The Second Census Optical Character Recognition Systems Conference

Jon Geist, R. Allen Wilkinson, Stanley Janet, Patrick J. Grother
Bob Hammond, Norman W. Larsen, Randy M. Klear, Mark J. Matsko,
Christopher J. C. Burges, Robert Creecy, Jonathan J. Hull,
Thomas P. Vogl, and Charles L. Wilson

Contents

1 Executive Summary

Bob Hammond and Jon Geist

1.1 Background

Since 1790, the United States has conducted a decennial census, or head count, of the American population. Over the last century, growth in the population and demand for quicker tabulations have presented very strenuous tasks for data capture and information technology. In the late 1800's, tabulating machines with punched cards were invented for Census use. In the 1950's, staff at Census and NBS helped develop the UNIVAC for general purpose computing. About the same time, they jointly developed the first optical scanning device for high speed mark recognition from microfilm. For over 40 years, this scanning technology has worked well for multiple-choice answers; however, the census still requires an enormous amount of paper handling and labor-intensive data entry operations to capture handwritten responses. Increasing workloads, rising labor costs and shrinking budgets have prompted this research into optical character recognition (OCR).

1.2 The Conferences

After the 1990 Census, NIST and Census sponsored a scientific experiment and set of conferences (hereafter referred to as the Conferences) to determine the state of the art in the optical character recognition industry. To organize each Conference, NIST and Census formed a Committee having representatives from government, industry, and academia, and NIST personnel ran the Conference. Twenty nine different groups from North America and Europe responded to the call for participation in the first Conference. Each party received an image database of isolated (segmented), handprinted, alpha and numeric characters for training their systems. Later, each party received a similar database for test purposes. Each attempted to recognize the characters, and all but three submitted their results to NIST for scoring. In late May 1992, all parties that submitted results convened in Gaithersburg, Maryland to discuss the results. Scientific and academic participation was encouraged, and marketing interests were discouraged. Attendance was strictly limited to sponsors, participants, and associates designated by each participant, along with a few observers from federal agencies (FBI, IRS, USPS) that are currently sponsoring work in the field.

The first Conference and related exercises focused on a single step in the process: machine recognition of individual (or segmented) characters with no context. With the single variable nature of this study, no valid comparisons can be made regarding cost or performance of systems designed to process entire forms or documents. Further, the efforts of participants were not proctored or monitored in any way by Census or NIST staff. Nor were any attempts made to distinguish between results obtained from experimental systems and those obtained from commercial systems.

The second Conference, whose results are described in this report, focused on a much more

realistic recognition task: reading answers from a digital image of forms scanned from paper and from microfilm. Lessons from the first conference also led to improvements in the design and preparation of materials for the second conference. The sample included handwriting from a much larger number of different writers and the training materials included several dictionaries to allow dictionary-based correction of OCR results.

The second Conference required a much more comprehensive OCR capability, and constituted a much more difficult OCR task than that of the first Conference. Otherwise, the second Conference was organized similarly to the first. Again, there were no efforts to proctor or monitor participation, nor any attempt to distinguish between results obtained from experimental systems and those obtained from commercial systems.

Twenty five different groups responded to the call to participate in the second Conference. Each party received training materials on two CD-ROMs. Some participants dropped out before receiving the test CD-ROM. Overall, ten groups submitted test results (eight were on time and two were late). All ten groups attended the meeting associated with the second Conference in mid February, 1994.

To establish a baseline for comparison, an independent set of reference data was created for a subsample of this test. This allowed scoring of the 1990 Census key entry operations as an eleventh participant in the test (at the zero percent rejection level). At rejection levels between 40 and 60%, several systems achieved accuracy levels that exceeded the human performance levels at the 0% rejection level (see Chapter 2, Section 5, Summary of Results).

1.3 General Conclusions

These Conferences demonstrate that the accuracy of optical character recognition systems for handwriting has improved dramatically over the last few years. Machine performance in reading words and phrases may now be good enough to decrease the cost and time needed to carry out a Census without decreasing the accuracy of the results. Improved techniques to separate text strings into individual characters (or pieces of characters for later reconstruction) and various algorithms to check spelling and context have contributed to these improvements (see Chapter 2, Section 3, OCR Methods Used).

It should be noted that some applications of OCR are easier than reading the Industry and Occupation answers for the Census. For example, reading and reconciling the legal and courtesy amounts on checks may be easier because it would require smaller dictionaries. Other applications that include many numeric fields might benefit from the use of checksums, and other techniques. Therefore, the results of this test suggest that OCR may already be good enough for these applications. Furthermore, this general conclusion is supported indirectly by the increase of commercial products and services that claim to perform OCR on handwritten text.

Of course, there are still a number of questions that were not addressed by these Conferences that remain to be answered. Also, more sophisticated, application-specific tests along with valid cost and benefit analyses are needed to answer the ultimate question of cost effectiveness. Fortunately, this conference has helped to frame most of the questions that need future

research and development:

1) What throughput rates can be achieved by OCR systems on various hardware platforms? The image processing associated with these systems is computationally intensive and the storage requirements for digital images are very large. Overall performance benefits and/or cost savings over traditional methods must be realized in order to justify the capital expense of sophisticated image processing systems.

2) What design attributes of forms will facilitate optimal performance of image processing systems while at the same time making the form easy for respondents to complete with minimal instruction and/or annoyance?

3) How much additional improvement can be achieved by constructing better language models and lexicons, and the techniques to employ them?

4) How much additional improvement can be achieved by using multiple recognition subsystems and constructing a voting algorithm of some kind? Initial tests show that using multiple systems to vote on the recognition might improve results, but the performance/cost of this concept is unknown.

5) Can other holistic intelligence be designed into future OCR systems? For example, could high confidence recognition of one answer help improve the recognition of answers to other questions on the same form (completed by the same writer)? Several participants suggested this possibility, but no one has implemented such techniques; thus, the performance/cost is also unknown.

6) There are a number of open questions about how to score OCR accuracy for any specific application (see Chapter 6). The answers to these questions depend heavily on the intended use of the outputs from the recognition system. Some applications are more tolerant of certain types of errors than other applications. The designers of each application must discover the most effective way to set rejection levels for their intended use.

7) Finally, what are the most efficient techniques to complement OCR systems with human correction of the handwritten answers rejected by the OCR system?

NIST made every effort to assure the accuracy of the measures computed from the submissions by the participants. Nevertheless, NIST and Census are aware that different tests, which may be more pertinent to real applications, might give different results than those reported here, and that other analyses of the submissions might give more complete results than those reported here.

Neither NIST nor Census are in any way responsible for how the results presented in this report may be used. While some results from this Conference may appear in marketing literature, under no circumstances should potential buyers use data from this study as a primary basis for purchasing decisions. Census and NIST can make only one recommendation to potential buyers: use your own application-specific data to thoroughly test the performance of any system (or component) in a realistic setting.

2 Introduction

Jon Geist

The goals of the First and Second Census Optical Character Recognition (OCR) Systems Conferences were scientific in nature. The first goal was to gauge the state of the art of OCR of handprinted characters with respect to the particular problems associated with entering census data into a computer database. The second was to learn what is currently limiting the state of the art. The third goal was to determine whether new databases of handprinted characters for use either in training or in testing could be expected to help to improve the state of the art of OCR for applications such as the census, and if so, what types of new databases are needed.

Neither the First nor the Second Conference had any marketing goals. In particular, the tests were not proctored, and no attempt was made to distinguish results obtained with commercial systems from those obtained with research systems. Also, participants were implicitly encouraged to carry out experiments that promoted the scientific goals of the Conference, even though they might not contribute to optimum system performance.

Neither Conference was designed to produce results that could be used as the basis for purchasing an OCR system. Anyone who does base a purchase on these results will probably encounter a number of serious problems. Decisions regarding the application of an OCR system to some specific task should be based on the results of proctored tests with test materials that are representative of that task.

On the other hand, the methodologies developed for these Conferences and the results obtained should prove quite useful in designing tests, both large and small, to support purchasing decisions. Furthermore, it was hoped that preparation for (including new databases) and participation in these Conferences would help to advance the state of the OCR art.

The full Census OCR task consists of document handling, document scanning, form identification, field isolation, character segmentation, character recognition, and context-based field correction. On the other hand, the recognition of properly segmented, isolated characters has been the bellwether of handprint OCR progress for some time. Therefore, the tests associated with the First Conference were limited to this task, and tests that were more typical of the full Census OCR task were postponed for future conferences.

It was decided that tests open to organizations having strong OCR programs would be cost-efficient tools for meeting the goals mentioned above. This would allow comparison of the results from a wide variety of systems employing different algorithms for the different OCR subtasks. Of course, it is not possible to control the variables as well as might otherwise be desirable with this type of experiment, but comparison of the results from a broad range of systems was thought to be more important than comparison of the results obtained from different variations of a single type of system.

The activities of the First Conference [?] were carried out from February through May of 1992. The tests consisted of classifying about 85,000 binary images of properly segmented, isolated characters (roughly 60,000 digits, 12,000 upper case, and 12,000 lower case letters)

that were distributed on a CD-ROM. All participants received identical tests, and none had seen any of the images on the CD before receiving it.

An important conclusion of the First Conference was that the OCR of isolated (properly segmented) characters was essentially a solved problem. Chapter 3 of this report presents the results of follow-up studies that strongly support this conclusion.

When planning the Second Conference at the end of the First Conference, there was an overwhelming consensus among the participants about most issues. First they wanted the training data to be more representative of the test data than in the First Conference. Second, they wanted data from many more writers than in the First Conference. Third, when given a choice between digital images scanned from forms designed to test segmentation accuracy, and digital images of non-sensitive answers scanned from microfilm copies of 1990 Census returns, they overwhelmingly chose the latter.

2.1 Organization of the Second Conference

The Second Conference was organized by a Committee consisting of the following individuals:

Bob Hammond, Norman W. Larsen, Randy M. Klear, Mark J. Matsko, and Robert Creecy: US Bureau of the Census

R. Allen Wilkinson, Stanley Janet, Charles L. Wilson and Jon Geist: National Institute of Standards and Technology

Dr. Jonathan J. Hull: Center of Excellence for Document Analysis And Recognition

Dr. Thomas P. Vogl: Environmental Research Institute of Michigan

Dr. Christopher J. C. Burges: AT&T Bell Laboratories

Jon Geist, the Committee Chairman, handled the planning of the Conference and the majority of the interaction with the participants. The Conference was run for the Committee by the Image Recognition Group (IRG) at the National Institute of Standards and Technology (NIST) under contract to the US Bureau of the Census. Bob Hammond administered the contract supporting this Conference and coordinated the Census Bureau work in support of the Conference.

The following individuals from the NIST IRG were instrumental in carrying out the NIST portion of the work of the Conference. Charles Wilson, the Leader of the NIST IRG, assured that resources were available when needed. Allen Wilkinson coordinated the technical activities of the Conference including the preparation of training and test materials, the receipt of participant submissions, scoring submissions, and software trouble shooting until accepting a new position at NIST. Stanley Janet modified the NIST scoring package to accumulate the measures chosen by the Committee for the Conference, and assumed Mr. Wilkinson's duties after the latter's departure. Patrick Grother provided valuable comments on various aspects of the Conference based on his role as a participant representing the NIST IRG OCR system. Mike Garris carefully reviewed this report before publication.

The following individuals from Census were instrumental in carrying out the Census portion

of work for the Second Conference. Stan Matchett and Bob Bair assured that resources were available when needed. Bob Hammond coordinated the technical activities at Census. Norm Larsen and John Rotegard designed and implemented the scanning systems, prepared and maintained the software for scanning operations, and provided technical analysis and trouble shooting for the image capture activities. Brian Washington operated and maintained the scanning equipment and performed a variety of general and special purpose tasks. Neal Bross installed and maintained the network and provided overall UNIX system administration and support. Dan Gillman developed software to extract the reference data from large archived data files and provided advice about the extracts and the related automated coding system. Randy Klear developed the key entry system and coordinated the production keying of reference data for the paper sample. About 30 volunteers (and/or conscripts) from 10 different divisions performed the independent keying operations.

The activities of the Second Conference started in January of 1993, with attempts to create a large sample of digital miniforms scanned from microfilm copies of a non-sensitive portion of the Industry and Occupation section of the 1990 Census Long Forms. Each miniform consisted of three answer boxes and the questions surrounding them, and covered an area of about 75 mm by 95 mm on the 1990 Long Form. Reference data for each miniform was obtained from the hand-keyed results of the 1990 Census. All of the reference data was screened by two independent methods to remove potentially sensitive information that, while not requested, was sometimes provided in answer to these questions. Finally, matching files of digital images and the ASCII transcriptions (references) of the answers written on them were prepared in a format chosen for the test.

During the Conference period, it became clear that the digital images obtained from microfilm had far inferior image quality to those that could be obtained by scanning the original paper forms. Therefore, the scope of the Conference was broadened to include a test with images scanned from forms that had been reserved from the 1990 Census, as well as a test with the images scanned from microfilm. Figures 1 and 2 at the end of this chapter compare a miniform scanned from microfilm with a miniform scanned from paper. Note that Fig. 1 is well above the average image quality for the miniforms scanned from microfilm, while Fig. 2 is of average image quality for the miniforms scanned from paper.

Later, it was decided to score the hand-keyed results from the 1990 Census against independently keyed reference data. Therefore, the answers on the paper forms were independently keyed twice, and rekeyed a third time when differences occurred. These references were used to score the participant submissions for the test scanned from paper. They were also used to score the answers keyed during the 1990 Census. This allowed fair comparison of the Conference submissions with human performance on the same task. The answers keyed during the 1990 Census were used as the reference data for the test scanned from microfilm.

By the end of June, which was well behind the original schedule, enough progress had been made to warrant issuing a Call for Participation on behalf of the Committee. A version of the Call is reproduced in Appendix A.

Twenty five organizations agreed to participate in the Conference. The first and second sets of training materials were shipped to the participants at the end of August and the beginning

PARTICIPATING ORGANIZATION	STATUS	SYSTEM	FROM PAPER	FROM μFILM
Adaptive Solutions, Inc. Beaverton, OR	NO SUBMISSION			
Aston University Birmingham, UK	NO SUBMISSION			
AT&T Bell Laboratories Holmdel, NJ	LATE LATE	ATT_0 ATT_1	X X	
CEDAR, SUNY Bufallo, NY	ON TIME LATE LATE	CEDAR_0 CEDAR_1 CEDAR_2	X X X	
CGA Gentilly Cedex, France	WITHDREW AFTER SDB13			
CGK mbH Konstanz, Germany	ON TIME LATE	CGK_0 CGK_2	X X	X X
Com Com Systems, Inc. Clearwater, FL	NO SUBMISSION			
Environmental Research Institute of Michigan Ann Arbor, MI	ON TIME ON TIME	ERIM_0 ERIM_1	X X	X X
Gamma Research, Inc. Huntsville, AL	WITHDREW BEFORE SDB13			
GTESS Corporation Richardson, TX	WITHDREW BEFORE SDB13			
Hughes Aircraft Company Reston, VA	ON TIME LATE LATE	HUGHES_0 HUGHES_1 HUGHES_2	X X X	
IBM Almaden Research Center, San Jose, CA	ON TIME ON TIME ON TIME ON TIME LATE	IBM_9 IBM_0 IBM_1 IBM_2 IBM_3	X X X X X	X X X X X

Table 1: Participating organizations, status, system names, and tests.

PARTICIPATING ORGANIZATION	STATUS	SYSTEM	FROM PAPER	FROM μFILM
IDIAP	ON TIME	IDIAP_0	X	
Martigny, Valais	LATE	IDIAP_1	X	
Switzerland	LATE	IDIAP_2	X	
	LATE	IDIAP_3	X	
INM, Inc.	NO SUBMISSION			
Waterloo, Ontario				
Intrafed, Inc.	NO SUBMISSION			
Bethesda, MD				
MCC	LATE	MCC_0	X	
Austin, TX	LATE	MCC_1	X	
Mimitecs SA	NO SUBMISSION			
Chatenay-Malabry				
Cedex, France				
Mitek, Inc.	WITHDREW			
San Diego, CA	BEFORE SDB13			
Nestor, Inc.	WITHDREW			
Providence, RI	BEFORE SDB13			
National Institute of	ON TIME	NIST_9	X	X
Standards and Tech.	ON TIME	NIST_0	X	X
Gaithersburg, MD	LATE	NIST_1	X	
	LATE	NIST_2	X	
	LATE	NIST_3	X	
RAF, Inc.	WITHDREW			
Redmond, WA	BEFORE SDB13			
Symbus Technology	WITHDREW			
Waltham, MA	BEFORE SDB13			
University of Bologna	ON TIME	UBOL_9	X	X
Bologna, Italy	ON TIME	UBOL_0	X	X
	LATE	UBOL_1	X	
U. of Florida	NO SUBMISSION			
Gainsville, FL				
U. of Michigan	NO SUBMISSION			
Dearborn, MI				

Table 2: Participating organizations, status, system names, and tests.

of October, 1993, respectively. The test materials were shipped to the participants by express carrier to arrive on December 1, 1993. The OCR results returned to NIST for scoring were to be received by the participant's express carrier by December 15, 1993 in order to be ON TIME. However, LATE results were accepted provided that an express carrier received them by January 31, 1994.

The test was very hard, and many participating organizations either withdrew from participation or did not submit results for scoring before the Conference deadlines. Only participants who submitted plausible entries before January 31, 1994 were permitted to attend the meeting run in connection with the Conference. Note that participation without attendance at the meeting could still be considered beneficial to the participants because it gave them early access to the databases prepared for the Conference. All participants, the names assigned by NIST to the systems for which they submitted results, and the type of results submitted are summarized in Tables 1 and 2. Note that some participants submitted results only for the test scanned from paper, while others submitted results for both tests.

2.2 Training and Test Materials

Chapter 4 describes the selection and scanning of both the microfilm and the paper samples of the industry and occupation answers from the 1990 Census that were used in the Conference.

The NIST multiple image set (mis) file format [?] was used for storing the miniform images after they were extracted from the images scanned from microfilm or paper, and complementary file formats were used for the reference, hypothesis, and confidence files used in conjunction with the training and test images.[?] Each mis file contained digital images of five miniforms, each of which had three answer fields. Each reference file had 15 lines of ASCII transcriptions of the answers entered into the answer fields on the five miniforms in the corresponding mis file. The hypothesis files produced by the participant's OCR systems had exactly the same format as the reference files, but contained the hypothetical answers produced by the OCR systems. Each confidence file had 15 lines of the ASCII representations of numbers ranging from 0.0 through 1.0 to convey the relative reliability of the corresponding hypotheses in the associated hypothesis file. More information about the file formats used to distribute the training and test materials and to return the test results can be found in Appendices A and B.

The training and test materials were distributed on CD-ROM. There were two training CD-ROMs and one test CD-ROM. The first training CD-ROM, Special Database (SD 11), had images scanned from microfilm, the associated references, some image manipulation software, and the first dictionaries for optional use in correcting the OCR results. These served to acquaint the participants with the data formats, and had the side effect of scaring some participants into withdrawing from the Conference. The second training CD-ROM (SD 12) had new microfilm-scanned images plus images from paper, the associated references, the image manipulation software, and dictionaries augmented by the addition of any words or phrases appearing in the images on SD 11. The test CD-ROM (SD 13) had everything that SD 12 had except the references. The participants returned hypothesis files and confidence files to NIST for scoring on floppy disks, using the same directory structure used on the test

CD-ROM.

SD 11 contained 25 subdirectories having 100 mis files in each directory and 5 miniforms per file for a total of 12,500 miniforms and 37,500 answer fields. SD 12 contained images from microfilm and images from paper. Its microfilm and its paper directories each had 12 subdirectories having 100 mis files in each directory and 5 miniforms per file for a total of 6000 miniforms and 18,000 answer fields for each test (microfilm and paper). SD 13 also contained images from microfilm and images from paper. Its microfilm and paper directories each had 6 subdirectories having 100 mis files in each directory and 5 miniforms per file for a total of 3000 miniforms and 9,000 answer fields for each test.

