother, "Handprinted Forms and Character Database, *NIST Special Database 19*," Technical Report and CD-ROM, National Institute of Standards and Technology, March 1995.

[5]. M. D. Garris, "Correlated run length algorithm (CURL) for detecting form structures within digitized documents," in *Proc. Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 413-424, Las Vegas, Apr. 1994.

[6]. P. J. Grother, "Karhunen Loève feature extraction for neural handwritten character recognition," In *Proc. Applications of Artificial Neural Networks III*, vol. 1709, pp. 155-166, SPIE, Orlando, Apr. 1992.

[7]. Donald F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3(1), pp. 109-119, 1990.

[8]. J. L. Blue, B. T. Candela, P. J. Grother, R. Chellappa, and C. L. Wilson, "Evaluation of pattern classifiers for fingerprint and OCR applications," *Pattern Recognit.*, vol. 27, no. 4, pp. 485-501, 1994.

[9]. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, Cambridge: MIT Press, 1986, pp. 318-362.

[10]. M. D. Garris and S. A. Janet, "Scoring Package release 1.0," *NIST Special Software 1*, vol. SP, National Institute of Standards and Technology, Oct. 1992.

FIGURE CAPTIONS

Figure 1. Example of a completed first page of a redesigned tax form.

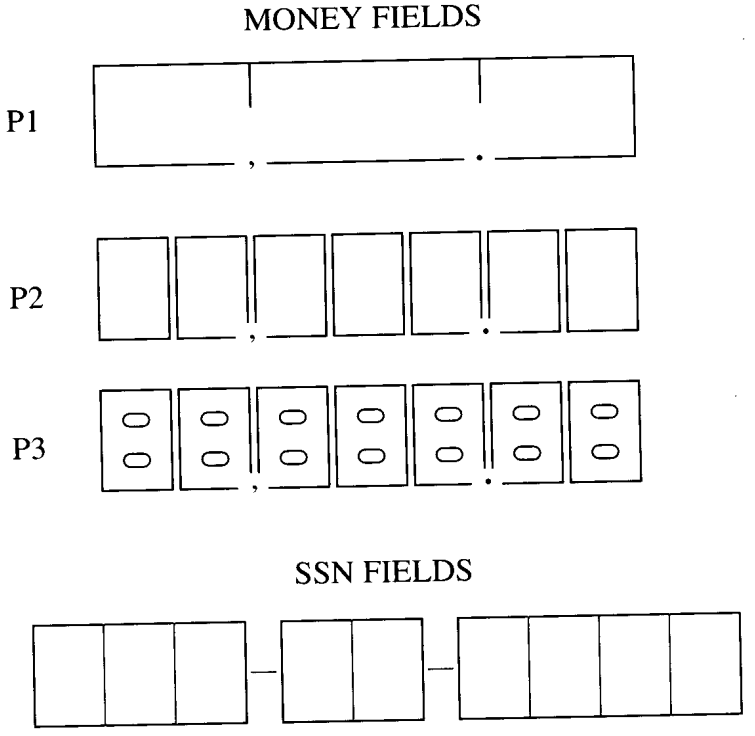
Figure 2. Numeric field representations on the redesigned forms.

Figure 3. Overall recognition performance on money fields.

Figure 4. Breakdown of writer idiosyncracies for a selected money field.

Figure 5. Recognition performance on a money field with idiosyncratic responses removed.

FIGURE 2



010

FIGURE 1

B01-10

Form **1040-T** Department of the Treasury—Internal Revenue Service **U.S. Individual Income Tax Return** **199X** OMB No. Version P1

Please Print or Type Here

Your first name and initial: TINA N.
Your last name: TAXPAYER
Spouse's first name and initial (if a joint return): TOM N.
Spouse's last name (if a joint return): TAXPAYER
Home address (number and street): 1100 MAIN STREET
Apt. number: 101
City, town or post office: NEWTOWN
State: KS
Country (if not the U.S.)
ZIP code: 71229

Presidential Election Campaign (Notes: Checking "Yes" will not ...)