The number of miniforms on the CD-ROM databases was decreased during the course of the Conference because it became apparent that many participants would not be able to process the larger data sets in the two weeks allowed for ON TIME test submissions. Instructions for the test phase of the Conference were sent with SD 13. One version of these is reproduced in Appendix B.

A number of different dictionaries for optional use with algorithms that correct the results of raw OCR were created from a 132,000 sample of answers to the Industry and Occupation questions obtained from the 1980 Census, and augmented with the references associated with the training images. Dictionary creation and format is discussed in Chapter 5.

In summary, the test required each participant to convert the letters, digits, and spaces between words in the answers on the images of the miniforms into their ASCII representations, retaining their order. Note that the participants were instructed not to have their OCR systems correct any misspellings that were detected. This same instruction was given to the key entry personnel during the 1990 Census. It resulted in the inclusion of a large fraction of misspellings as well as abbreviations in the dictionaries. This makes the dictionaries much larger than they would otherwise be. It is possible that dictionaries containing only correct spelling and abbreviations would have improved the results of the Conference, but neither NIST nor Census knew how to remove the misspellings without also removing the abbreviations from the dictionaries in the time available. Removal of abbreviations would probably have hurt the test results. This issue is discussed in more detail in Chapter 5.

The original plans were to carry out the test phase of the Second Conference during September of 1993, but it soon became clear that neither Census nor NIST could complete their role in support of the Conference soon enough for this time scale, particularly after the scope of the Conference was extended to include images scanned from paper and the generation of reference answers independent of the answers collected during the 1990 Census. Also, most, but not all, participants indicated a desire to see the time frame for the Conference extended. Thus the test phase was postponed until December 1 of 1993, with all other schedules slipping proportionally. The actual meeting where the results were discussed was held on February 15 and 16 of 1994.

2.3 OCR Methods Used

A wide variety of preprocessing, feature extraction, and classification algorithms were employed by the OCR systems used for the recognition of isolated characters in the First Conference. The overall task of the Second Conference was far more complex. Nevertheless, it appears that all of the systems used in the Second Conference can be roughly described in terms of the following subtasks:

1) FORM IDENTIFICATION: This subtask consists of identifying certain expected features on each miniform image presented for recognition. The output of this subtask is either the rejection of the miniform form as unrecognizable or a set of locations of key features that identified the miniform as acceptable for further processing. (Two slightly different miniforms were used in the Second Conference test. Figure 1 is one type and Fig. 2 is the other type.)

2) FIELD ISOLATION: This subtask consists of extracting the text image of each answer from the form. The output of this subtask consists of one or more images of text minus the surrounding portions of the form. (There were three answer fields, each marked by a dashed rectangle, on each miniform. There were no guidelines for participants about what to do with parts of answers written outside the boxes.)

3) LINE ISOLATION: This subtask consists of extracting images of single lines of text from each answer field. The output of this subtask is one or more images of a single line of text. (Some answers were written on more than one line in each answer field.)

4) SEGMENTATION: This subtask consists of breaking each image of a line of text into smaller units for recognition. The output of this subtask is one or more image segments. Each segment is either the image of an isolated character, an image of an isolated piece of a character, or the image of an isolated group of connected or otherwise undersegmented characters. (Images of characters that are broken into more than one segment are oversegmented and images of characters that are grouped together are undersegmented.)

5) RECOMBINING SEGMENTS: This subtask consists of selecting various combinations of segments (including single segments) as plausible candidates for isolated character images. The output of this subtask is one or more isolated character-image candidates. (Some systems purposely oversegment enough to assure that no undersegmentation occurs. In this case, it is necessary to recombine segments in different ways to be sure that all isolated character images occur among the different combinations.)

6) RECOGNITION: This subtask consists of one of two slightly different functions depending upon the OCR system. The more general consists of assigning relative confidences to all of the allowed classes for each character-image candidate; the less general consists of assigning a single character class to each character-image candidate. The output of this subtask is either a single class (possibly with a confidence) or a set of ordered pairs consisting of character class and associated confidence. (In either case, this output is called raw OCR to emphasize that it has been generated without the help of any context other than that existing in the isolated character-image candidates.)

7) ORGANIZING CHARACTER CANDIDATES: This subtask consists of organizing the

output of subtasks 5 and 6 (sometimes just 6) into a form useful for the dictionary input stage. The output of this task is just the output of those task in a format suitable for the particular dictionary look-up method being used. (Examples from the Conference include various combinations and modifications of well known techniques such as hidden Markov, Viterbi, and Levenstein-distance algorithms.[?]).

8) DICTIONARY-BASED CORRECTION: This subtask consists of selecting the dictionary entries that best match the properly organized character-image candidates according to some set of criteria. The output of this subtask is the hypothetical answer provided by the OCR system as its final result and (usually) a confidence for the field. (Different systems used different dictionaries or combinations of dictionaries.)

In order for the above set of subtasks to properly describe all of the Conference systems, it is necessary to consider some of the subtasks to be empty tasks for some of the systems, that is, to return their input as output. For instance, some systems did not carry out segmentation, but attempted to recognize each character or stroke without isolating it from the other characters or strokes in the same line. For those systems, subtasks 4) and 5) were empty tasks. It is also necessary to allow some of the subtasks to be carried out simultaneously with other subtasks or in iterative loops containing one or more subtasks, and to allow for decision points and alternative paths through the subtasks.

Notice that it was only subtask 6) that was tested in the First Conference. Recall that this subtask has been the bellwether of OCR progress for some time. For the First Conference, subtask 6) was divided into three smaller subtasks: preprocessing, feature extraction, and classification. No new preprocessing or classification techniques were reported as being used for the Second Conference, but one new feature extraction process similar to computer tomography was reported by one participant. It is illustrated for the letter C in the System Summary for CGK in Appendix C.

Each participant was requested to fill out the questionnaire shown as Enclosure 6 in Appendix A about the algorithms used in his or her OCR system. It turned out that the questionnaire was very poor at extracting the key ideas about the different systems. Therefore, the questionnaires returned by the participants are not included anywhere in this report. On the other hand, many participants presented quite detailed descriptions of the systems they used at the Conference meeting, so their viewgraphs are reproduced in either Appendix C or Appendix D along with graphs of their test results. Appendix C contains the ON TIME results, Appendix D the LATE results.

It is well beyond the scope of this report to compare and discuss the details of the algorithms used for the different subtasks by the different systems. However, some comments and general conclusions seem warranted. Most systems were empty for one or more of subtasks 1), 4), 5) and 7). It appears that making 1) empty caused little or no error for this test. On the other hand, both of the lowest error systems carried out non-empty versions of subtasks 4), 5), and 7).

None of the systems using empty (no) segmentation were among the most accurate. This suggests that segmentation is an important subtask for this type of test, at least at the current state of the art. All systems that attempted segmentation except NIST's used intentional

oversegmentation as a means of avoiding undersegmentation, and the best performing systems had sophisticated means for recombining segments prior to dictionary-based correction. This suggests that segmentation is the most challenging subtask for this type of test, and that intentional oversegmentation followed by sophisticated recombination methods, possibly in more than one of the downstream subtasks, is the best solution to the segmentation problem at the current state of the art.

Finally, there is rather fundamental trade-off between dictionary coverage (size) over the set of test phrases, on one hand, and confusion among the dictionary entries, on the other. This limits the accuracy that can be achieved with purely dictionary-based methods. Use of language models, particularly more sophisticated models based on a much larger set of training-reference phrases should allow this problem to be solved. More details about this problem are presented in Chapter 5.

2.4 Summary of Results

Classification, recognition, hypothesis, reference, rejection, and confidence are general ideas of importance in the OCR of words and phrases. The precise definitions of these terms as used in this report are given in Chapter 6. Two different measures of classification error were calculated for this Conference: the field error rate and the field distance rate. These too are defined in Chapter 6.

Notice that the field error rate does not distinguish among different ways that fields can be incorrect. For instance, the incorrect hypothesis CARRIES BAGS for the reference DRIVES THE TRUCK makes the same contribution to the field error rate as the incorrect hypothesis DRIVES THE TRUCKS even though the first hypothesis is completely wrong and the second almost right. In many applications (such as the generation of Industry and Occupation codes from Census Long Forms), the error in the second hypothesis will have no effect on the final use of the ASCII version of the answer. The field distance rate is much less sensitive to this type of problem, but it does not have a unique definition, as discussed in Chapter 6.

The introduction of two different OCR accuracy measures raises the question of which should be used in any given application or how to use both. The short answer is that this remains an open question.

There are a number of unresolved issues associated with scoring field hypotheses in contrast to the scoring of isolated character hypotheses, which is relatively straightforward. These include insensitivities to important properties of word and phrase alignments in the algorithms used to align references and hypotheses before scoring, as well as alternate error measures. As an example, suppose that two systems produce the same field distance at a given rejection rate, but that the first produces a much smaller field error rate than the second. Intuition says that the first should be the superior system, but in some applications the second may actually produce more useful results. This and related issues are also addressed in more detail in Chapter 6.

Tables 3 and 4 list the field error and field distance rates at rejection rates of 60%, 50%, 40%, and 0% for both the paper and the microfilm tests for all of the results returned to

Scanned	Field Rejection Rate							
From Paper	60%		50%		40%		0%	
	Percent Classification Field Error And Distance Rates							
On Time System	error rate	distance rate	error rate	distance rate	error rate	distance rate	error rate	distance rate
CEDAR_0	37.6	20.5	38.7	20.4	41.5	21.3	58.7	37.3
CGK_0	13.8	4.7	19.6	6.2	26.3	9.0	50.5	24.6
ERIM_0	3.6	0.8	6.3	1.6	12.2	4.3	39.7	18.7
ERIM_1	3.9	1.0	7.5	2.4	14.1	5.4	41.9	20.8
HUGHES_0	61.6	26.0	69.2*	38.7*	74.3*	47.5*	84.6	63.4
IBM_0	49.6	24.3	56.9	28.8	62.4	32.9	75.0	44.8
IBM_1	53.6	24.7	60.3	29.4	65.0	33.0	76.8	44.8
IBM_2	86.6	58.6	88.4	58.8	89.9	59.4	93.1	63.4
IDIAP_0	10.7	2.6	16.8	5.2	25.8	10.5	52.6	33.4
NIST_0	46.6	16.2	53.8	21.7	60.1	27.2	75.3	46.2
UBOL_0	57.1	36.0	61.8	38.6	64.9	39.9	71.7	43.1
KEY_90	NA	NA	NA	NA	NA	NA	8.5	1.6

Table 3: Field error and distance rates at certain field rejection rates for the classifications that were carried out on the images scanned from paper and submitted to NIST on time for scoring, including the results for the 1990 Census key entry operation. *These results contain hypotheses with 0.0 confidence.

Scanned	Field Rejection Rate							
From μfilm	60%		50%		40%		0%	
	Percent Classification Field Error And Distance Rates							
On Time System	error rate	distance rate	error rate	distance rate	error rate	distance rate	error rate	distance rate
CGK_0	23.4	7.1	31.6	11.3	38.7	15.5	60.7	32.6
ERIM_0	9.7	5.0	16.2	7.5	24.6	11.8	50.0	25.7
ERIM_1	10.1	5.2	16.7	8.0	25.4	12.4	50.9	26.7
IBM_0	66.8	36.6	71.6	40.6	75.1	43.7	83.7	53.3
IBM_1	69.5	36.7	73.8	40.4	77.2	43.9	85.1	53.4
IBM_2	91.3	65.1	92.6	65.3	93.3	64.9	95.6	66.9
NIST_0	77.3	42.3	81.4	48.2	84.3	52.4	90.4	62.6
UBOL_0	70.8	47.6	73.6	48.5	77.1	52.4	82.0	55.4

Table 4: Field error and distance rates at certain field rejection rates for the classifications that were carried out on the images scanned from microfilm and submitted to NIST on time for scoring.

NIST for scoring by December 15. Figures 3 and 4 at the end of this Chapter plot this data over the entire range of the rejection rate. As mentioned above, the individual results for each organization can be found in Appendices C and D. The latter Appendix also explains why some of the curves turn up with increasing rejection rate.

Table 3 also gives the field distance and error rates at 0% rejection rate for the 1990 Census key entry operation (KEY_90) for comparison with the machine results for the paper test.[?] Since the human key entry operators did not produce confidences for their entries, no data is available for the greater rejection rates.

The best machine system in the Conference does not reach the 8.5% field error rate achieved by the key operators at 0% rejection rate until over 45% of the fields have been rejected. Similarly, the best machine system in the Conference does not reach the 1.6% field distance rate achieved by the key operators at 0% rejection rate until 50% of the fields have been rejected. Clearly, machines cannot yet read handprint phrases as well as people can even though Chapter 3 presents evidence that machines can classify isolated characters about as well as humans can, at least in economically significant applications. This is not surprising. People do not read by first isolating the characters in a word and then classifying them, but by reading whole words and phrases using many different types of context.

Note, however, that it is not necessary for OCR systems to read handprint words and phrases as well as people before they can be economically viable for use in a Census. What is necessary is a combination of an OCR system and human keyers that provides no greater error and costs less than doing the whole job with human keyers. This might be achieved by having the OCR system classify all of the Census forms and reject a fraction of the forms as having unreliable hypotheses. The rejected forms could then be turned over to humans for keying.

Figures 3 and 4 at the end of this Chapter and Table 3 show that the field error rates for two ON TIME systems fall below 8.5% at rejection rates between 40% and 50%, while the field distance rates for two ON TIME systems fall below 1.6% at rejection rates between 50% and 55%. This is the reason that Tables 3 and 4 list results at 60%, 50%, and 40% rejection rate, as well as at 0% rejection rate.

A hybrid system was made from the ERIM_0 and 1990 Census results. The 60% of the ERIM_0 hypotheses that had the lowest confidences were replaced by the corresponding 1990 Census hypotheses. This simulates a system where a machine classifies all of the Census fileds but humans only classify the 60% of them rejected by the machine. To two decimal place accuracy, the resulting field distance and error rates were 1.55% and 8.34%, respectively, while the KEY_90 field distance and error rates were 1.56% and 8.58%, respectively.[?] However, the hybrid system got 96 fields incorrect that KEY_90 got correct, and KEY_90 got 116 fields incorrect that the hybrid system got correct. This highlights the possibility that the errors made by the hybrid system could have a more or less detrimental impact on the Census application than those made by the 1990 Census key operators. In fact, this can be the case even for the fields that both systems got incorrect since they need not get them incorrect in the same way.

All but one of the participants providing ON TIME submissions also provided LATE sub-

missions. Two participants provided only LATE submissions. The system summaries for the LATE submissions and the viewgraphs for the two systems providing only LATE submissions are contained in Appendix D.

Most of the LATE submissions were significant improvements over the ON TIME systems from the same participant. In fact, there were considerably more late submissions than ON TIME submissions. However, the significant improvements achieved by the LATE submissions do not affect the overall conclusions of the Conference. On the other hand, the fact that significant improvements to system performance were still being achieved right up to the end of the Conference suggest that further improvements will be obtained in the future. For this reason, it was decided to allow the participants (and anyone else) purchasing the test data base to submit their results to a NIST anonymous ftp site for scoring. The only purpose of this is to assure uniform scoring and to keep NIST apprised of the state of the art. NIST has reserved enough materials to run a second test if the state of the art improves enough to warrant it. More details are given in Chapter 8.

The fact that for rejection rates greater than 40% some OCR systems had lower error and distance rates than humans at 0% rejection rate does not mean that OCR systems are now good enough to be used in a Census, but it does mean that they might be. Only more sophisticated, application-specific tests that have not yet been developed will be able to resolve this question. To emphasize this point, consider the following specific question which was brought up earlier in this report. Ignoring the rejection rates, which is better, the combination of a field distance rate of 1.6% with a field error rate of 8.5% that was achieved by the human classifiers or the combination of a field distance rate of 1.6% with a field error rate of 6.3% that was achieved by the ERIM system at 50% rejection rate? There is no way to answer this question without further work on scoring.

It has often been suggested that a consensus OCR system formed from the results of two or more OCR systems might out-perform the systems from which it was made. Chapter 7 presents the results of a small and preliminary study using some systems in the Conference in simple voting schemes that requires no training. Significant improvements were found at some rejection rates, but decreased performance was found at the most economically significant rejection rates. It is possible that more sophisticated voting schemes that require training to renormalize the confidences produced by the individual OCR systems may produce improved results, but further work is necessary to investigate this possibility.

2.5 Major Conclusions

The major conclusions of the first two OCR Systems Conferences are listed below. The justifications for these conclusions are contained in the Chapters referred to above.

1) Humans are no longer clearly better than machines at the recognition of isolated characters, and certainly not in any economically significant way. (See Chapter 3.)

2) Humans are still much better than machines at reading words and phrases drawn from a large sample of possible words and phrases. (See this Chapter.)

3) Despite the above conclusion, machine performance in reading words and phrases may

now be good enough to decrease the cost and time needed to carry out a Census without decreasing the accuracy of the results. Only more sophisticated application-specific tests can answer this question. (See this Chapter.)

4) It should be recognized that there are a number of applications of the OCR of words and phrases, such as reading and reconciling the legal and courtesy amounts on checks, that are much less difficult than reading the Industry and Occupation answers for the Census because they use smaller dictionaries. Therefore, the results of this test suggest that OCR may already be good enough for these applications.

5) There are a number of open questions about how to score OCR output for any specific application. (See Chapter 6.)

6) Segmentation was probably the most difficult subtask required for the Second Conference tests. (See this Chapter.)

7) Intentional oversegmentation (more precisely, intentional avoidance of undersegmentation) at the character level seems to be the best solution to the segmentation problem, at least at the current state of the art. (See this Chapter.)

8) As dictionary size is increased, coverage over some domain can also be increased. At the same time, the average similarity among the entries in the dictionary will also increase. If only a small fraction of the words or phrases added to the dictionary actually contribute to increased coverage over some test domain, the increase in field error due to increased confusion may be larger than the decrease in field error due to increased coverage. A similar phenomena may also occur with the field distance. (See Chapter 5.)

9) The trade-off between dictionary size (coverage) and dictionary confusion can probably be overcome with more sophisticated context-based correction schemes such as language models based on the statistical properties of a suitably typical set of training data. (See Chapter 5.)

10) There seemed to be a consensus among the Second Conference Participants and Committee that improved form design and utilization of the context provided in all three fields would decrease the field error and distance rates significantly. The systems in the Conference recognized each field separately and did not utilize context across the three fields.