Do you want $1 to go to this fund? (•) Yes ( ) No
If joint return, does spouse want $1 to go to this fund? ( ) Yes (•) No

Filing Status and Exemptions

1 ( ) Single
2 (•) Married Filing Joint
3 ( ) Married Filing Separate
4 ( ) Head of Household
5 ( ) Widow(er) Year
6a (•) Yourself
6b (•) Spouse
6d ( ) Pre-1985 agreement
6e Total Exemptions: 4

6c List of dependents

(1) Name (first, initial, and last name): TONY N. TAXPAYER
(2) ( ) Under age 1
(3) If age 1 or older, dependent's SSN: 567-89-1234
(4) Relationship: SON
(5) Months in home (1992): 12

(1) Name (first, initial, and last name): TANYA N. TAXPAYER
(2) ( ) Under age 1
(3) If age 1 or older, dependent's SSN: 456-78-9123
(4) Relationship: DAUGHTER
(5) Months in home (1992): 12

(1) Name (first, initial, and last name):
(2) ( ) Under age 1
(3) If age 1 or older, dependent's SSN:
(4) Relationship:
(5) Months in home (1992):

Number of your children on 6c who:
• lived with you: 2
• didn't live with you due to divorce or separation:
Number of other dependents on 6c:

Social Security Number, Signature, and Occupation

Your social security number: 123-45-6789
Spouse's social security number: 678-91-2345

Under penalties of perjury, I declare that I have examined this return and accompanying schedules and statements, and to the best of my knowledge and belief, they are true, correct, and complete. Declaration of preparer (other than taxpayer) is based on all information of which preparer has any knowledge.

Your signature
Date
Your occupation
Spouse's signature
Date
Spouse's occupation

Paid Preparer's Use Only

Preparer's signature
Date
( ) Self-employed
SSN
Firm's name (or yours if self-employed) and address
EIN

Attach Form W-2 here.

7	Wages	21,724.90
8a	Taxable interest	25.32
8b	Tax-exempt interest	
9	Dividend income	150.89
10	Taxable refunds, etc.	
16a	Total IRA distribution	
16b	Taxable amount	
17a	Pensions & annuities	
17b	Taxable amount	
20	Unemployment compensation	
21a	Social security benefits	
21b	Taxable amount	
22	Other income	
23	Total income	21,901.11

Adjustments to Income

24a	Your IRA deduction	750.00
24b	Spouse's IRA deduction	500.00

Tax Computation

32	AGI	20,651.11
33a	( ) 65 or older ( ) Spouse 65 or older ( ) Blind ( ) Spouse blind Total	
33b	( ) Claimed elsewhere	
33c	( ) MFS & itemized or dual-status	
37	Taxable income	4,112.31
38	a (•) Tax Table b ( ) Schedules d ( ) Form 8615	
38e	Form 8814	
38	Tax	619.00

FIGURE 3

System Configuration Performance Money Fields			
	Character Output Error	Character Decision Error	Field Error
P1	20.4%	12.3%	19.6%
P2	10.9%	8.1%	11.9%
P3	10.1%	7.3%	10.4%