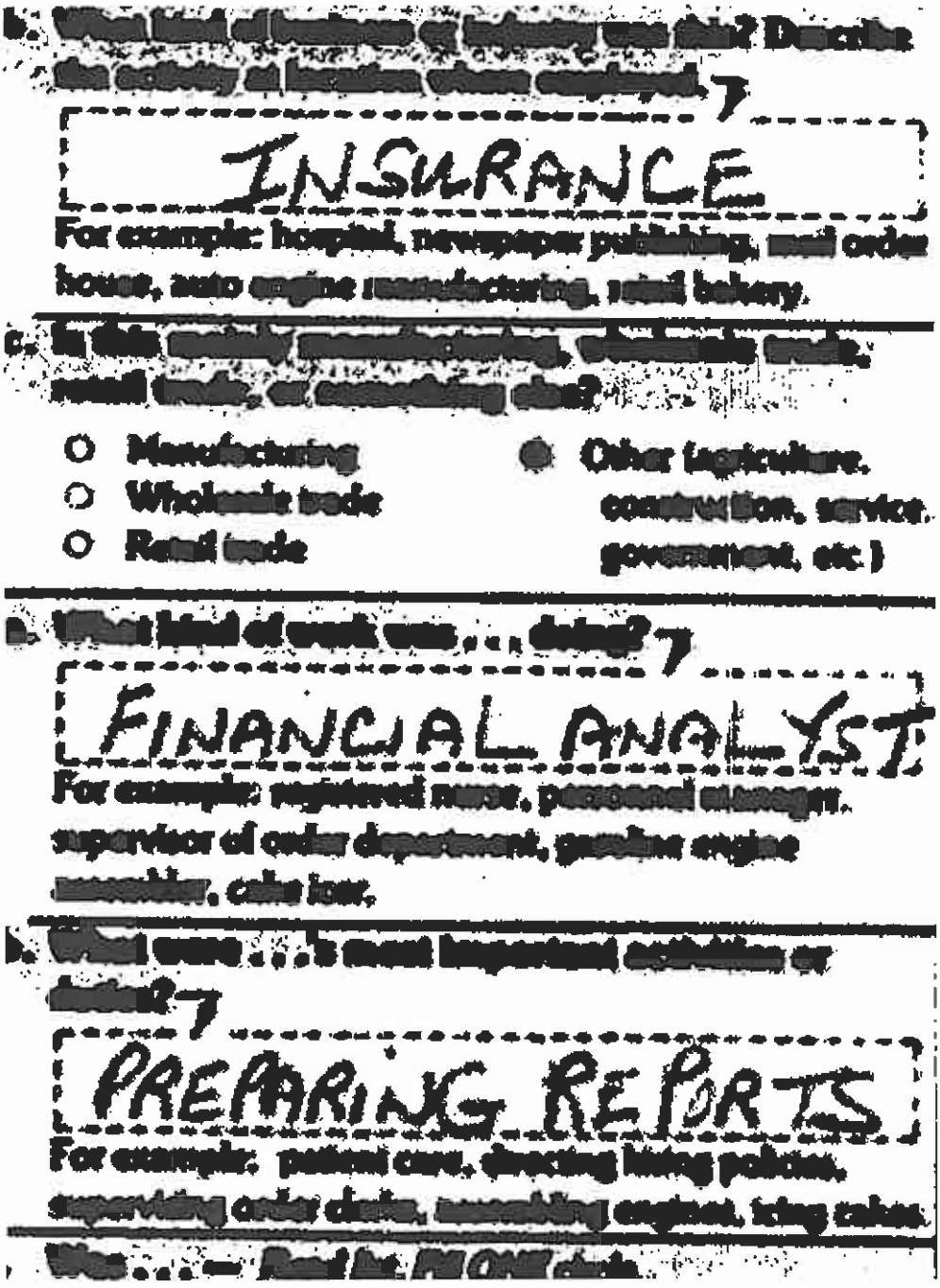


Figure 1: A well above average quality image scanned from microfilm.

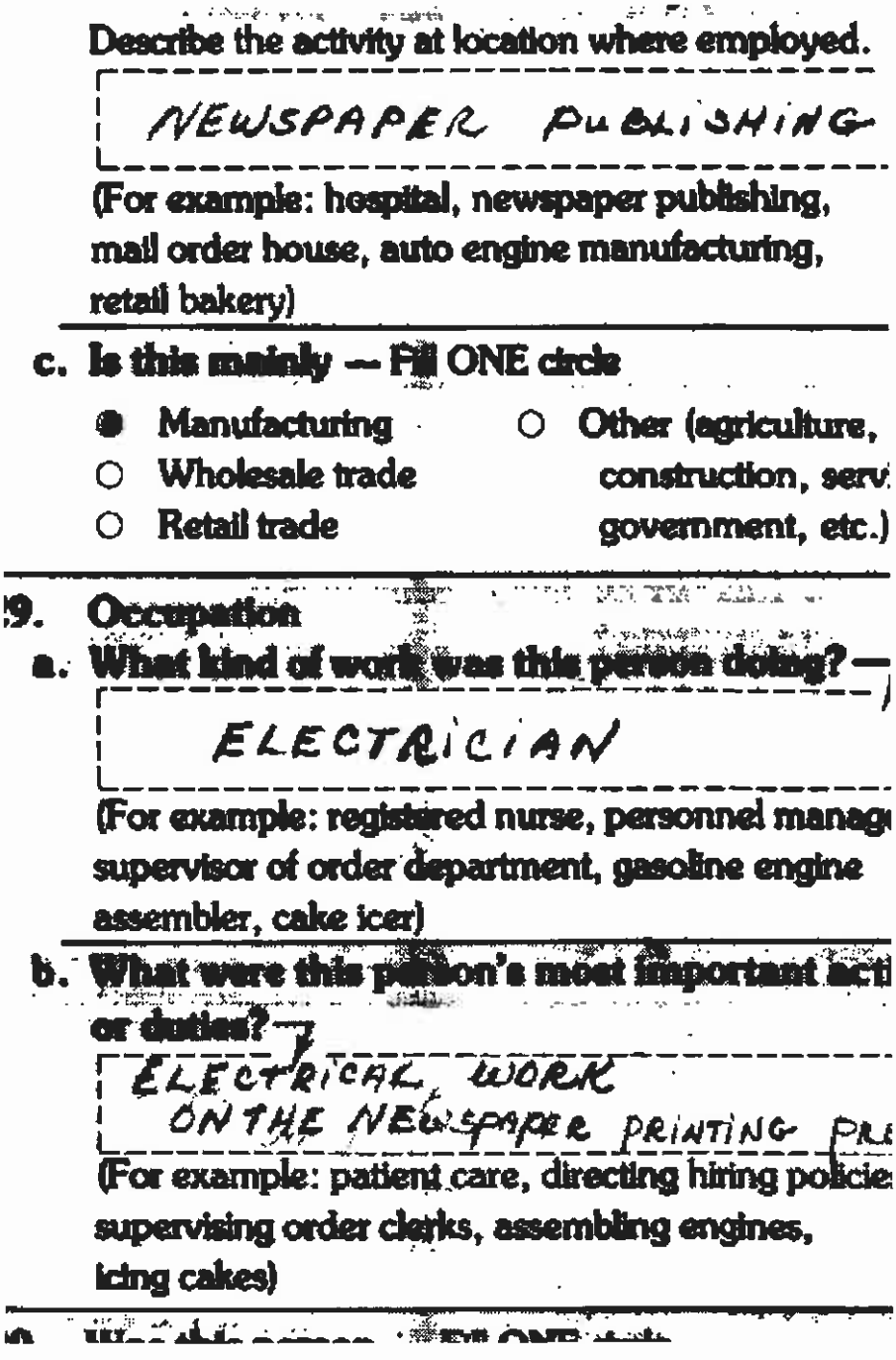


Figure 2: A typical image scanned from paper.

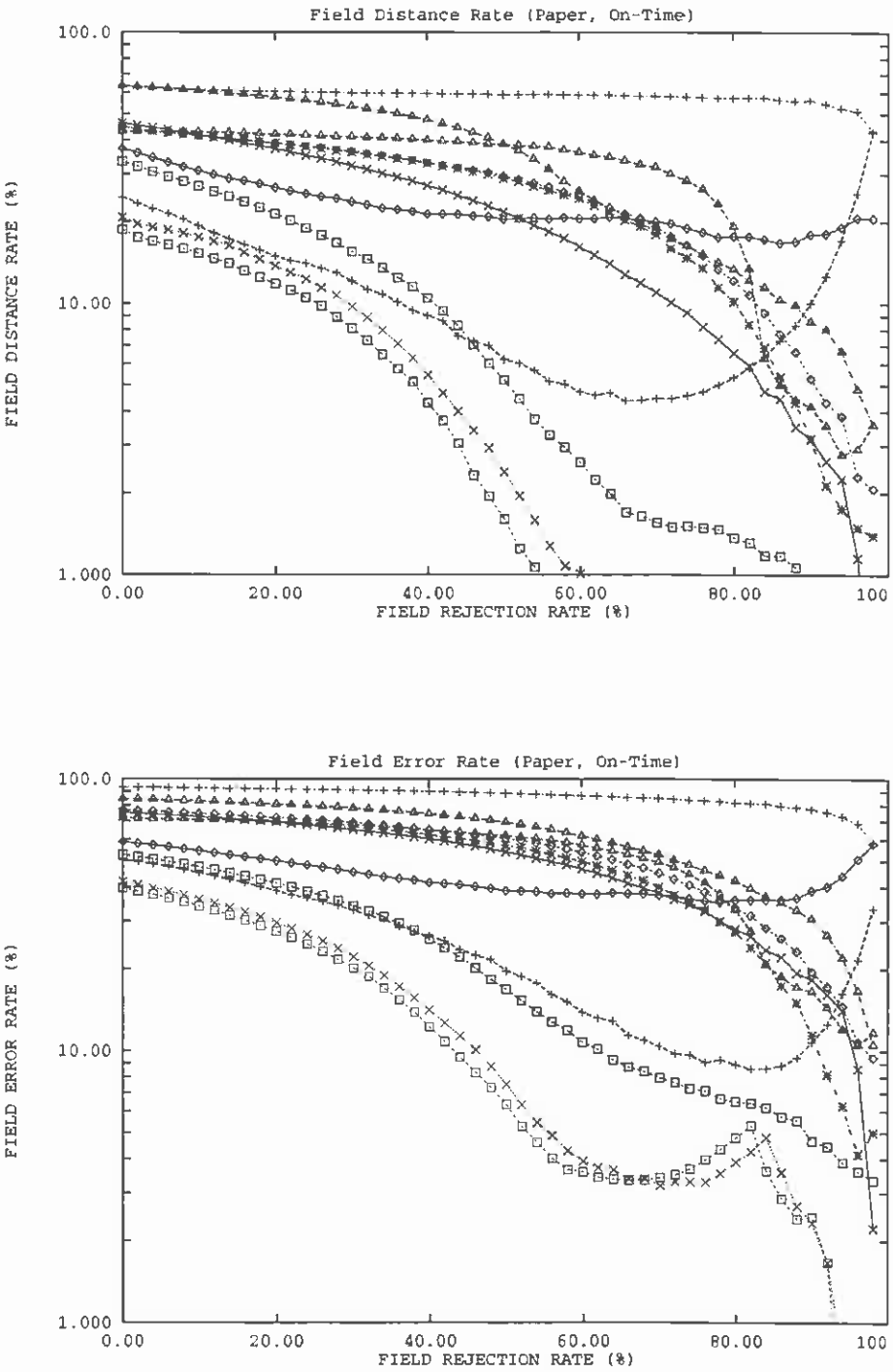


Figure 3: Field distance and error rates versus field rejection rate for all on-time system submissions for the test images scanned from paper.

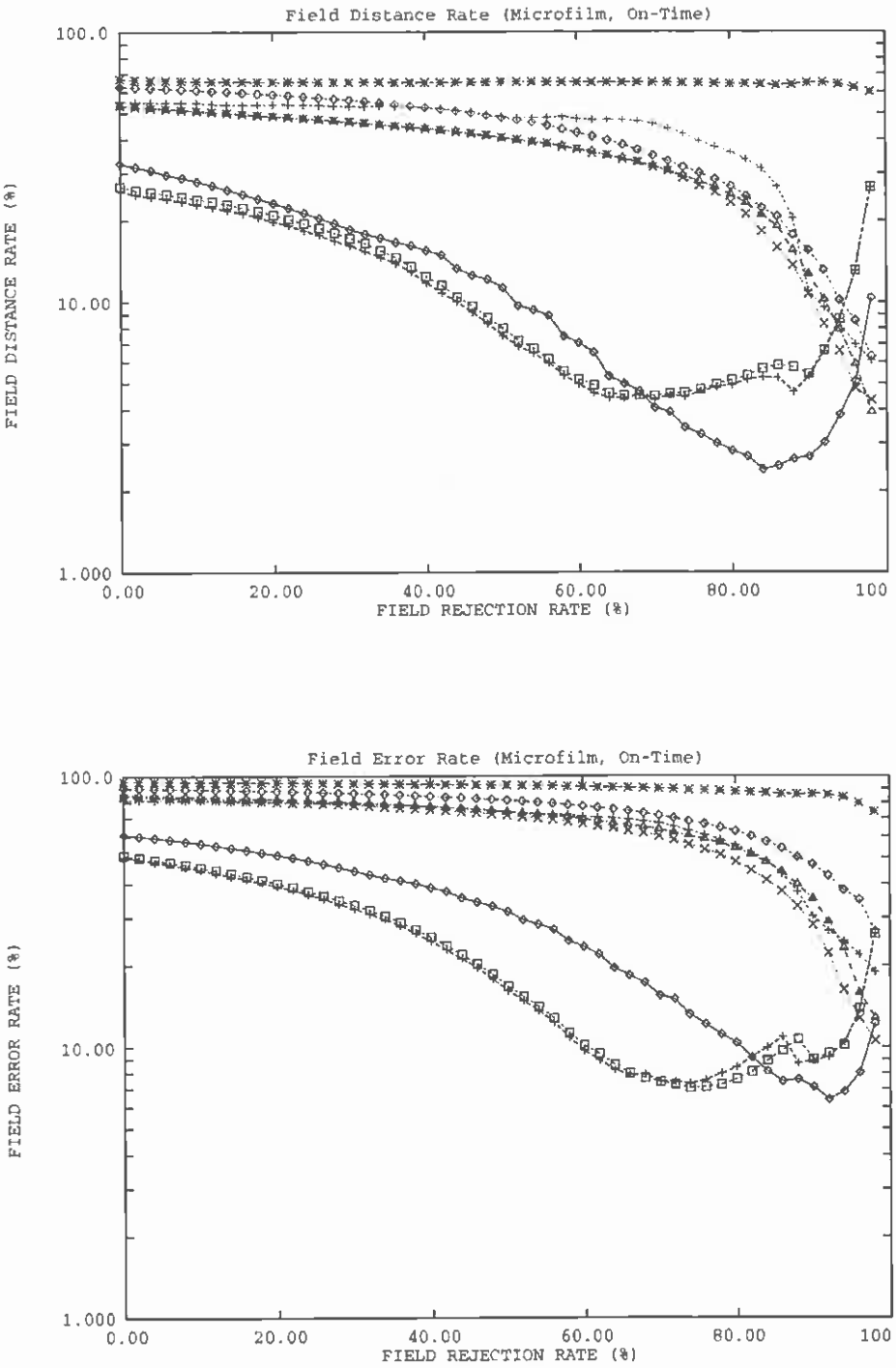


Figure 4: Field distance and error rates versus field rejection rate for all on-time system submissions for the test images scanned from microfilm.

3 Machine Recognition of Isolated Characters

Jon Geist

3.1 Background

Within the context of this report, isolated characters are single characters presented in isolation from all context not contained within the character itself. Even though the recognition of isolated handprint characters is not a task with a real application, it was the task chosen for the First Census OCR Systems Conference. By contrast, the Second Conference had a real application as its task: reading handprint word and phrase answers from a form. In this case, the recognition of isolated characters is just one of many subtasks of the real application, and it is not necessarily the error limiting subtask as mentioned in Chapter 2.

In this connection, it is important to understand that people do not read words and phrases by isolating each of the characters in the word or phrase, recognizing each of them in isolation from the other characters, and then reconstructing the word or phrase from the isolated recognitions. They do not even read multidigit numbers this way.

For example, some people print fours that look like other people's nines. It is much more common for an isolated handprint 4 to be mistaken for a 9 or vice versa than for a handprint 49 to be mistaken for a 94, 44, or 99. Each digit of 49 carries information not only about its own identity but also about how the adjacent digit should be decoded. When people read words, phrases, and multidigit numbers they use a large number of contextual clues carrying syntactic, semantic, and other types of information such as how nearby characters are formed.

Even though the OCR of isolated characters has no real application except as part of a larger process, it has long been considered the bellwether of OCR capability. This is the reason that the First Conference tested this capability. The results of that conference convinced most of the participants that it was time to see how well words and phrases could be read. Implicit in this decision was the assumption that machine recognition of isolated characters would probably not be the performance-limiting step in this much more complex task. Chapter 8 presents data that support this assumption.

The present chapter presents data obtained following the First Conference that suggests that there is no reason to believe that humans are currently superior to machines in the classification of isolated handprint characters in any economically significant way. This is so despite the fact that the results of the Second Conference show that humans are still clearly superior to machines in the reading of handprint words and phrases. Before presenting this data, it is necessary to discuss human classification results obtained on the First Conference test materials.

3.2 Human Classification of the First Conference Tests

Before the First Conference, a machine assisted human classification process [?] was used to classify the 60,000 digits, 12,000 upper case letters, and 12,000 lower case letters used for the Conference tests. This procedure allowed context bias to influence the classification of the characters, but it also presented the most difficult characters to more than one human for classification. Only in the case where there was agreement between the last two humans who saw a character that had been flagged as having a problem was that character included in the test materials. The classifications obtained by this procedure were used as the references for the test.

During the First Conference, the author (JG) hand-classified the first 10,000 digits of the test data in a few periods of roughly one hour duration separated by a day or more. Each unknown digit was presented free of any external context on a computer terminal in the same random order as on the test CD-ROM. One of the digit keys was pressed followed by the Return key to indicate the classification, or the question-mark key was pressed followed by the Return key to reject the unknown digit as unclassifiable. Rates of about one character per second were sustained for anywhere from 10 minutes to over an hour. (The Report on the First Conference incorrectly stated the rate as two characters per second.) The time needed for the computer to display the image of each character took a substantial portion of the classification time.

It is the author's impression that the possibility of rejecting a digit was important to both speed and job satisfaction in this task. In most cases the digits were recognized faster than the keys could be punched. Usually, if a digit was not immediately recognized, it was not perceived as being ambiguous between two or more characters, but as being completely unknown. On the other hand, there were a few occasions where a possible ambiguity with another digit was noticed just as a key was being punched. By then it was too late to hit the question-mark key.

Whenever no digit was immediately recognized, pressing any key other than the question mark would have required stopping an apparently reflexive recognition process and starting some higher level cognitive process. Not only would this have been very frustrating, but it would have slowed the classification process significantly.

For instance, Ref. [?] describes a human recognition procedure that required a great deal of human cognitive effort to identify the 360 most difficult digits from a 17,000 sample of isolated digits. Thinking about an unknown digit, assigning normalized confidences to each of the ten digit classes, and then explaining the reasons for the choices that were made took the human classifiers on the average of between two and three minutes per digit. Neither this procedure, nor the reflexive recognition of a very large set of isolated digits, is particularly frustrating, but mixing the two seems to be.

Following the First Conference, a technician used the same system that was used by JG during that Conference to classify all of the characters of the Conference test data. For this experiment a two-pass process was used. In the first pass, the technician was instructed that accuracy rather than speed was the main goal, but that speed was desirable, and that he was free to reject any characters that he couldn't immediately classify. This task was carried

human 1 (1st pass)		human 1 (2nd pass)		human 2 (JG)	
rej. rate	error rate	rej. rate	error rate	rej. rate	error rate
DIGIT TEST					
0.0000	0.0361	0.0000	0.0071	0.0000	0.0157*
0.0334	0.0028	0.0056	0.0045	0.0122	0.0035*
UPPER CASE LETTER TEST					
0.0000	0.0848	0.0000	0.0377	NA	NA
0.0599	0.0316	0.0046	0.0354	NA	NA
LOWER CASE LETTER TEST					
0.0000	0.1388	0.0000	0.0862	NA	NA
0.0855	0.0583	0.0317	0.0697	NA	NA

Table 5: Results of human classification of all of the isolated characters in the digit, upper case letter, and lower case letter tests used for the First Census OCR Systems Conference. (* First 10,000 digits only; NA = Not Available)

out at a much slower rate than what was done by JG during the Conference. After the first pass was completed, the results were scored to give two data points: one for zero rejection and one for whatever fraction the classifier chose to reject during the first pass through the data.

In the second pass, the characters that were rejected on the first pass were presented for classification as in the first pass: one at a time, without context, and in the random order in which they occurred in the test. For this pass, however, the technician was instructed that he could spend as much time as needed thinking about each character that he had rejected during the first pass before classifying it or rejecting it again. Thus, this pass was carried out at an even slower (presumably more cognitive) pace. After the second pass, the accepted character classifications produced during the first pass were combined with all of the classification produced during the second pass, and the combination was scored, again giving two data points.

The results of human classification of the First Conference tests are summarized in Table 5. At first glance, there seems to be a great deal of difference between the results produced for the digit test by the two different human classifiers. However, most of this is explained by the fact that the two humans chose to reject very different fractions of the data set on the first pass. If the first-pass zero-rejection rate data point for each human is ignored as too strongly affected by the subjective choice of which digits to accept and which to reject, then the remaining data points appear roughly to fall on the same curve. Therefore they were combined to give the composite curve for human digit recognition that is shown in Fig. 5. (It had been demonstrated previously that the first 10,000 digits were typical of the full 60,000 digit test. [?])

On the first pass of the digit test, human 1 rejected over twice the fraction of digits as human 2. Recall that human 2 sometimes noticed possible two-character ambiguities during classification of some of the unknown digits, but too late to reject them. Also, recall that human 1 carried out the the first pass classification process at a much slower rate than human

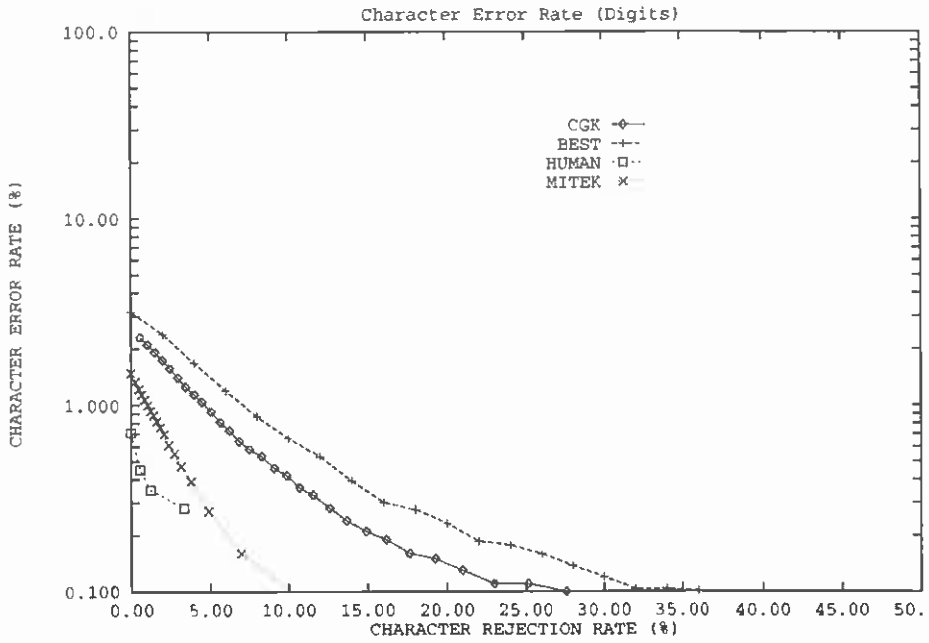


Figure 5: Post-conference results for the First Conference digit test: BEST = lowest value at each rejection rate from the First Conference; HUMAN = composite of two-pass data from one human and one-pass data from another; CGK = data submitted by CGK; MITEK = data submitted by Mitek.

2. Therefore, it is possible that the difference in rejection rate reflects the ability to notice and reject more possible two-character ambiguities as the classification rate is decreased. However, if this is a correct interpretation, it means that the first class that comes to mind when viewing unknown images that are ambiguous between two characters is usually the correct class. Otherwise the error rate for human 2 at the rejection rate of 0.0122 would have to be much greater than 0.0035.

3.3 Machine Results Obtained Following the First Conference

About a year after the First Conference, two organizations that did not participate in that conference sent unsolicited results obtained on one or more of the Conference tests. [1] Mitek submitted results only for the digit test, while CGK submitted results for all three tests.

[1]The test data and references for the First Conference are on available on CD-ROM and DOS-format floppy disk, respectively, as Special Database (SD) 7 for US $1000 from Joan Sauerwein, Standard Reference Data, NIST 221/A320, G'burg, MD 20899, (301)975-2208 (voice), (301)926-0416 (FAX), srdata@enh.nist.gov (e-mail). Note that some of the more basic utilities needed for handling the data formats on SD 7 are available on SD 3, which sells for US $895.

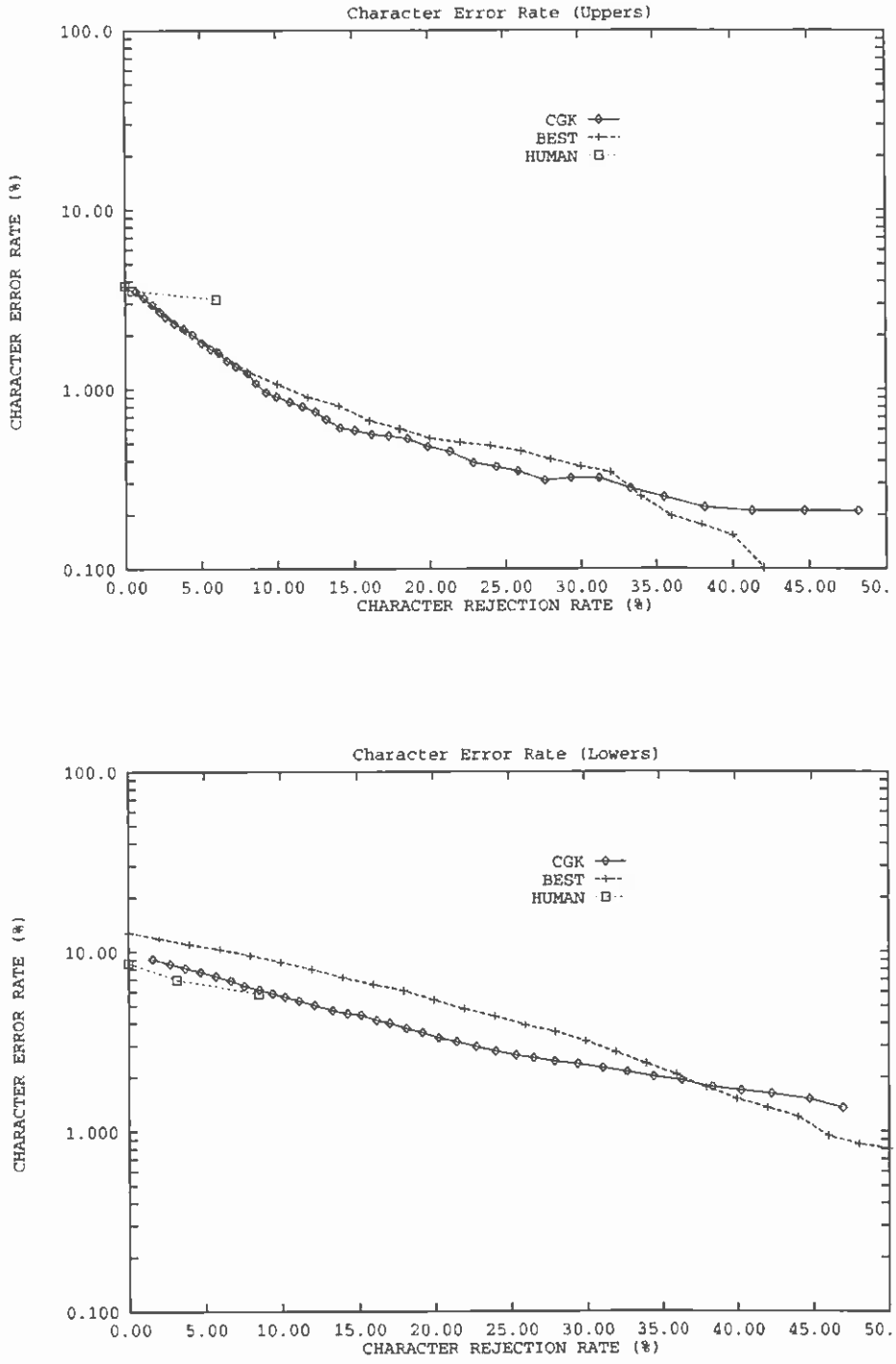


Figure 6: Post-conference results for the First Conference upper case and lower case letter tests: BEST = lowest value at each rejection rate from the First Conference; HUMAN = two-pass data from one human; CGK = data submitted by CGK.

The lowest value at each rejection rate (BEST) of the results from the First Conference, the human results, and the CGK results for the digit test are compared in Fig. 5 and for the upper and lower case letter tests in Fig. 6. The Mitek results for the digit test are also included in Fig. 5.

For very low rejection rates, the human error rate for digits that is shown in Fig. 5 is over a factor of two lower than the lowest machine results. For upper case letters, the error rates for two machines are less than or comparable to the human results for all rejection rates. For lower case letters, the human results are marginally better at low rejection rates. Even in the case of digits, the human results are not superior to the machine results in any economically significant way. The human advantage in accuracy is more than compensated by the machine's ability to quickly generate both classifications and reasonably reliable confidences. For instance, the Mitek test results were obtained at 15.5 characters per second.

Requiring humans to generate confidences as they classify the characters is far too slow a process to be economically feasible. However, without confidences, the maximum rejection rate is determined by the arbitrary number of unknown images that different human classifiers choose to reject on the first pass through the images. Once the human has completed the first pass, the maximum rejection rate has been set, and further rejection to obtain a lower error rate is not possible. Also, the low human error rate at zero-rejection rate is not even obtained until after the second human pass is complete. For example, comparison of Table 5 and Fig. 5 shows that the first-pass error rate at zero-rejection rates for both human 1 and human 2 were significantly greater than that for Mitek.

The real advantage of human classification does not appear until after the second pass, so there is actually no reason to have a human do the first pass. Machines can do it faster, and machines can generate confidences as part of the classification process. This means that a hybrid machine/human classification process should be better than either alone. For instance, suppose a human reclassified the 10% of the unknown images having the lowest confidence following the Mitek classification process, and these new classifications were added to the other 90% produced by Mitek. It is hard to imagine that the error rate would be as large at any rejection rate as that obtained by either Mitek or the human alone. Furthermore, it should take significantly less time since the human needs to classify only 10% of the unknown images, even though it is the most difficult 10%. Finally, any first-pass rejection rates could be used just as easily as 10% in contrast to the case with human first-pass classification, where this value is set by the human during classification in an uncontrollable way.

3.4 Conclusions

The information presented in this Chapter suggests a number of conclusions about economically significant human classification (ESHC) compared to machine OCR of isolated handprint characters. However, all of these conclusions must be considered tentative because they are based on the very limited and incomplete data reported in this Chapter.

1) ESHC of isolated handprint characters is a mostly reflexive process.

2) The error rate at zero rejection rate for ESHC of isolated handprint characters is unrealistically high and variable due to the subjective choice of human classifiers to either reject or think about the rare occurrences of characters that they cannot immediately (reflexively) classify. The problem is that the action of stopping a fast reflexive process, initiating a hierarchy of slow cognitive processes with an unknown end point, and then restarting the reflexive process is not a comfortable way to work when under the least bit of time pressure.

3) The second pass of a two-pass classification process produces a more realistic estimate of the zero-rejection rate error rate for ESHC. This process may also reduce human-to-human variability in the error versus rejection rate data.

4) Two machine classifiers outperformed ESHC on a test consisting of about 12,000 handprint letters consisting mostly of upper case letters but containing a significant fraction of lower case letters.

5) One machine classifier produced comparable performance to ESHC on a test consisting of about 12,000 lower case handprint letters.

6) Two machines produced error rates at rejection rates of 5 and 10%, respectively, that were less than that produced by ESHC at a rejection rate of 3.5% on a test consisting of about 60,000 handprint digits. Extension of these ESHC results beyond a rejection rate of 3.5% would require adoption of a forced cognitive process like that described in Ref. [?], but such a process would be far too slow to be economically significant. Therefore, processing a set of isolated handprint digits by machine, and having humans reclassify a low confidence subset should produce a lower error rate at less cost than conventional ESHC of the same set.

7) There is no reason to believe that ESHC capabilities are inherently superior to machine capabilities in the classification of isolated handprint characters. Currently, machines appear to be superior at some aspects of the task, humans at others. Machine performance will improve within the next few years; human performance will not.

4 Sample Selection, Image Capture and Reference Data

Norman Larsen and Bob Hammond

4.1 Introduction

This section describes the selection of sample forms and the creation of the digital images and reference data for the training and test materials for the Second Census OCR Systems Conference. The primary objective was to create a reliable statistical sample of a large number of different writers.

During the 1990 Census, over 100 million forms were collected from the American public. Most households received a "short form" booklet that included seven questions about each person and several questions about the housing and living arrangements; most of the questions were multiple choice. About 17 million households received a "long form" booklet that included up to 33 questions about each person and up to 26 questions about the housing and living arrangements. Many of these questions, such as those about ethnic origin, migration, cost of utilities, and occupation required handprinted responses.

All of the images for this Conference came from the 1990 Census long form's Industry and Occupation questions. Initial plans included only images scanned from microfilm. The average quality of these images was sufficiently poor to warrant midstream adjustments to plans. A subsample of images were then created from the original paper documents. The methods and techniques to capture these images, and the related reference data, are described in separate sections below. The bulk of the dictionary data came from the 1980 Census, but it was augmented with training data from the 1990 Census as described in Chapter 5. All census forms were processed in seven regional processing offices. In general, after extensive handling and clerical edits, the forms were microfilmed and the mark sense answers were captured from the film by FOSDIC scanners (Film Optical Sensing Devices for Input to Computers). The long form booklets were then sent to key entry operators who captured the handprinted answers.

The industry and occupation questions were the basis for the miniforms created for this Conference. Questions 28b: "What kind of business or industry was this? Describe the activity at location where employed.", 29a: "What kind of work was this person doing?", and 29b: "What were this person's most important activities or duties?" were the basis of the training and test miniform. The answers to these questions along with question 28a, "For whom did this person work?", were keyed during the 1990 Census and used by a semiautomatic process to determine a 3 digit industry and 3 digit occupation code for each member of the labor force. These coding lists were derived from the Census Bureau Standard Industrial Classification (SIC) and the Standard Occupational Classification (SOC) coding schemes.

Question 28a was not included in the test to comply with the privacy provisions required by

Title 13 of the U.S. Code. A manual check was also made on the other fields to ensure that they did not contain information that might identify any individual.

Two templates were used to print the forms: one was mailed directly to the household and the second was filled in by census enumerators who visited those households that did not return a form by mail. These two form types had small differences in the layout that the OCR systems had to recognize.

4.2 Images Digitized from Microfilm

In total, about 120,000 rolls of microfilm were created during the 1990 Census; approximately 53,000 rolls contained images of the long forms. After the census, the film was packed in boxes (maximum 92 rolls per box) and shipped to the archives. The sample for this Conference consisted of the 25th box of long form film from each of the seven processing centers (giving 644 rolls of microfilm). This arbitrary selection rule was selected to include a representative mix of questionnaires that were completed by respondents and returned by mail and those that were completed by census enumerators. Each roll of film had a control number for census processing. By coin toss, the odd numbered rolls were used for the training images and the even numbered rolls were used for the test images.

Each roll contained images from up to 400 forms. From this film, one image was created from each long form questionnaire that met the following conditions: 1) the form was for an occupied housing unit (not vacant), 2) the form was the Census "form of record" that was used to generate the final population count for the census, and 3) the form contained writing in the answer box for question 29a.

A Kodak Imagelink Digital Workstation was used to generate a TIFF image of each selected form. The scanner was set at 200 dots per inch and conversion was set to transpose black and white from the silver halide negative microfilm. The microfilm density was used to set the scanner binarization parameters for the entire roll. This was not entirely successful as there was considerable variance in the quality of the images obtained from a single roll. This was partially caused by a shadow created during filming by a fold in the form that passed through the questions.

A crude check was made on the number of pixels in a portion of the general area where the response was expected. If the number of pixels was too high (too much noise) or too low (too faint an image) the image was rejected. Each acceptable image was then cropped further and converted into the NIST IHead format.[?]

A file containing the keyed data for each form was used to extract the reference data for each image in the sample. These keying operations had some verification for quality control purposes, but an unknown amount of error still remained in this reference data. See the discussion of reference data for the paper subsample.

4.3 Images Digitized from Paper

During the preparation of the images digitized from microfilm, the Conference Committee concluded that the average quality of the images was sufficiently poor to cause substantial problems for many OCR systems. Some committee members suggested that a test of images from microfilm only would be more a test of image enhancement tools than of OCR techniques. Indeed, some participants dropped out of the conference after seeing only the microfilm images. To avoid the risk of invalidating the entire study, a subsample of images was created from the original paper forms (even though this decision delayed the original conference schedule by several months).

The original paper booklets for those forms captured in the Jeffersonville Indiana processing office had been retained for future OCR testing. These forms were collected from Indiana, Illinois, Michigan, Ohio, and parts of Missouri. About 28,000 forms corresponding to the previously chosen microfilm forms were pulled from storage. Census staff in Jeffersonville located each form and page in the sample, attached a bar code label for future identification, and cut the labeled page from the booklet. These pages were shipped to Census headquarters and scanned on a Fujitsu 3096 scanner using dynamic thresholding at 200 dots per inch. Staff at NIST made appropriate changes to the postprocessing software which was used to count pixels, crop and convert the resulting TIFF image files to the NIST IHead format.

For some bar-code labels, the label paper was not opaque enough to prevent information under the label from being captured by the scanner. This smudged a fraction of the labels to the extent that the NIST software could not read the bar code. These bar codes were keyed into the system manually.

The data for the microfilm sample was keyed and an 8% sample was verified during the Census, and it was known that the results of the keying were not error free. To eliminate this source of error in the scoring, the paper form data was keyed twice by different keyers and discrepancies were resolved by a third keyer. The resulting transcriptions were used as the reference data for the portion of the test scanned from paper. Both the machine OCR results and the results keyed during 1990 Census were scored against this reference data.

4.4 Conclusions

The effort to produce images and reference data for this conference reiterate [or confirm] the need to evaluate plans and equipment along many different dimensions before selecting a method. An integrator should attempt to select the methods and equipment that produce the "best" result for the intended application. Some of these considerations are described below:

1) Many different CCD scanners (or components) are now available in the commercial market. This equipment varies widely in cost, speed, resolution, flexibility and transport capabilities.

2) The algorithm for converting the grayscale bit pattern to a binary image is critical if binary images are going to be recognized. Lower cost binarization schemes do not provide much flexibility for source materials that have significant variability. Future algorithms

that perform adaptive thresholding may solve this problem, but may add cost to the initial equipment acquisition and to the system performance.

3) There are moderate differences in the performance of the various methods of lossless compression of binary data that are commonly used today. However, this variable had minimal impact on the production efforts for this conference.

5 Dictionary Production for the Conference

Jon Geist and R. Allen Wilkinson

5.1 Introduction

The miniforms used as test and training material in the Second Conference were made from a portion of the Industry and Occupation section of the 1990 Census Long form (D-2) that contained questions 28b, 28c, 29a, and 29b. Each participant was asked to recognize the handprint word and phrase answers on the test miniforms, remove all punctuation, convert all lower case letters to upper case, and return the results to NIST for scoring.

As discussed in Chapter 3, humans do not usually recognize the characters within a word or phase in isolation from the surrounding characters and then reconstruct the word or phrase. Instead, they recognize the word or phrase as an entity using all sorts of context besides the individual letters. Similarly, OCR systems need context to improve their performance. A dictionary of allowed or expected words or phrases is one type of context that can be used to improve the accuracy of current OCR systems.

Two types of data were used in the creation of the dictionaries for the Second Conference. The first type was obtained from a sample of 132,247 of the forms containing the Industry and Occupation answers obtained during the 1980 Census. The use of this sample simulates the actual Census situation. Even though there is no way to predict exactly what words and phrases will appear in any future census, the preceding census should be a good statistical predictor of a large subset of the answers.

The second type of data used in dictionary creation were phrases from the training miniforms used for the Conference. In fact, the dictionaries on the second training CD-ROM (SD 12) and on the test CD-ROM (SD 13) had been augmented by inclusion of any new answers from the CD-ROMs that had preceded them. Notice that these answers came from the 1990 Census, which was the sample from which the test data was drawn. Note also that the dictionaries on any given CD-ROM were not augmented with answers appearing in the images on that same CD-ROM, only with those from previously issued CD-ROMs. This simulates the incremental extension of the dictionaries during a census by using data obtained earlier in that census.