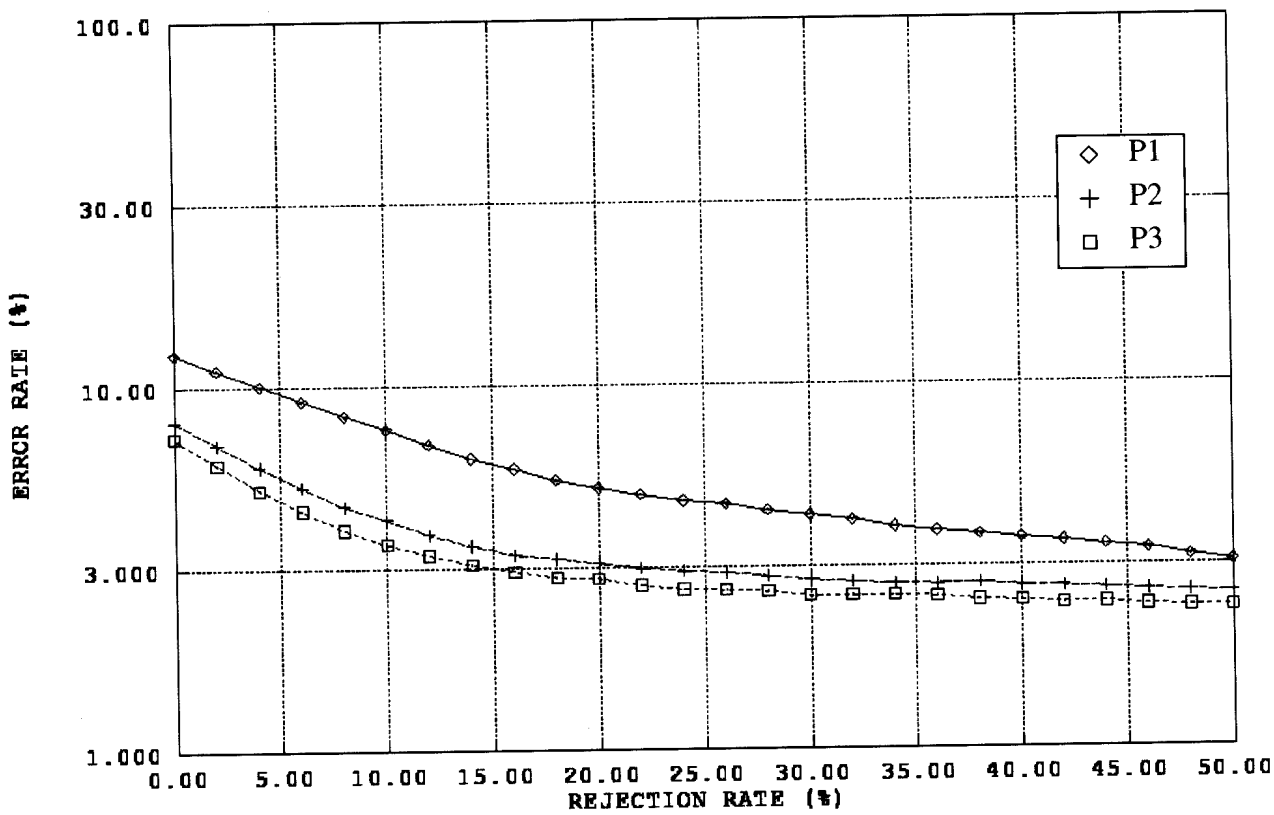


FIGURE 4

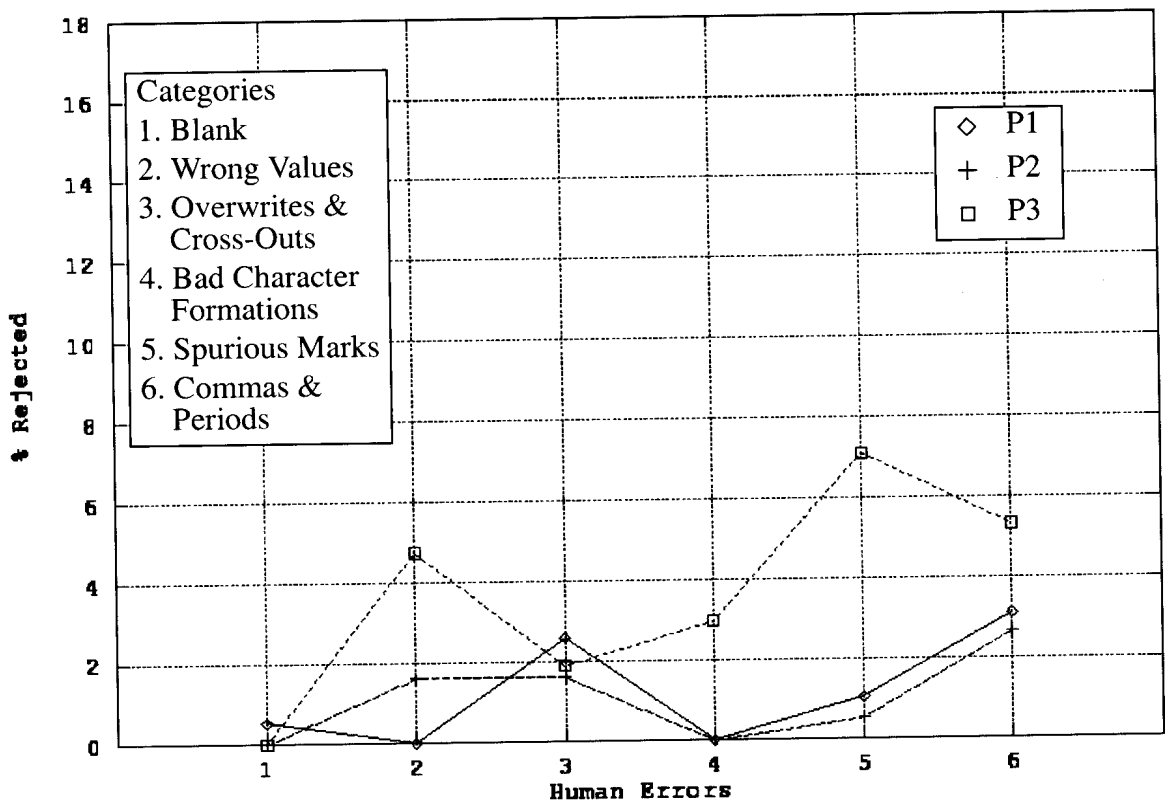
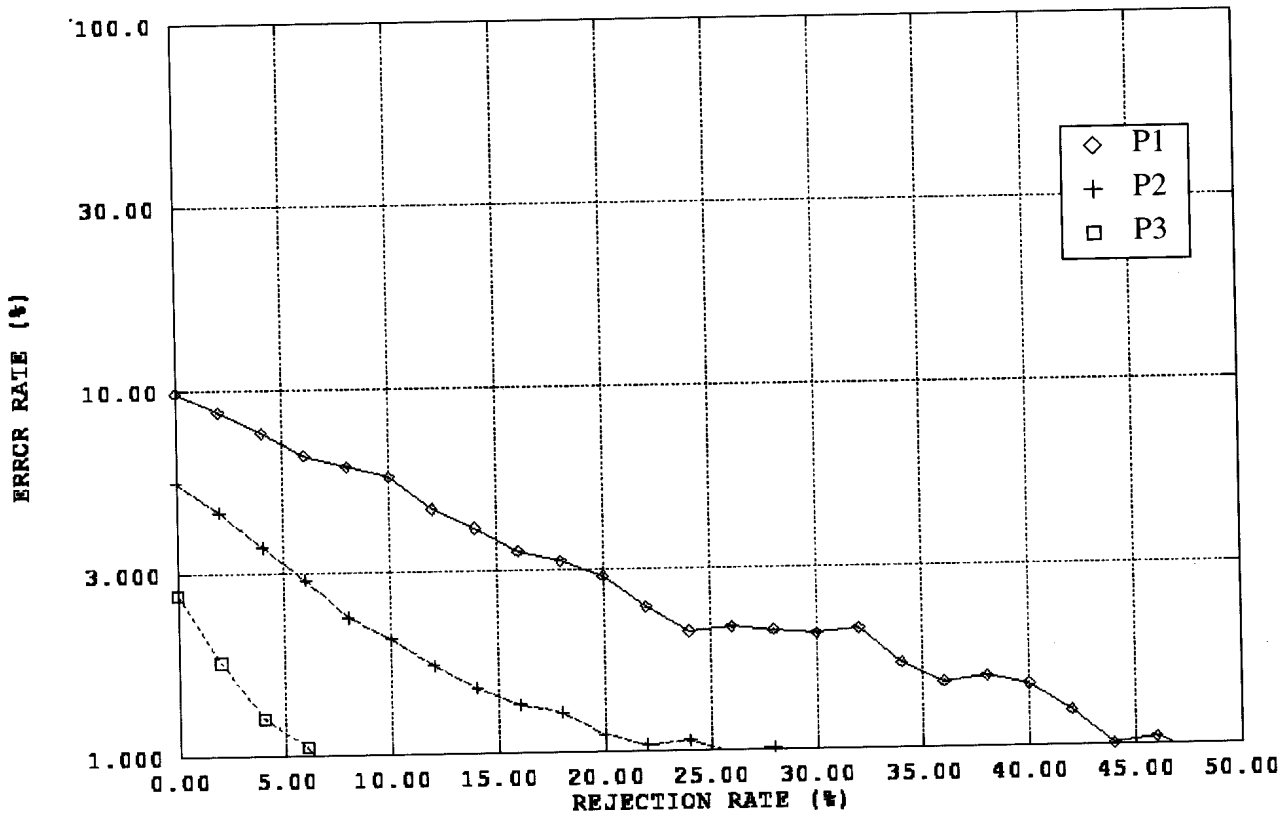