Within the above categories, three different types of dictionary were produced: word dictionaries, long phrase dictionaries, and short phrase dictionaries. One of each of these three different types of dictionaries was made for each of the three answer fields on the miniforms. The word dictionaries contained every unique word from a master list of phrase answers (after removing punctuation). The master list for each new CD-ROM was made from the union of the 132,000 sample from the 1980 Census and the answers on every miniform that had already been distributed to the participants on any earlier CD-ROM. The long phrase dictionaries were similar to the word dictionaries except that they contained full phrases instead of words. The short phrase dictionaries contained only those phrases that occurred at least twice in the master list.

There is a fundamental trade-off associated with dictionary size. Adding words or phrases to a dictionary may improve the coverage of the dictionary over that application. Any such increase in coverage will contribute to improved OCR accuracy. On the other hand, every additional word or phrase in the dictionary increases the confusion among the different words or phrases in the dictionary. This increase in confusion may contribute to reduced OCR accuracy.

The coverage of the word dictionaries over the test answers was much greater than that of the phrase dictionaries. On the other hand, there is no need to segment each phrase answer into its constituent words, and there is less confusion among the entries in the phrase dictionaries. This is particularly true of the short phrase dictionaries, which are much smaller than the long phrase dictionaries with only slightly reduced coverage over the test answers. This fact suggests that the Industry and Occupation answers come from a long-tailed distribution that can be approximated as the union of two sets: a relatively small set containing commonly occurring answers, and a very large set containing rarely occurring answers.

Obviously the addition of a word or phrase that does not actually appear in an application can never improve coverage, only confusion. On the other hand, the addition of a word or phrase that occurs a great number of times will almost certainly improve coverage more than enough to compensate for any increase in confusion. The problem is that it is never possible to know exactly which words or phrases will appear and which will not in any particular application. Preliminary experiments that illustrate the competition between coverage and confusion with increasing dictionary size are described in Section 5.3.

Because they carry little or no extra information, both punctuation and misspellings tend to increase dictionary confusion. Furthermore, it seems unlikely that the Industry and Occupation codes assigned by the Census Bureau would be changed by removal of the punctuation and correction of the misspellings. Therefore, punctuation was removed, not only from the dictionaries, but from the reference data as well. This significantly reduced dictionary confusion while at the same time increasing dictionary coverage. Unfortunately, the same sort of gain was not possible with misspellings due to the presence of abbreviations.

Misspellings tend to be very similar to the word from which they are derived, and therefore contribute very significantly to dictionary confusion. With but a few exceptions, the fact that a misspelling is common means that the correctly spelled word is even more common. Since the misspelled word and the correct word are easily confused, there are probably more cases where the correctly spelled word is incorrectly recognized as the misspelled word and where the misspelled word is incorrectly recognized as the correctly spelled word than there are of the misspelled word being correctly recognized. Therefore, correcting misspelled words in dictionaries should improve overall system accuracy in context-based correction of raw OCR.

Abbreviations, according to their purpose, tend to be much shorter than the words or phrases from which they are derived, so there is very little similarity. Therefore, they do not contribute strongly to dictionary confusion unless they happen to be similar to some other entirely different word or phrase (which is, of course, sometimes the case). As a result they are unlikely to be recognized as the word or phrase from which they are derived. In contrast to the case with misspellings, removing common abbreviations from dictionaries should

reduce the usefulness of the dictionaries.

In any case, neither the misspellings nor the abbreviations were removed. There were two reasons for this. The first is that neither NIST nor Census could come up with a foolproof procedure to automatically remove the misspellings in the time allotted without also removing the abbreviations. The second is that the Census key entry operators were instructed to preserve misspellings and abbreviations, rather than correct them in their work. So in a certain sense, this made the test more realistic.

5.2 Producing the Dictionaries

Figures 7, 8, and 9 in this section illustrate some of the steps in the dictionary creation process, and Figs. 10, and 11 present small samples of the actual dictionaries at various stages in the process.

```
 0 0 4 1 412 354
ZNT-OPERATOR
OPEATOR
OPERATOR
 0 0 1 1 250 259
GLASS MARUF
MANUFACTURING GLASS
MANUFACTURING GLASS
 0 0 4 1  11 274
FFED LOOT
SELLING CATTLE
SUPERVISOR
 0 0 0 1 391 674
LAMPSHADE MARV
MAKING LAMPSHADE
GIVING
 0 0 4 2 910 179
BKING JUDGE
JUDICAL
JUDGE
```

Figure 7: Sample entries used to illustrate the dictionary creation process.

A list containing the keyed responses to three of the Industry and Occupation questions from a sample of 132,247 1980 Census Long Forms was used to make the dictionaries. The list has four lines of Industry and Occupation data for each Census form in the sample. The first line, Line 0, is the Census-Bureau assigned Industry and Occupation code for the information that follows on the next three lines as shown in Fig. 7. Line 1 is the response to question 28b: "What kind of business or industry was this? Describe the activity at

```
BKING JUDGE
FFED LOOT
GLASS MARUF
LAMPSHADE MARV
ZNT OPERATOR
```

Figure 8: Examples of phrases extracted from Line 1 entries of the sample list after removal of punctuation and alphabetic sorting.

```
BKING
FFED
GLASS
JUDGE
LAMPSHADE
LOOT
MARUF
MARV
OPERATOR
ZNT
```

Figure 9: Examples of words extracted from phrases in the sample list.

location where employed." Line 2 is the response to question 29a: "What kind of work was this person doing?". Line 3 is the response to question 29b "What were this person's most important activities or duties?".

Line 0 of the 1980 sample of Industry and Occupation answers was not used for dictionary production. The other three lines, Line 1, Line 2 and Line 3 were treated as phrases, and processed to produce separate phrase and word dictionaries for each line.

The very small (short sample) list in Fig. 7 will be used to demonstrate the process of dictionary creation. The phrase lists are generated by removing the same line of each field from each record in the 1980 answer sample. For instance, one phrase list contains all the Line 1 responses. All characters other than digits and upper case letters were replaced with spaces in these phrases. Fig. 8 lists all the Line 1 phrases for the short sample list with punctuation removed. Notice, for example, that the hyphen in ZNT-OPERATOR has been removed.

Even though not shown in Fig. 8, the master lists created from the original Census data have many entries with multiple consecutive spaces. Most of these were caused by the previously mentioned replacement of punctuation with spaces. Multiple spaces were converted into one space to clean up each phrase list. Each phrase list was then sorted in alphabetical order and all duplicate entries removed. The sorted lists were then visually edited by humans to

```
ACADEMIC INSTUTUTION
ACADEMIC LIBRARY BOOK JOBBER
ACADEMIC PEDIATRIC PRACTICE
ACADEMIC PHYSICS
ACADEMIC RESEARCH
ACADEMIC RESEARCH CENTER
ACADEMIC SCIENCE DEPTS
ACADEMIC TEACHING
ACADEMIC UNIVERSITY
ACADEMIC ZOOLOGY RESEARCH TEACHING
ACCESS CONTROL MFG
ACCESS FLOOR SERVICE CENTER
ACCESSIBILITY SURVEYOR ANALYST
ACCESSORIES FOR KNITTING HILLS
ACCIDENT INSURANCE FIRM
ACCOOPIONAL TABLES
ACCOUNBEE HEATING
ACCOUNT
ACCOUNT PAYABLE DIVISION
ACCOUNT REP
```

Figure 10: Examples of real phrases found as entries in the 1980 Census sample.

remove any sensitive information such as addresses, personal names, small-company names that might reveal the identity of the person filling out the form, and small town names. As a final check the new list was sorted and all remaining duplicates were removed. The resulting lists were very large. For Line 1 there were 46,593 unique entries, while Line 2 had 46,813 and Line 3 had 61,384. A small subset of real Line 1 phrases from the original 1980 Census sample is shown in Fig. 10.

The dictionaries described so far are called long phrase dictionaries. This means every unique phrase from the original list is represented in one of the dictionaries. Shorter dictionaries were made from the long phrase dictionaries. Word dictionaries that contain only the unique words in the long phrase dictionaries are shorter, as are dictionaries made by keeping only phrases which occur more than once in the long phrase dictionaries. For instance, the short phrase dictionaries for Line 1 had 8,216 entries; for Line 2, 8,516 entries; for Line 3, 7,831 entries. This is an excellent example of reduction in dictionary size, while still maintaining good coverage over the original list of answers. The short phrase dictionaries are 15% to 20% of the size of the long phrase dictionaries, but contain 60% to 70% of the phrases in the latter. It turns out that the coverage of the short phrase dictionaries was almost as large over the test data, while the long phrase dictionaries provided only a few percent more coverage.

The conversion of phrase lists into word lists is rather simple. Using UNIX utilities, it is possible to replace all spaces with newlines. This puts each word on a line of its own. The

```
ADVICORY
ADVISEMENT
ADVISING
ADVISMENT
ADVISOR
ADVISORS
ADVISORY
ADVOCACY
ADVOCATE
ADVRTISING
AENCY
AEOROSPACE
AERATOR
AERESOL
AERIAL
AERO
AEROBIC
AEROCPACE
AERONAUTICAL
AEROPLANE
```

Figure 11: Examples of words extracted from phrases in the 1980 Census sample.

word lists are sorted in alphabetical order and duplicate entries removed. The word list for Line 1 had 13,745 words, Line 2 had 13,879 words, and Line 3 had 16,333 words. Again there is a significant reduction in dictionary size compared to the long phrase dictionaries. This can be attributed to the redundancy of words within the phrase lists.

Fig. 9 shows the words produced from the phrases in the short sample shown in Fig. 8. Fig. 11 shows a small sampling of the words from the real Line 1 Census data. Notice that both of these examples contain misspelled words, some of which are misspellings, some of which are abbreviations, and a few of which might be either.

5.3 Dictionary Coverage versus Confusion

During the Conference test period, Wolfgang Lellman of CGK used one set of on-time and one set of late entries to carry out an experiment on dictionary coverage. He had all of the images on the test CD-ROM hand-keyed and added the phrases and words to the dictionaries provided on the test CD-ROM to make augmented dictionaries. The net effect was to substantially increase the coverage of dictionaries over the test data to nearly 100% with only a little increase in dictionary size. It is important to understand that this approach cannot be used in a real Census. This is the reason that CGK_1 and CGK_3 were not included among the submissions listed in Chapter 2. Nevertheless, the results of this experiment are

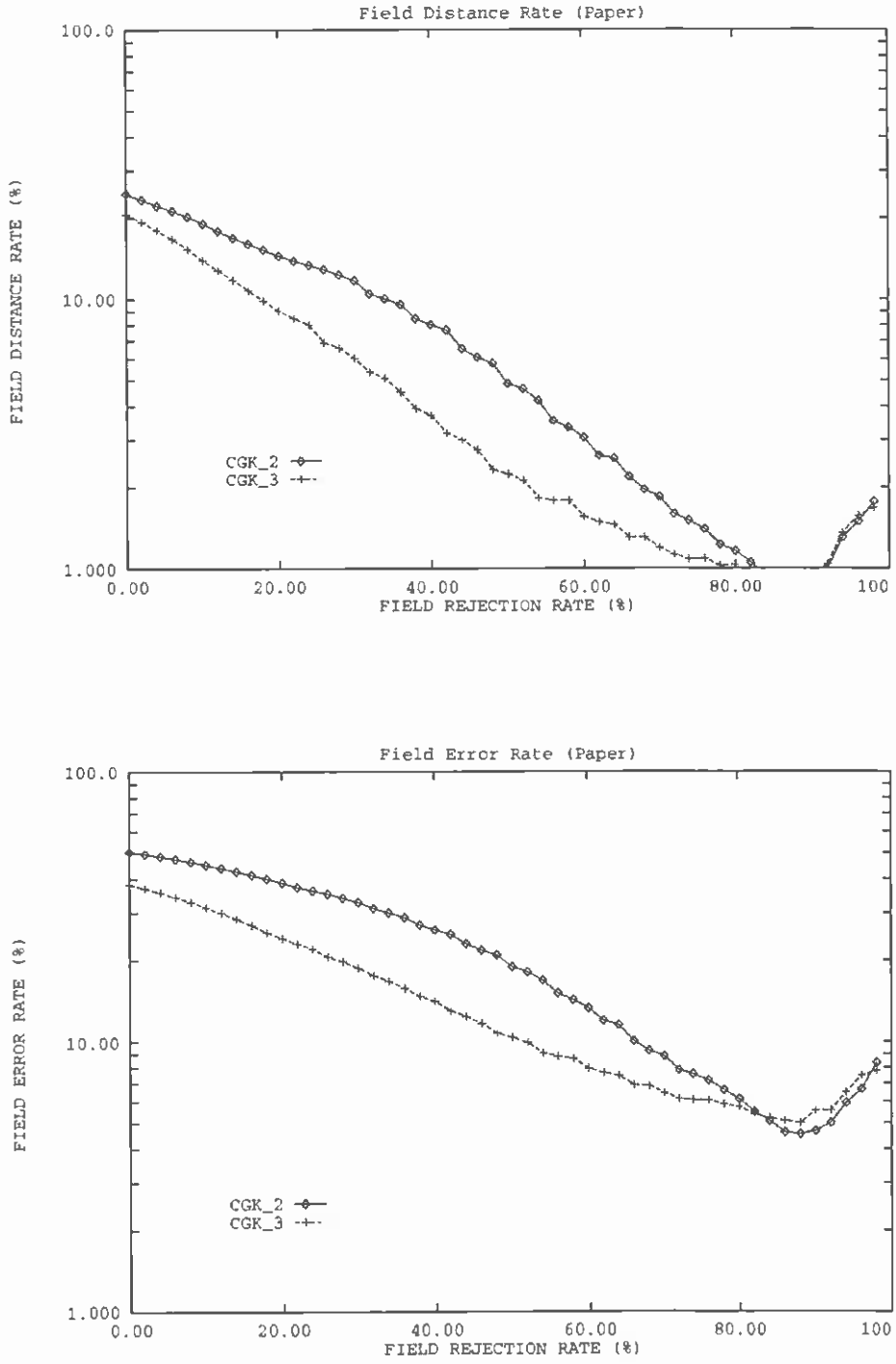


Figure 12: Field distance and error rates versus field rejection rate for two late CGK systems that differ only in dictionary coverage.

SYSTEM	FIELD ERROR RATE	FIELD DISTANCE RATE
CGK_2	0.5050	0.2459
CGK_3	0.3812	0.2066
NIST/CGK_3	0.2100	0.1542

Table 6: Comparison of effects of competition between dictionary coverage and confusion. The dictionaries used to create CGK_2 had moderate coverage over the test data set and moderate confusion. Those used to create CGK_3 had high coverage and moderate confusion. Those used to create NIST/CGK_3 had 100% coverage and the minimum confusion consistent with this level of coverage.

interesting as discussed below.

CGK_0 (on-time) and CGK_2 (late) were obtained from raw OCR results corrected with word and phrase dictionaries that contained between about 16,000 and 19,000 words and phrases, respectively, per field type. These dictionaries were made from the dictionaries provided on the test CD-ROM. CGK_1 (on-time) and CGK_3 (late) were obtained from the same raw OCR results corrected with the augmented dictionaries created as described above, and having nearly 100% coverage over the test miniforms. The augmented word and phrase dictionaries were about 300 words and 2000 phrases larger than the original word and phrase dictionaries, respectively. The results for CGK_2 and CGK_3 are compared in Fig. 12. It is important to understand that the only difference between CGK_2 and CGK_3 is the content of the dictionaries used in correcting the raw OCR. The additional confusion among the entries in the augmented dictionaries due to the addition of about 2% more words and 11% more phrases was more than offset by the improved coverage obtained with the additional size. This was assured because only those entries needed to achieve nearly 100% coverage were added to the dictionaries.

On the other hand, Thomas Breuel of IDIAP built a phrase dictionary with significantly greater coverage in a way that could be used in a real Census application. In essence, he combined all of the unique words in the three different word dictionaries in different plausible combinations to create different phrases. The resulting phrase dictionary, which consisted of about 150,000 phrases, had about 87% coverage over all three fields of the test miniform. This is to be compared with about 66% for a dictionary of 20,000 phrases made from the unique phrases in the short phrase dictionaries. The 20,000 phrase dictionary resulted in a significantly lower error rate than the 150,000 phrase dictionary. In this case, the additional confusion among the entries in the larger dictionary due to its greatly increased size was not offset by the improved coverage. The problem was the following: For every additional entry that improved the coverage over the set of test images, there were roughly 20 additional entries that did not. Therefore, the increase in error caused by increased confusion was greater than the decrease in error caused by increased coverage.

After the Conference test period, Patrick Grother of NIST carried out a related experiment. He passed the CGK results through the NIST dictionary-based correction algorithm using the minimum-size long phrase dictionaries needed to get 100% coverage over the test images, namely the reference data for the test. These dictionaries had only about 2,000 phrases per

field. The zero-rejection rate field-error and distance rates are compared with those for CGK_2 and CGK_3 in Table 6. Notice that there is a significant improvement over CGK_3, and that this improvement is associated entirely with decreasing the dictionary confusion to the minimum level consistent with 100% coverage.

5.4 Conclusions

A few tentative conclusions can be drawn from the material presented in this chapter.

1) The US Census Industry and Occupation answers, including misspellings and abbreviations, appear to come from a long tailed distribution that can be approximated as the union of a rather small set of commonly occurring words and phrases and a very large set of rarely occurring words and phrases.

2) Increasing dictionary coverage will not improve OCR accuracy if the increased coverage is purchased at the price of too large an increase in dictionary confusion.

3) The removal of punctuation from the Second Conference test removed confusion from the dictionaries and contributed to lower error rates than would otherwise have been obtained.

4) Misspelled words tend to contribute more strongly to dictionary confusion than to dictionary coverage, and the failure to remove them from the dictionaries probably contributed to higher error rates than would otherwise have been obtained.

5) Language models should help overcome the trade-off between dictionary size and confusion.

6 Scoring Procedures and Issues

Jon Geist

6.1 Introduction

Scoring the OCR of words and phrases is not as straightforward as scoring that of isolated, properly segmented characters. An isolated character is either correct or incorrect, and even though some character images are truly ambiguous, the problems caused by this fact are readily handled by a statistical analysis. [?] [?]

Each field containing a word or phrase is also either correct or incorrect, and the field error rate for fields of words and phrases is analogous to the error rate for isolated characters. Unfortunately, systems can achieve the same field error rate in very different ways that can have profoundly different effects on any given application. For instance, fields with only one incorrect character make the same contribution to the field error rate as fields in which every character is incorrect. Yet the former are probably useful in most applications, at least in the case of long phrase fields, while the latter are clearly useless in all applications.

One solution to the above problem is to perfectly simulate an application when testing OCR systems for that application, and to score according to the performance achieved in that simulation. Another solution would be to use a generic scoring method that could be tailored for use with any given application. These represent the opposite ends of the spectrum of possible solutions to the scoring problem.

The disadvantage of the first solution is that it requires creative effort every time a new application is considered or even when some aspect of an application is changed in a significant way. The difficulty with the second solution is that it does not currently exist, and short of developing and testing it, there is no way to be sure that it is even possible.

A second accuracy measure, the field distance rate, was defined and used in scoring the test results of the Second Conference to partially compensate for the problem with the field error rate that was mentioned above. This chapter defines both of these accuracy measures. It also discusses limitations of the field distance rate with respect to the Census use of the Industry and Occupation data, and possible extensions in the direction of a generic scoring method.

The next section of this chapter defines the field error and distance rates. One problem with the field distance rate is that it cannot be calculated until the characters in the hypothesis and reference phrases are aligned making allowances for characters being deleted from the reference and characters being inserted into the hypothesis by the OCR process. Therefore, the section defining the field error and distance rates is followed by a section discussing string alignment and related problems, then by a section discussing problems with the field distance rate and their possible solution through generalization, and then by a concluding section. Before starting the next section it is, however, important to give more precise definitions of some terms used in this report.