% Rejected
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
Human Errors
Categories
1. Blank
2. Wrong Values
3. Overwrites & Cross-Outs
4. Bad Character Formations
5. Spurious Marks
6. Commas & Periods
P1
P2
P3

FIGURE 5

System Configuration Performance Money Fields with Idiosyncracies Removed		
	Character Output Error	Character Decision Error
P1	11.0%	9.9%
P2	5.8%	5.6%
P3	2.8%	2.8%



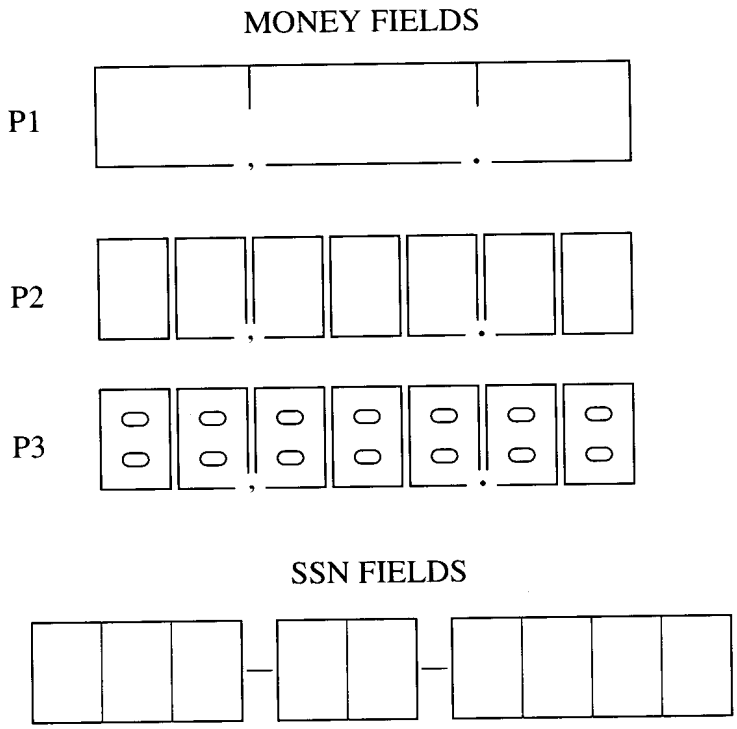


Figure 1

System Configuration Performance Money Fields			
	Character Output Error	Character Decision Error	Field Error
P1	20.4%	12.3%	19.6%
P2	10.9%	8.1%	11.9%
P3	10.1%	7.3%	10.4%