A classification or recognition process assigns an ASCII character to an image of a character, or a set of ASCII characters to an image of a word, phrase or some other similar string of characters. A classification may be obtained from human key-entry or from some machine process, and it may be correct or incorrect. If a classification is defined to be correct, then it is a reference, otherwise it is a hypothesis.

A rejection process divides a set of classifications into rejected classifications and accepted classifications. Only the accepted classifications are considered valid. It was the classification of isolated characters that was of interest in the First Conference, so rejection and acceptance were carried out character by character. It is the classification of complete phrases that is the task of interest for this Conference, so rejection and acceptance are carried out field (word or phrase) by field rather than character by character.

All submissions to the Second Conference provided a single confidence value for each answer field. This is a number (between zero and one for this Conference) that orders the classifications according to expected reliability. Example 1 of Appendix B shows a representative hypotheses file and its related confidence file.

6.2 The Field Error and Distance Rates

Two different measures of classification error were calculated for this Conference: the field error rate and the field distance rate. The field error rate $R_{f\epsilon}(r_f)$ as a function of field rejection rate r_f is defined as

$$R_{f\epsilon}(r_f) = \frac{F_\epsilon(r_f)}{F_c(r_f) + F_\epsilon(r_f)}. \tag{1}$$

where the field rejection rate r_f for a set of field classifications is defined as the ratio of the number of fields rejected by the rejection process to the total number of fields presented for classification. In eq. (1), $F_c(r_f)$ is the number of accepted fields that are correctly classified and $F_\epsilon(r_f)$ is the number of accepted fields that are in error, that is to say differ in any way from the corresponding reference fields. For the Second-Conference test, all characters except digits, letters, and single spaces were filtered out of hand-keyed classifications to make the references. Therefore, the filtered out characters did not contribute to the field error or distance rates.

If a confidence is associated with each field hypothesis, any desired field rejection rate r_f can be approximated by choosing a confidence threshold and rejecting any field hypotheses having confidences less than or equal to the threshold and accepting any having confidences greater than the threshold.

As an example of the field error rate, suppose that an OCR test requires reading two images, image 1 and image 2, and that each image contains a single answer field. Further suppose, that the references for images 1 and 2 are `DRIVES TRUCKS` and `WAITS ON TABLES`, respectively, and that four different OCR systems, SYSTEM A, SYSTEM B, SYSTEM C, and SYSTEM D produce the hypotheses shown in Table 7 for those images. The number of field errors and the field error rates for these systems are as shown in Table 8.

image	SYSTEM A hypothesis	SYSTEM B hypothesis	SYSTEM C hypothesis	SYSTEM D hypothesis
1	DRIVES TRUCKS	DRIVES TRUCKS	DRIVES TRUCK	DROP FORGING
2	WAITS ON TABLES	WAITS TABLES	WAITS TABLES	WRITES TABLOIDS

Table 7: Hypotheses from four OCR systems for two images having DRIVES TRUCKS and WAITS ON TABLES as references.

	Contribution to field error $F_\epsilon(r_f)$			
image	SYSTEM A hypotheses	SYSTEM B hypotheses	SYSTEM C hypotheses	SYSTEM D hypotheses
1	0	0	1	1
2	0	1	1	1
field error rate	0.0	0.5	1.0	1.0

Table 8: Contribution to the field error $R_{f\epsilon}(r_f)$ for hypotheses of the preceding table, and resulting field error rates.

Notice that the field error rate for SYSTEM C, which is almost correct, is the same as that for SYSTEM D, which is completely wrong. Clearly, the field error rate cannot distinguish between hypotheses that will have no adverse effect on an application and those that will, even in the absence of fine distinctions. This is the reason for introducing the field distance rate as an alternative accuracy measure that complements the field error rate. In fact, the only reason for reporting the field error rate at all is that it is the most intuitive error measure and is, therefore, of general interest.

For this Conference, the field distance $R_{fd}(r_f)$ as a function of the field rejection rate r_f was defined as

$$R_{fd}(r_f) = \frac{C_\epsilon(r_f)}{C_c(r_f) + C_\epsilon(r_f)}, \tag{2}$$

where $C_c(r_f)$ is the number of characters in the field that are correctly classified when the hypothesis and reference fields are aligned as described in Section 6.3, and where

$$C_\epsilon(r_f) = C_d(r_f) + C_i(r_f) + C_s(r_f), \tag{3}$$

with $C_d(r_f)$, $C_i(r_f)$, and $C_s(r_f)$ representing the respective number of character deletion, insertion, and substitution transformations needed to convert each reference into the associated hypothesis according to some particular alignment of the two strings.

Therefore, before calculating the field distance between a hypothesis and its reference, it is necessary to align the hypothesis and reference to identify the characters that have been deleted from the reference and the characters that have been inserted into the hypothesis. This step is also a prerequisite for identifying the characters in the hypothesis that are correct and those have been substituted for characters occurring in the reference. String alignment

reference	WAITS ON TABLES
hypothesis	WRITES TABLOIDS
aligned ref.	WAITS ON TABLiiES
alignment	WsITddss TABLiisS
aligned hyp.	WRITddES TABLOIDS

Table 9: Example of an alignment of a hypothesis with its reference.

reference	WAITS ON TABLES
hypothesis	WRITES TABLOIDS
aligned ref.	WAITiS ON TABLiiES
alignment	WsITiSddd TABLiisS
aligned hyp.	WRITESddd TABLOIDS

Table 10: Example of an alternate alignment of a hypothesis with its reference.

is discussed in the next section.

6.3 String Alignment

This section discusses how the Levenstein distance is used to align to field hypotheses and references prior to scoring, and some of the problems with using the Levenstein distance for string alignment. The fact that many different strings alignments, some intuitively satisfying and some clearly wrong, can have the same Levenstein distance, is a problem. This problem is further complicated by the fact that is is often impossible to choose between equally intuitive alignments except by statistical arguments.

Tables 9 and 10 show alternative alignments of `WRITES TABLOID`, which is the SYSTEM D hypothesis for image 2 of Table 7, with `WAITS ON TABLES`, which is the reference for that image. Each lower case d, i, or s that appears in the alignments, aligned references, and aligned hypotheses represents a deletion, insertion, or substitution error, respectively.

Notice that there are 2 deletion errors, 2 insertion errors, and 4 substitution errors in the alignment of Table 9, but that there are 3 deletion errors, 3 insertion errors, and 2 substitution errors in the alignment of TABLE 10. Therefore, the contribution of `WRITES TABLOIDS` to $C_e(r_f)$ in eq. (3) is $2 + 2 + 4 = 3 + 3 + 2 = 8$, for either alignment. Notice also that the alignment of Table 10 has one more correct character (the `S` in `WAITS`).

The first step in obtaining both alignments was to calculate the Levenstein distance matrix [?] [?] [?] with 5, 3, and 1 set as the penalties for deletion, insertion, and substitution transformations from the reference `WAITS ON TABLES` to the hypothesis `WRITES TABLOIDS`. The second step was to trace a minimum penalty path backwards through the Levenstein matrix (backtracking). For Table 9, the path corresponding to a substitution was chosen instead of the path corresponding to either an insertion or a deletion whenever choosing to move in the direction of a substitution during backtracking would give an equal penalty path (a tie) through the matrix. For Table 10, a substitution was chosen instead of an insertion,

WsITddss TABLiisS
WsITdsds TABLiisS
WsITdssd TABLiisS
WsITsdsd TABLiisS
WsITssdd TABLiisS
WsITddss TABLisiS
WsITddss TABLsiiS

Table 11: Some of the equal penalty alignments of WRITES TABLOIDS with WAITS ON TABLES when the deletion, insertion, and substitution penalties are 5, 1, and 3, respectively.

reference	SOLD OLD YELLOW RUSTED STEEL
hypothesis	SOLD STEEL
alignment 1	SOLDddddddddddddddddd STEEL
alignment 1	SddddOdddddLddddddddD STEEL

Table 12: Example of two equal penalty alignments that are not equally intuitive.

but a deletion was chosen instead of a substitution, whenever these choices occurred.

For any given set of deletion, insertion, and substitution penalties, the Levenstein algorithm allows creation of a matrix from which all of the equal penalty alignments may be obtained by backtracking. Table 11 gives some of the other equal penalty alignments for the d = 5, i = 1, s = 3 case discussed above. The total penalty for all of these alignments is 24. Notice however, that if the penalty for substitutions were changed from s = 3 to s = 4, then the total penalty for the alignment in TABLE 9 would be increased to 28, while the total penalty for the alignment in Table 10 would be increased to only to 26, despite the fact that both of these alignments make the same contribution to $C_\epsilon(r_f)$.

Even though the alignments of Tables 9 and 10 make equal contributions to $C_\epsilon(r_f f)$, they do not make equal contributions to the field distance rate $R_{fd}(r_f)$ because they have different numbers of correct characters $C_c(r_f)$. If the sum of the penalties for deletion and insertion is equal to twice the penalty for substitution, different alignments with the same Levenstein distance may have different numbers of correct characters. For each extra correct character, the longer alignment has one more deletion, one more insertion, and two less substitution errors. The net effect is that the alignment of Table 9 has a field distance rate of 0.4706, while that of Table 10 has a field distance rate of 0.4444. In some cases this produces a more plausible alignment, but in other cases it produces a much less plausible alignment. This problem is illustrated in Table 12, which compares two alternative alignments of SOLD OLD YELLOW RUSTED STEEL and SOLD STEEL. As far as the Levenstein distance is concerned, each of these alignments is equally good, and whichever one is chosen for scoring will depend upon the choices made during backtracking.

Both alignments in the example of Table 12 result in the same Levenstein distance and the same contribution to the field distance rate as defined in eq. (2). However, this need not be the case when words are torn apart like SOLD to acheive an alignment. This is illustrated

reference	COLD ROLLED STEEL PLATE
hypothesis 1	COLD ROLLED STEEL
hypothesis 2	COLD ROAST EEL PLATE
aligned ref. 1	COLD ROLLED STEEL PLATE
alignment 2	COLD ROLLED STEELddddddd
aligned hyp. 2	COLD ROLLED STEELddddddd
aligned ref. 2	COLD ROLLED STiEEL PLATE
alignment 2	COLD ROsddd STiEEL PLATE
aligned hyp. 2	COLD ROAddd STiEEL PLATE

Table 13: Example of a very incorrect alignment having a lower Levenstein and field distance that a more correct alignment.

in Table 13. Despite the fact that COLD ROLLED STEEL gets three of the four words of COLD ROLLED STEEL PLATE correct while COLD ROAST EEL PLATE gets only two words correct, it is the latter that has the lowest Levenstein penalty and the most correct letters.

In summary, the problem with all of the ambiguities associated with using the Levenstein distance and backtracking as an alignment algorithm is the following. Suppose that two OCR systems have comparable accuracy, but tend to make different types of errors. It is possible under these conditions that one system will score higher than the other for some choices of Levenstein distance penalties and backtracking strategies, but lower for other choices. Since these choices are quite arbitrary, there is no objective way to choose one choice over another, and there are far too many combinations (proportional to the square of the number of letters in the alignment) to allow scoring all possible choices and averaging the results.

More complex (and time consuming) alignment algorithms that can be used to minimize the field distance rate as defined in eq. (2) have been described.[?] Even these, however, do not solve the problem illustrated in Table 13, which is associated with the actual definition of the field distance rate. Possible solutions to this and other problems are discussed in the next section.

In scoring the Second Conference tests the field distance rate was calculated from eqs. (2) and (3) from the alignment of a hypothesis with its reference obtained by backtracking through the associated Levenstein matrix. The penalties for deletion, insertion, and substitution transformations of the reference into the hypothesis were 5, 1, and 3 for generation of the Levenstein matrix, and the order of priority for breaking backtracking ties was correct, followed by substitution, followed by insertion, followed by deletion.

6.4 Field Distance Rate: Problems and Generalizations

The preceding section discussed some some of the problems associated with the use of the backtracking through the Levenstein matrix to obtain an alignment for use in calculating the field distance rate. Even with these problems, and others mentioned in this section, the field distance rate appears to be superior to the field error rate as an OCR error measure.

	Contribution to field distance $C_t(r_f)$			
image transformation	SYSTEM A hypotheses	SYSTEM B hypotheses	SYSTEM C hypotheses	SYSTEM D hypotheses
1				
deletions	0	0	1	1
insertions	0	0	0	0
substitutions	0	0	0	9
correct	13	13	12	3
2				
deletions	0	3	3	2
insertions	0	0	0	2
substitutions	0	0	0	4
correct	15	12	12	9
field distance rate	0.0000	0.1017	0.1429	0.6000

Table 14: Contribution to the field distance $R_{fd}(r_f)$ for hypotheses being used as examples in this Chapter, and resulting field distance rates.

This is illustrated in Table 14, which lists the contributions to the field distance and the field distance rates for SYSTEMS A, B, C, and D of Table 7. Comparison of the data of this table with that of Table 8 for the field error rate shows that the field distance rate gives much more intuitively satisfying rankings of the results from different OCR systems.

Even so, the field distance rate, as defined in eq. (3), still suffers from the problem illustrated in Table 13 that was mentioned in the last section. It also suffers, but with much less severe effects, from the same type of problem affecting the field error rate that prompted the introduction of the field distance rate in the first place. While it is able to distinguish gross differences between the suitability of different hypotheses for a given reference, the field distance rate as defined in eq. (3) is not able to distinguish finer distinctions. For example, the hypotheses `WRITER` and `WAITOR` are equally good approximations to the reference `WAITER` as far as the field distance rate is concerned, but not as far as most applications are concerned. For most applications, a `WAITER` is something completely different from a `WRITER`, whereas `WAITOR` is easily understood as a misspelling of `WAITER`.

Various generalizations of the field distance rate can be defined to solve these problems in principle, but problems with the details of the implementation remain. For example,

$$R_{fd}(r_f) = \frac{C_g(r_f)}{C_c(r_f) + C_g(r_f)}, \tag{4}$$

where

$$C_g(r_f) = \sum_{i,j}^{m,n[i]} Q(t_i[j], w_i[j]) + P(a_j) \tag{5}$$

is a generalized version of the field distance defined in eq. (3).

j	$t_2[j]$	$w[j]$
1	(W,W)	WAITS
2	(A,R)	WAITS
3	(I,I)	WAITS
4	(T,T)	WAITS
5	(,E)	WAITS
6	(S,S)	WAITS
7	(,)	
8	(O,)	ON
9	(N,)	ON
10	(,)	

Table 15: First ten members of the ordered set of transformations $t[j]$ of the reference WAITS ON TABLES into the hypothesis WRITES TABLOIDS, and first ten members of the current word $w[j]$, both as a function of the alignment position j.

In eq. (5), $P(a_i)$ represents a generalized penalty that depends upon the alignment $a_i = a_i[1]a_i[2]...a_i[n[i]]$ of the hypothesis for the ith field, $i = 1, 2, ..., m$, with its reference, and $Q(t_i[j], w_i[j])$ represents a generalized penalty that depends upon the transformation $t_i[j]$ and current reference word $w_i[j]$ at position $j = 1, 2, ..., n[i]$ of the alignment a_i. The transformation $t_i[j]$ is the ordered pair $(x[j], y[j])_i$, where $x[j]$ is the reference character that is converted into the hypothesis character $y[j]$ at alignment position j of alignment a_i. Table 15 gives examples of $t_2[j]$ and $w_2[j]$ for alignment positions $i = 1, ..., 10$ of the transformation of the reference `WAITS ON TABLES` into the hypothesis `WRITES TABLOIDS`.

The generalized penalty $Q(t_i[j], w_i[j])$ allows each character transformation at each location in every different word of a test to be scored differently depending upon the impact on the application of that particular transformation at that particular location. The generalized penalty $P(a_i)$ allows alignments that tear words apart, as illustrated in Tables 12 and 13, to be penalized more than alignments that preserve words. Ref. [?] and references therein describe a generalization of the Levenstein algorithm that can be used to minimize the generalized field distance rate of eq. 4 over all possible alignments.

Suppose that $P(a_i) = 0$ for all fields indexed by i, and $Q(t_i[j], w_i[j]) = 1 - \delta[x, y]$, where $\delta[x, y]$ is the Kronecker delta (equal to 1 when character x = character y, but zero otherwise). Then, $C_g(r_f) = C_\epsilon(r_f)$. Therefore, the generalized field distance is identical to the field distance of eq. 3 in this case.

Unfortunately, just because we are able to propose definitions of a generalized field distance does not mean that we actually know how to tailor the generalized penalties $Q(t_i[j], w_i[j])$ and $P(a_i)$ to any given application. It is clear that lower penalties will be associated with transformations involving character strings like `S`, `ES`, and `ING` at the end of nouns than with the `A` in `WAITER`. It is also clear that the penalties associated with all transformations of words like `ON`, `A`, and `THE` will be relatively low. Finally, it appears that a language model for a given application will be necessary to tailor the penalty matrices to that application, but the details how to do it are not obvious.

6.5 Conclusions

Some conclusions can be drawn from the material presented in this Chapter.

1) The field error rate is not a good measure of OCR performance on words and phrases because it cannot distinguish hypotheses that are almost correct from hypotheses that are completely wrong.

2) The field distance rate as defined in eq. (2) in this Chapter is a better measure of OCR performance on words and phrases because it can usually distinguish between hypotheses that are almost correct and those that are completely wrong. Occasionally, however, coincidental similarities between a completely wrong hypothesis and its reference will produce a misleadingly low error rate.

3) The field distance rate is not an optimum measure of OCR performance on words and phrases because it cannot make fine distinctions between different hypotheses, and cannot be tuned to capture the requirements of different applications.

4) It is probably possible to generalize the field distance rate so that it can make fine distinctions between different hypotheses and be tuned to capture the requirements of any particular application, but a fairly sophisticated language model will probably be required, and the details of how to carry out this project are not obvious at the time of the writing of this report.

7 Voting Systems

Patrick J. Grother

Described elsewhere in this report are performance results for all the entrants in the Second Census OCR Conference. Three particpants submitted classifications markedly superior to the average, though such systems still fail to classify almost two out of five fields correctly at zero rejection rate. A frequently asked question in pattern recognition is whether competing recognition systems fail to classify the same examples. Given the large error rates here, the apparent complexity or the problem and the diversity of algorithms used, there seems to be some suggestion that recognition failure between systems may be different and therefore exploitable. However for a given image, recognition failure may not be systemic; rather it could be a property of the imaged writing itself. In the systemic case it is possible that different algorithms will not universally fail on a given field, and complementarity of the classifiers will help. If however the handwriting is unreadable then no gain can be expected. Generally both types of error are present.

The systems used to form the voting systems given below are ERIM_0, IDIAP_2, and CGK_2, and only those fields imaged from paper are considered. An earlier effort, presented at the conference is not discussed here since one of its voting members, CGK_3, was found to have used dictionaries augmented by the real reference phrases, which are generally unavailable *a priori*, at least for the Census application.

This Chapter describes elementary voting systems using both the hypotheses *and* the confidences of the contributing systems. However the use of confidence values is problematical; each voting system generally derives confidence values on arbitrary and utterly disparate criteria obtained from the recognition and dictionary retrieval algorithms. This inconsistency makes sensible voting schemes more difficult to realize. Although the regulations for participation in this conference stipulated that confidences must lie on the range [0.0,1.0] there was certainly no specification of the *distribution* of the supplied confidences. The relevant procedure here is to transform confidences to yield values with some fixed distribution.

7.1 Normalization of Confidences

Figure 13 shows the very different confidence distributions of the three voting systems. A simple transformation is applied to realize uniformly distributed confidences on the range [0.0,1.0]. If a deviate x has some arbitrary distribution $p(x)$ then by applying some function $y = y(x)$ a distribution $p(y)$ is obtained. If we further specify that y should be uniformly distributed

$$p(y)\,dy = \begin{cases} dy & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

then by transforming x according to the fundamental transformation law of probabilities

$$p(y)\,dy = p(x)\,dx$$

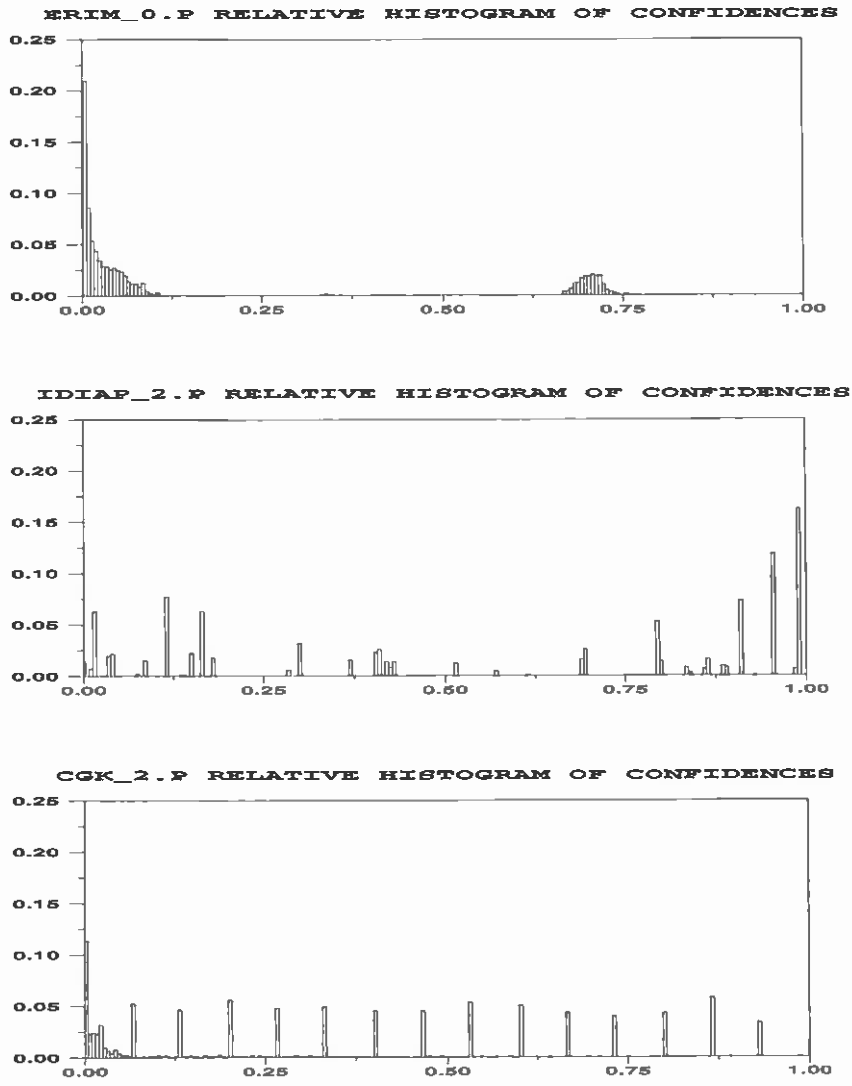


Figure 13: Comparison of distribution of confidence values generated by three OCR systems.

3 Voters	ERIM_0	IDIAP_2	CGK_2	RAW	NORM
Field Error	39.7	42.1	50.5	36.4	34.5
Field Distance	18.7	22.0	24.6	18.5	17.0

Table 16: Performance of three systems and two voting systems based on them.

2 Voters	Reject %	ERIM_0	IDIAP_2	RAW	NORM
Field Error	0	39.7	42.1	39.3	34.4
Field Distance	0	18.7	22.0	20.6	17.3

Table 17: Performance of two systems and two voting systems based on them.

quickly yields the solution for y as being just the indefinite integral of the original distribution.

$$y = \int p(x)\, dx$$

Thus a confidence x is transformed to a value y which is just the area to the left of x under the original distribution. This area is the value of the cumulative relative distribution.

The normalization transformation was applied to the confidences of the three systems. In the cases of ERIM_0 and IDIAP_2, approximately uniform confidence distributions were sucessfully obtained. However the CGK_2 system supplied confidences obtained from commercial hardware representing values in 4 bits, meaning that mostly only 16 discrete values were submitted for the confidences of the 9000 fields. This is shown in Figure 13. One consequence of this is that rejection is random over the examples hypothesized with identical confidences. Another consequence is that confidence normalization is not numerically possible. One solution is to order the confidences around the fixed value, by adding small unique amounts of noise correlated to the hypothesis string lengths. The only constraint being that the maximum perturbation of the confidences should be much less than 2^{-4}.

7.2 Voter Systems

With three contributing systems the decision rules are simple. If all voters agree on the hypothesis then there is no other choice. If two of three agree then the majority's hypothesis is used. In both of those cases the scalar confidence value is taken as the largest of the majority's confidences. If all three disagree, then that hypothesis with the largest confidence is used. Table 16 shows the performances of the three contributing systems as reported elsewhere, and of the raw and normalized confidence voting systems.

If the weakest of the three voting systems, CGK_2, is discarded an even simpler two voter system is obtained. The decision is to use the output of the system that asserts its hypothesis most confidently whether the two systems agree or not. The use of unnormalized confidences is clearly futile for a two voting system: Table 17 shows that field distance deteriorates toward the level of the inferior system's (IDIAP_2) because it happens to have the higher distribution of raw confidences.

When the normalized confidences of the ERIM_0 and IDIAP_2 entrants are used, the voting

	ERIM_0	IDIAP_2	CGK_2	NORM	GAIN
Reject for Field Distance < 1%	55	74	84	66	-11
Reject for Field Distance < 3%	45	50	61	38	7
Reject for Field Error < 10%	44	48	67	36	8

Table 18: Rejection percentages required to attain given field distance and field error rates.

system with best performance is obtained. The same large improvement in zero rejection performance is obtained using two voters rather than three, as shown in Figs. 14 and 15. The gains in performance at realistic rejection rates are more relevant to a real classification system.

The graphs in Figs. 16 and 17 show the additive (VOTER-ERIM_0) and the multiplicative (ERIM_0/VOTER) increases in the performance of the voting system compared to the better lone system as a function of rejection rate. On both field error and distance measures the voting system increasingly outperforms the ERIM_0 system until 43% of the hypotheses are rejected. The improvement in the field distance remains superior until a higher rejection rate than field error improvement because hypothesis strings can still be improved by voting even when neither system is classifying the field entirely correctly.

However the gain statistics can be deceptive. A more economically pertinent question for the Census application is what additional percentage of the hypotheses can be accepted using a voting system while maintaining a constant tolerable error level. If one accepts that the human-key-data-entry field distance rate of the dicennial census is 1.6% then what percentage of ERIM_0's and the voting system's hypotheses must be rejected to achieve some nominally lower rate like 1%? What are the reject rates for a 3% tolerable field distance, or for a 10% field error rate? Table 18 gives the required rejection rates to achieve various recognition criteria. The "NORM" column refers to the normalized ERIM_0 and IDIAP_2 voting system and "GAIN" is the improvement of that system over ERIM_0.

Although performance gains are available at low rejection levels, the first line of the table is noteworthy. The best voting system is *worse* than the ERIM_0 system at realistic rejection levels; 11% more hypotheses must be given to human key-data-entry personnel to achieve a 1% field distance rate on the accepted classifications.

The implication of this negative result is that the method used for combination of hypotheses, including their confidence normalization, is partially retrograde. The ordering of the confidences according to the indefinite integral of the confidence distribution takes no account of the *value* of that confidence, in terms of its relation to the probability of correct classification of a field. This can only be determined if the by-field performance statistics are available. For a fair voting system to be created, the confidences and hypotheses of contributing systems would have to be available from some training set. Aside from partitioning the testing materials used for the conference, such data is not available.

7.3 Conclusions

The voting system offers both improvement and degradation of recognition depending on the performance criteria used. These criteria are likely to be economically determined. The voting systems described here are not sophisticated, and more research into trainable voting systems is necessary. It should be possible to develop a confidence-normalizing method that can be optimized on the training data. If so, it should produce results that are no worse than the best individual system at all rejection rates, with greater improvement at targeted rejection rates.

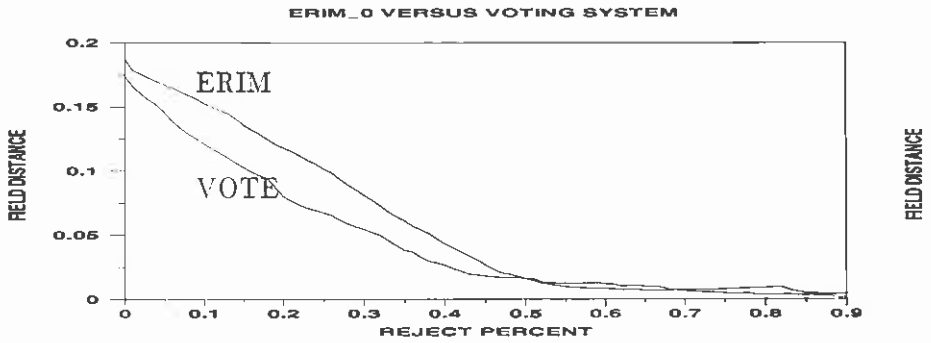


Figure 14: Comparison of field distance rates for VOTER system and ERIM_0 system.

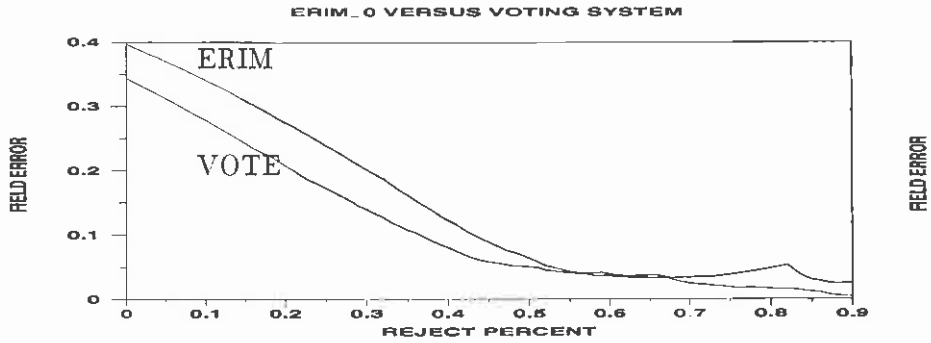


Figure 15: Comparison of field error rates for VOTER system and ERIM_0 system.

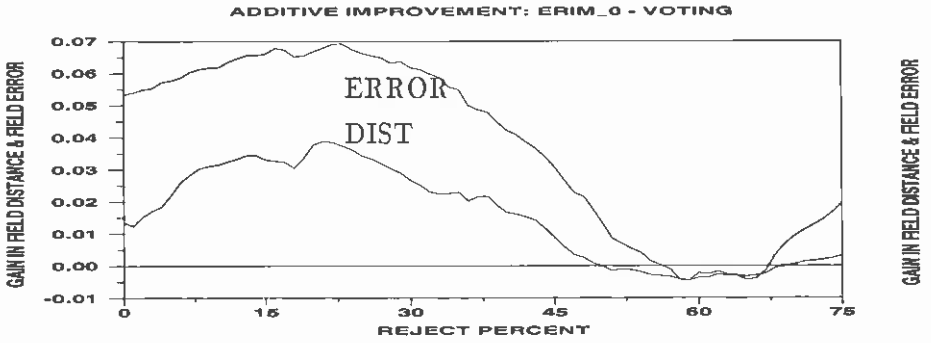


Figure 16: Difference between field error and distance rates for VOTER system and for ERIM_0 system.

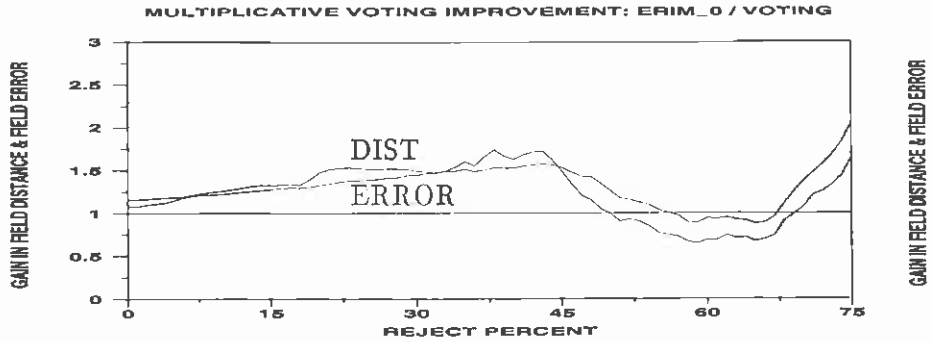


Figure 17: Ratio of field error and distance rates for VOTER system and for ERIM_0 system.

8 Dictionary-based correction of raw OCR results

Jon Geist

8.1 Introduction

Some of the OCR systems in the Second Census OCR Systems Conference did not produce field hypotheses before attempting dictionary-based corrections, and others did. Some of the latter were submitted to NIST for scoring. Such hypotheses will be referred to as raw OCR results to distinguish them from results that have been obtained after language-model or dictionary-based context correction.

It is interesting to compare the raw OCR results obtained at the Second Conference with the results obtained at the First Conference to assess the relative importance and difficulty of different OCR tasks. It is also interesting to compare the raw OCR results from the Second Conference with the results obtained after dictionary-based correction to focus on the importance of the correction. This chapter describes these comparisons, and briefly describes the method used by NIST for dictionary-based correction of raw OCR.

8.2 Raw OCR Results for the Second Conference Test

IBM, NIST and the University of Bologna submitted raw OCR results as well as the results obtained after dictionary-based correction. Figures 18, 19, and 20 at the end of this Chapter compare the on-time raw OCR results with the on-time dictionary-corrected results for these three organizations, respectively. Neither NIST's nor the University of Bologna's late submissions showed significant improvement over their on-time submissions, so they are not shown here. On the other hand, the IBM late submission was a significant improvement, and it is compared in Fig. 21 with the raw results from which it was obtained (the same on-time raw results shown in Fig. 20).

The raw OCR results submitted by IBM and the University of Bologna were also run through the NIST dictionary-based correction methods to compare the NIST methods with those of the other two organizations. Those results are also shown in the figures mentioned above.

Perhaps the most significant results in Figs. 18 through 21 are the very high error rates for the raw OCR. IBM, NIST and the University of Bologna all participated in the First OCR Systems Conference. Table 19 compares the field distance rates for the raw OCR results obtained on words and phrases at the Second Conference with the character error rates obtained on isolated lower case letters at the First Conference for these three organizations. The comparisons are shown for zero and fifty percent field rejection rates.

The field distance rate is the appropriate generalization of the character error rate for use with words and phrases. If the only errors made in the Second Conference were substitution errors, then the field distance rate would be equal to the character error rate, but the field distance rate is more general in that it also accounts for deletion and insertion errors.

ORGANIZATION	50% Rejection Rate		0% Rejection Rate	
	2nd Conf.	1st Conf.	2nd Conf.	1st Conf.
IBM	0.45	0.0183	0.50	0.1542
NIST	0.47	0.0458	0.54	0.2029
U. Bologna	0.52	0.0287	0.55	0.1548

Table 19: Comparisons of field distance rate for raw OCR results submitted by three organizations in the Second OCR Systems Conference with lower-case letter error rates for isolated character OCR results submitted by the same three organizations in the First Conference.

ORGANIZATION	50% Rejection Rate		0% Rejjection Rate	
	Before	After	Before	After
IBM (ON TIME)	0.45	0.38	0.50	0.45
NIST	0.47	0.23	0.54	0.46
U. Bologna	0.52	0.40	0.55	0.43
IBM (LATE)	0.45	0.16	0.50	0.40

Table 20: Comparisons of field distance rates of raw (Before) and dictionary-corrected (After) OCR results submitted by three organizations in the Second OCR Systems Conference.

The field distance rates at zero rejection rate for the Second Conference are roughly a factor of three greater than the character error rates for the First Conference. These differences are striking, and the differences at 50% rejection rate are even larger. This strongly suggests that the OCR of isolated (properly segmented characters) is not the accuracy limiting step in the recognition of the handprint words and phrases from the 1990 Census.

There was probably more cursive writing in the Second Conference test than in the First. This, however, does not explain the differences shown in Table 19. Only 6% of the characters in the Second Conference test were cursive. Even if there was no cursive in the First Conference test and all of the cursive in the Second Conference test was recognized incorrectly, the error rate would grow by less than a factor of 1.5, not by a factor of 3.

It is likely that the error limiting step in the Second Conference was associated with the proper segmentation of words and phrases into characters or character fragments. This is almost certainly true for the systems that carried out segmentation separately from recognition, but it is also likely for the other systems as well. Poor segmentation precludes accurate recognition unless special measures are taken to compensate for under and oversegmentation. Most of the participants described special measures that they used to compensate for these problems, and most of these involved intentional avoidance of under-segmentation, and means for dealing with the resulting oversegmented character fragments.

Table 20 compares the field distance rate for the raw OCR (before dictionary-based correction) with field distance rate after dictionary-based correction for the same three organizations. The improvements at zero rejection rate are very modest. The improvements at 50% rejection rate for two of the systems are somewhat better. Note that the top scoring systems in the Second Conference had substantially lower field distances at zero rejection rate as well

as greater improvements with increasing rejection rate than the "After" results in Table 19.

Also, note that the late IBM "After" results, which are substantially improved over the on-time IBM "After" results, were obtained from the same raw OCR data. The only difference is the dictionary-based correction method. This level of improvement illustrates that many of the participants were still making substantial changes to their systems after the completion of the on-time test period. Chapter 3 showed that there is still room for improvement in the state of the art in the recognition of isolated characters beyond that demonstrated in the First Conference. For both of the reasons just mentioned, and for other reasons mentioned elsewhere in this report, it seems likely that substantial improvements beyond those obtained in the Second Conference are also possible.

8.3 Future Availability of Test Materials

It was decided at the Conference to allow the participants to submit new results on the Second Conference test by anonymous ftp to NIST for scoring. By the time that this report is available, the NIST Office of Standard Reference Data should be selling the Second Conference training and test materials as Standard Database 11 (SD 11, training: references and microfilm images on a CD-ROM); Standard Database 12 (SD 12, training: references, paper images, and microfilm images on a CD-ROM); and Standard Database 13 (SD 13, test: paper and microfilm images on a CD-ROM, reference on a separate disk). Anyone who purchases the test data is free to submit the results to NIST for scoring. Submissions can be made as follows in the format described in Appendix B. First, call Stan Janet at (301) 975-2916 to get a system (tar file) name. Then,

- ftp sequoyah.ncsl.nist.gov or ftp 129.6.61.25
- cd incoming
- put /etc/motd *filename*

Sample images are available at the same ftp site to help interested parties to decide whether or not they want to purchase one or more of the Special Databases. Sample images from paper and their associated references can be found in

- ind_occ/data3/d00

Sample images from microfilm and their associated references can be found in

- ind_occ2/data/d00

More sample images from microfilm can be found in

- ind_occ/data/d00 or ind_occ/data/d01

but only one of the mis files in these directories has an associated reference file.

It was also decided at the Second Conference meeting to provide the references for the test data on SD13 at the same ftp site so that the participants could examine the types of errors their systems had made. The references for the test can be found in

- ocr_conf_2/refs_paper.tar.Z

- ocr_conf_2/refs_microfilm.tar.Z

Since the references for the test are being distributed on the ftp site, the significance of future submissions of test results will not be the same as for the original test. The only reason for NIST to score these submissions is to assure uniformity of scoring and to keep NIST apprised of the improving state of the art. If significant improvements in the state of the art are observed, NIST may run a new test with materials that were reserved for this purpose. Anyone who receives this report as a result of a direct mailing from the Image Recognition Group at NIST will also receive a notice by letter of such a test, if and when one is planned.