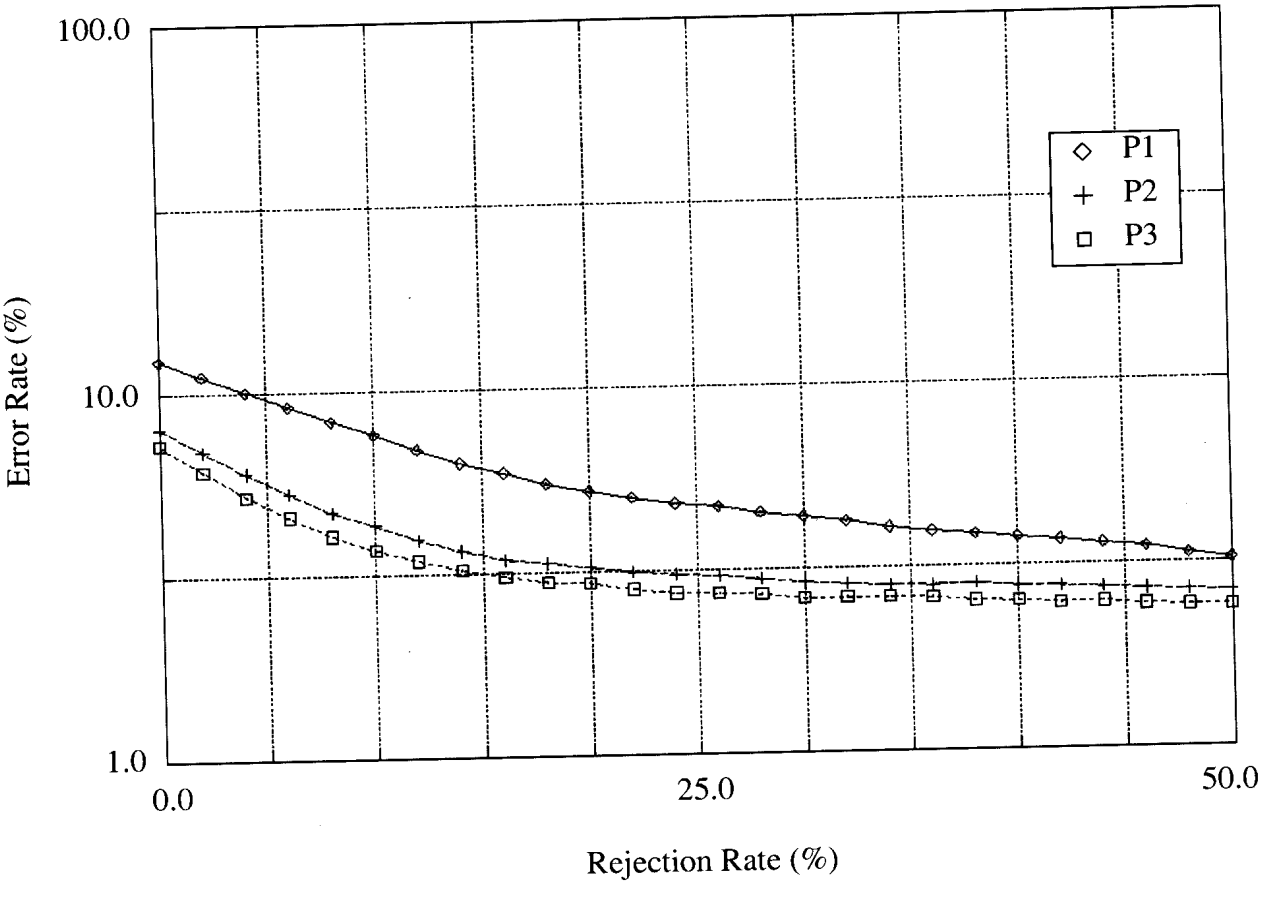


Figure 2

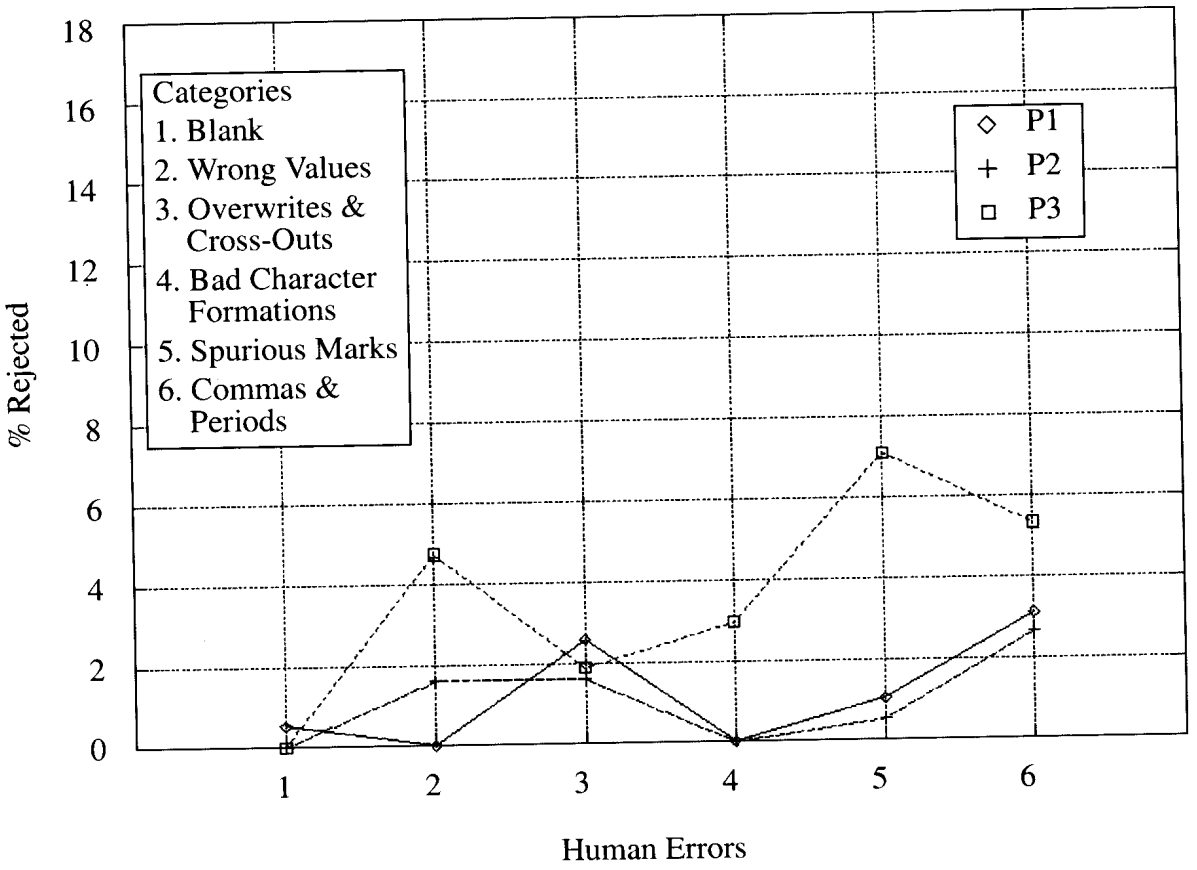

18
16
14
12
10
8
6
4
2
0
% Rejected
Categories
1. Blank
2. Wrong Values
3. Overwrites & Cross-Outs
4. Bad Character Formations
5. Spurious Marks
6. Commas & Periods
P1
P2
P3
1
2
3
4
5
6
Human Errors

Figure 3

System Configuration Performance Money Fields with Idiosyncracies Removed		
	Character Output Error	Character Decision Error
P1	11.0%	9.9%
P2	5.8%	5.6%
P3	2.8%	2.8%

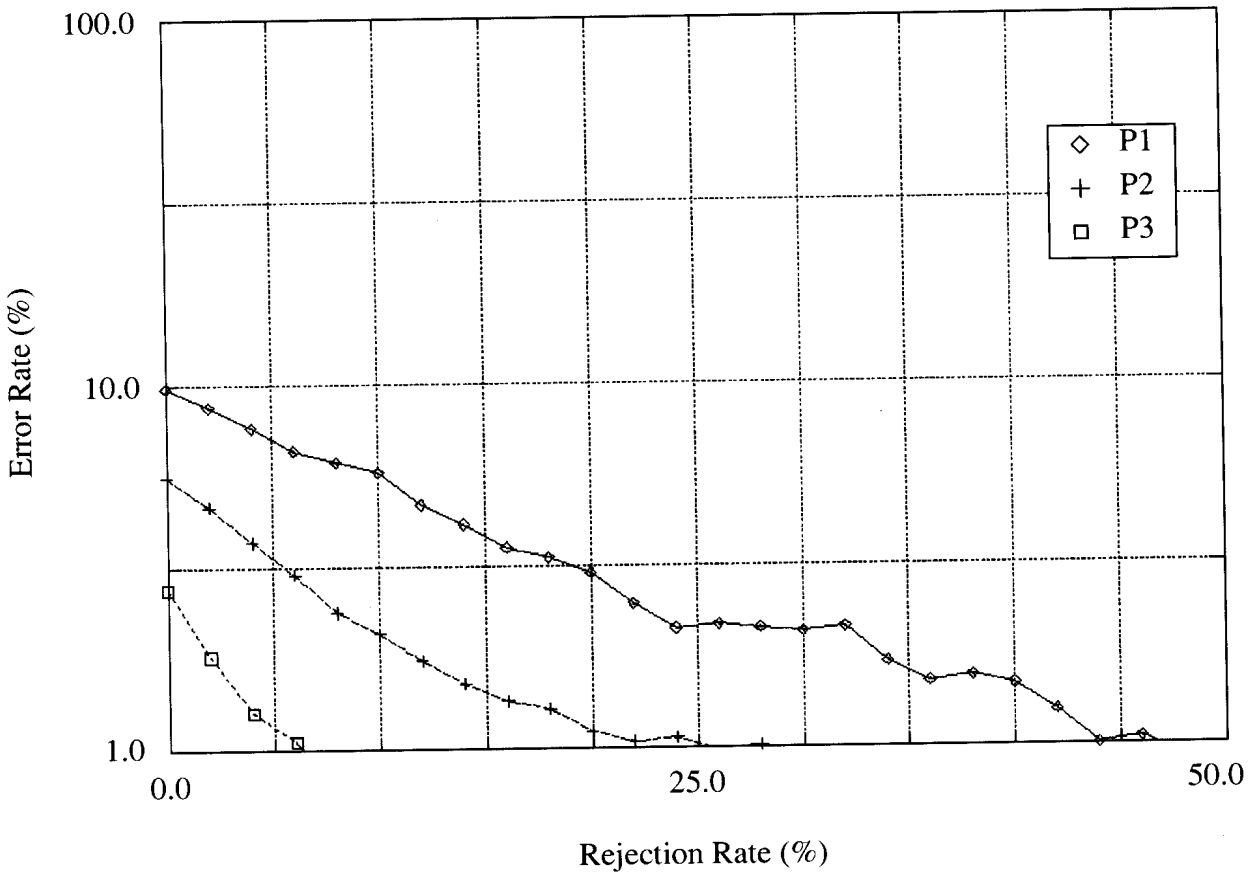


Figure 4