8.4 Dictionary-based Correction Methods

The character-recognition methods used by NIST in the Second Conference have been described previously[?] [?], but the dictionary-based correction methods have not. Since the latter may not be described elsewhere in the future, and since they are not very different from many of the other methods used in the Conference, they will be briefly described here.

The NIST dictionary-based correction system is not too different from the ERIM system, even though the NIST system was designed to work with raw OCR and the ERIM system was not. The ERIM OCR system intentionally over-segments words and phrases into character fragments, and does not produce a string of hypothetical characters for each field as an output from the character recognition process. Instead, it produces a string of confidence vectors for a subset of the unions (combinations) of the character fragments (including isolated characters) that were obtained from the intentional oversegmentation of each field. Each confidence vector consists of an ordered set of confidences, one for each possible character class. Clearly, this string of confidence vectors cannot serve as a hypothesis for the string of characters in the field. On the other hand, it can be used very effectively in dictionary-based correction as shown by the ERIM results.

The NIST method was designed to work on character hypotheses rather than character fragments, and with only the highest confidence hypothesis rather than a vector of confidences for all possible hypotheses. Nevertheless, it has enough in common with the ERIM method and most of the other methods to give some insight into how they function.

The NIST method required two dictionary searches. The first search was carried out on a bit-mapped, digraph-encoded version of a short phrase dictionary to pick out some likely candidate phrases from which to construct a much smaller dictionary for the second pass. This first-pass search was very fast due to the way it was implemented even though the entire dictionary was searched.

The first-pass search was followed by a second search through the candidate phrases selected on the first pass. The candidate phrase with the lowest Levenstein distance [?] [?] [?] was chosen as the field hypothesis. The first-pass method, which is covered by a US Patent that has been allowed but not yet issued, is similar to the method described in Ref. [?].

How the first-pass method works is illustrated in Table 21 for the phrase `LINE RUNNER`. This example uses an eight-character alphabet consisting only of the letters E, H, I, L, N, R, and

Alphabet	@ E H I L N R U
Phrase	@LINE@RUNNER@
Letters	@ L I N E @ R U N N E R
Digraphs	@L LI IN NE E@ @R RU UN NN NE ER R@
Bit-Map Coding of Digraphs	
Bit	@ E H I L N R U
@-byte	0 0 0 0 1 0 1 0
E-byte	1 0 0 0 0 0 1 0
H-byte	0 0 0 0 0 0 0 0
I-byte	0 0 0 0 0 1 0 0
L-byte	0 0 0 1 0 0 0 0
N-byte	0 1 0 0 0 1 0 0
R-byte	1 0 0 0 0 0 0 1
U-byte	0 0 0 0 0 1 0 0

Table 21: Example of bit coding of the digraphs in the phrase LINE RUNNER into eight bytes for rapid dictionary search.

U, and the symbol @, which is used to designate the beginning of the phrase, the end of the phrase, or a space between words in the phrase. Table 21 illustrates the rewriting of the phrase LINE RUNNER in terms of digraphs. The digraphs are mapped onto eight bytes (called the bit map) that represent the phrase as a bit-mapped code. One byte of the bit map is allocated for each possible first letter of a digraph, and one bit in each byte for each possible second letter of a digraph. Thus a 1 in some location in the eight-byte bit map indicates the presence of a particular digraph in the phrase, and a 0 indicates its absence.

All of the phrases in the short-phrase dictionaries were coded as illustrated in Table 21 into 27, 32-bit machine words with a 27 letter alphabet consisting of the 26 English letters and the @ symbol. Any other characters occurring in the phrases were deleted from the phrases before coding. All of this was done off-line, and the resulting dictionary of bit maps was stored for use during dictionary-based correction of the raw OCR.

During the correction process, each ASCII string that was obtained from the raw OCR process was coded as described above. The resulting bit map was combined 32-bit word by 32-bit word with each dictionary bit-map with the logical AND operation, and separately with the logical XOR operation. The 1 bits in all of the words in the bit map resulting from the AND operation (hits) were summed to determine how many digraphs were common between the raw OCR bit map and each dictionary bit map. The 1 bits in all of the bytes in the bit map resulting from the XOR operation (misses) were summed to determine how many digraphs were different between the raw OCR bit map and each dictionary bit map. The confidence c for each dictionary phrase was calculated as $c = 1 - m/(h + m)$ where h and m are the sums of the hits and misses for the phrase, respectively. A small number (15 for example) of the dictionary phrases having the greatest c were selected for use in the second search pass.

The second pass of the NIST dictionary-based correction uses the Levenstein distance method. This method is quite robust, but it is too slow to be used on the entire dictionary. For this

reason the digraph-based method is used first. Not only is the latter very fast when bit mapped as described above, but it is also a very local method. It looks only at which letters follow immediately after other letters, and has no global information except the phrase start and end digraphs. This locality complements the Levenstein method very well because the Levenstein distance, while using both global and local information, depends heavily on global information and is very tolerant of local errors.

Use of only the top-ranked phrase from the digraph-based first pass produced field distances that were worse than the raw OCR. Addition of the Levenstein distance-based second pass produced results that were essentially as accurate as those obtained from use of the Levenstein method on an entire phrase dictionary, but in a small fraction of the time. The fractions of the time devoted to the digraph search and to the Levenstein distance are somewhat dependent upon details of implementation. As a rough approximation, it takes about 5% as long to calculate the digraph confidence for two strings as it takes for the Levenstein distance. The digraph algorithm is also readily modified for IC chip implementation in a highly parallel architecture as described in a pending NIST patent.

The Levenstein distance method has been discussed in connection with its use in determining string alignments for scoring in Chapter 6, and is well documented in the references cited earlier in this section. In short, it takes two strings of characters as input, an unknown string and a string from a dictionary, and it uses user-assigned penalties to minimize the sum of the penalties for the alignment of the two strings through dynamic programming. These penalties are arbitrary, but they can be interpreted as negative logarithms of probabilities.

As pointed out in the viewgraphs in the Systems Summaries, the method used by ERIM can be interpreted as a Levenstein distance method. In this interpretation, the probability that a given union of character fragments represents a particular character in a dictionary phrase is approximated by the confidence of the recognition of the union as that character during the OCR process. It would be fairly straight forward to extend the NIST Levenstein distance method in this direction. However, this would be of little use without also developing satisfactory methods for oversegmentation and for selecting a reasonably small number of unions of character fragments for use in the Levenstein-like method.

The NIST method could also be extended to use different types of probabilities in a way that would complement an ERIM-like method. The prior probability of occurrence will be reasonably reliable for any raw OCR result that is encountered more than once during training. A generalized Levenstein distance penalty formula can be devised to account for insertions and deletions as well as substitutions. In essence it treats all such combinations of these errors as general transformations of small groups of letters into other small groups of letters. Given the existence of the prior probabilities mentioned above, all of the other probabilities necessary to build a Bayes classifier could be obtained from the same training data.

Such a classifier would be poor competition for an ERIM-like classifier on any raw OCR string for which a reliable approximation to the prior probability was not available from the training data. On the other hand, if a raw OCR string were represented a number of times in the training data, and if the training data were representative of the application, then

the estimate of the prior probability should be quite reliable with an uncertainty that can be estimated by standard statistical procedures. In this case, a Bayes classifier, with its statistical language model of the training data, could be more accurate than an ERIM-like classifier.

Notice in this connection that an ERIM-like classifier has no more language model than the dictionaries it uses. Its performance is ultimately limited by the relative accuracy of the probability estimates (confidences) generated by the OCR classification of the unions of character fragments. Actually it is quite tolerant of moderate errors in these confidences due to the constraints imposed by the dictionary words. Since ERIM-like methods and Bayes classifiers use very different types of information, their results should be somewhat complementary instead of completely redundant. Thus, when combined properly with a more sophisticated confidence normalization than that demonstrated in Chapter 7, they might produce substantially better results than either one alone.

8.5 Conclusions

A few tentative conclusions can be drawn from the material in this chapter:

1) The recognition of isolated (properly segmented characters) is not the accuracy limiting step at the current state of the art for OCR of handprint words and phrases.

2) Segmentation of fields into characters (or character fragments and the subsequent recombining of the fragments) is probably the accuracy limiting step at the current state of the art for OCR of handprint words and phrases.

3) One possible solution to this problem is to use sophisticated languages models or dictionary-based methods to correct the output of the character recognition step.

4) Over the next few years, the state of the art in the OCR of words and phrases for Census-like applications will probably be improved substantially beyond what was demonstrated at the Second Conference.

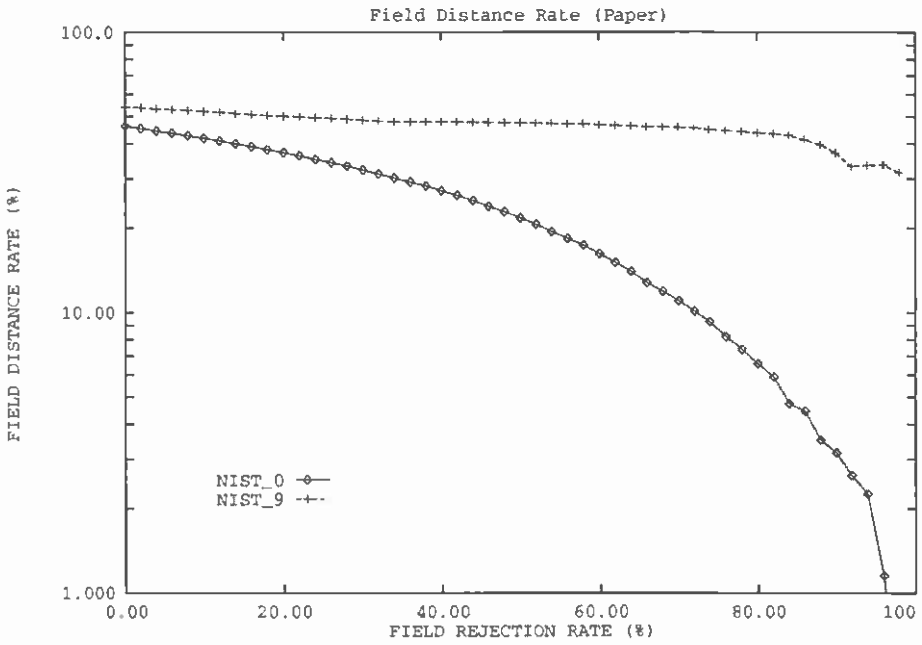


Figure 18: Comparison of on-time raw OCR results (NIST_9) with results of dictionary-based correction by NIST system (NIST_0).

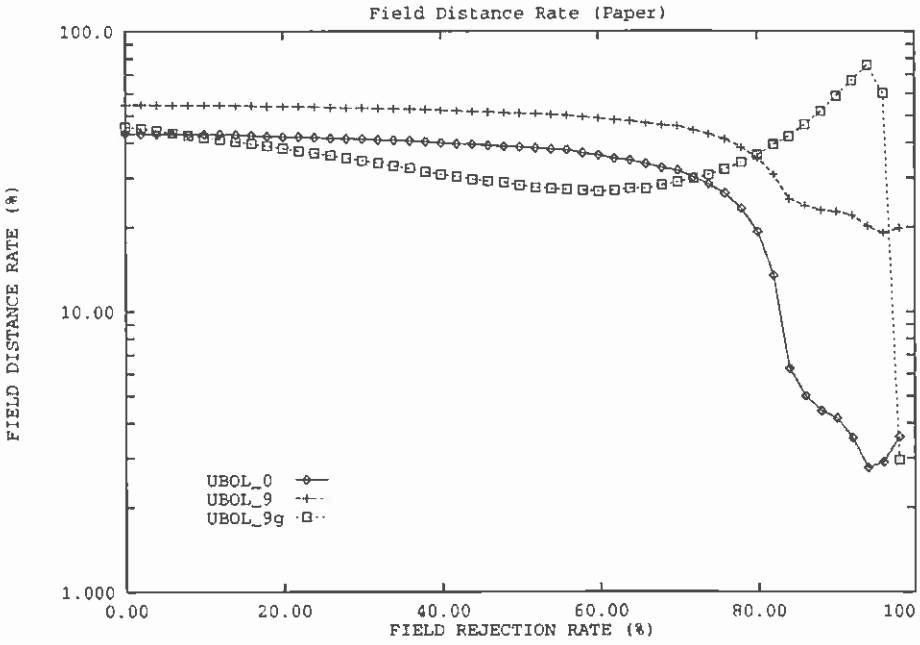


Figure 19: Comparison of on-time raw OCR results (UBOL_9) with results of on-time dictionary-based correction by UBOL system (UBOL_0) and by NIST dictionary-based correction system (UBOL_9g).

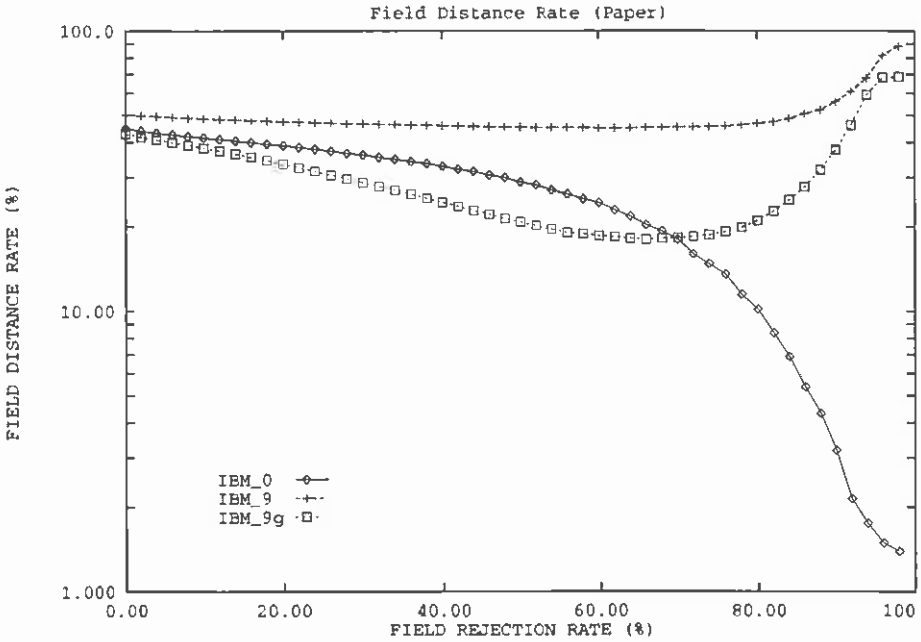


Figure 20: Comparison of on-time raw OCR results (IBM_9) with results of on-time dictionary-based correction by IBM system (IBM_0) and by NIST dictionary-based correction system (IBM_9g).

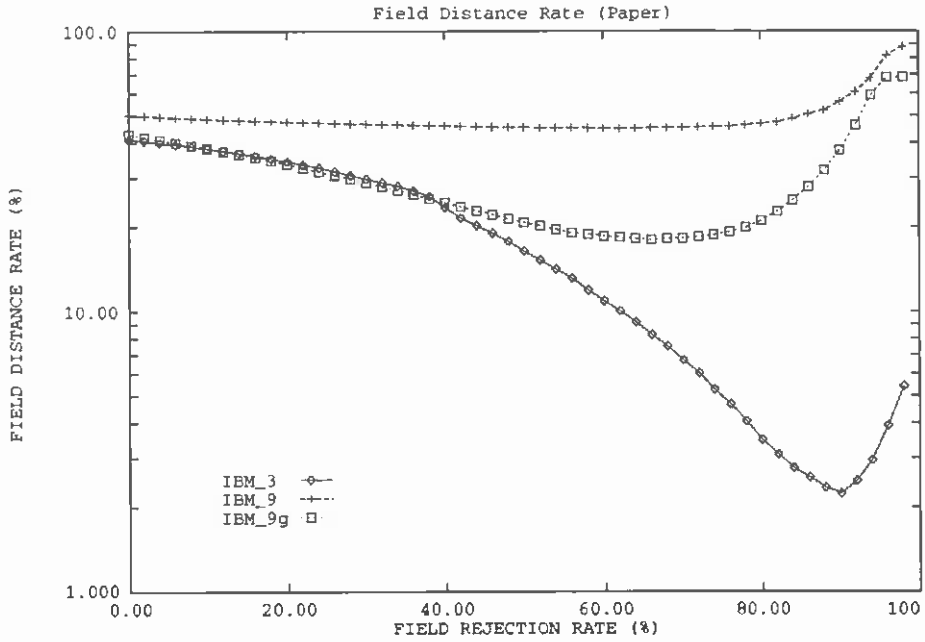


Figure 21: Comparison of on-time raw OCR results (IBM_9) with results of late dictionary-based correction by IBM system (IBM_3) and on-time NIST dictionary-based correction system (IBM_9g).

References

[1] R. A. Wilkinson et al. The first census optical character recognition systems conference. Nistir 4912, National Institute of Standards and Technology, Gaithersburg, MD 20899, 1992.

[2] M. D. Garris and R. A. Wilkinson. Nist special database 3. handwritten segmented characters. NIST, Advanced Systems Division, Image Recognition Group, Gaithersburg, MD 20899, 1992. page 6.

[3] S. N. Srihari, editor. *Computer text recognition and error correction.* IEEE, Piscataway, NJ 08854, 1985.

[4] No computer records of the reference results keyed from the 1990 Census returns were available for 32 out of 3000 mini-forms in the Conference test based on the images scanned from paper. (Had this been known before creation of the test CD-ROM, these mini-forms would not have been used.) These mini-forms were removed from the test when scoring the 1990 Census results, but not when scoring the machine results. As a result, 8904 fields were used in scoring the Census results, 8931 in scoring hybrid results, and 9000 when scoring the machine results. Examination of the left-out mini-forms suggests that they were typical of the remaining 99% of the miniforms in the Conference test, so the scores for the 1990 Census and hybrid results should be comparable to the machine results even though they are based on slightly different samples.

[5] R. A. Wilkinson, M. D. Garris, and J. Geist. Machine-assisted human classification of segmented characters for ocr testing and training. *SPIE*, 1906:208–217, 1993.

[6] R. Legault, C. Y. Suen, and C. Nadal. Difficult cases in handwritten numeral recognition. In H. S. Baird, H. Bunke, and K. Yamamoto, editors, *Structured Document Image Analysis*, pages 236–249. Springer-Verlag, NY, 1992.

[7] P. J. Grother. Cross validation comparison of nist ocr databases. *SPIE*, 1906:296–307, 1993.

[8] J. Geist and R. A. Wilkinson. Ocr error rate versus rejection rate for isolated handprint characters. *SPIE*, 1906:267–278, 1993.

[9] R. A. Wagner and M. J. Fisher. The string-to-string correction problem. *J. ACM*, 21:168, 1974.

[10] H. G. Zwankenburg. Inexact alphanumeric comparisons. *The C Users Journal*, page 127, May 1991.

[11] R. Valdés. Finding string distances. *Dr. Dobb's Journal*, page 56, April 1992.

[12] B. J. Oommen. Constrained string editing. *Information Sciences*, 40:267–284, 1986.

[13] P. J. Grother. Karhunen loève feature extraction for neural handwritten character recognition. *SPIE*, 1709:155–166, 1992.

[14] W. B. Cavnar and A. J. Vayda. Using superimposed coding of n-grams lists for efficient inexact matching. In *Proceedings of the Fifth USPS Advanced Technology Conference*, pages 253–267. Washington, DC, 1992.

A The Call For Participation

Jon Geist
225/B063/NIST
G'burg MD 20832
June 23, 1993
To whom it may concern:

Subject: 2nd Census OCR Systems Conference

In May of 1992, the U.S. Bureau of the Census and the National Institute of Standards and Technology (NIST) held a conference on optical character recognition (OCR) of hand-printed characters. Various tests focused on the recognition of individual characters, and the results were encouraging. However, most participants agreed that larger samples of handwriting were needed to tackle the more realistic problem of processing images of forms that contain unconstrained hand print.

Census and NIST are now preparing for a 2nd OCR Systems Conference to further advance this research. They are constructing a data base to contain: images of handwriting from the 1990 Census forms; ASCII text answers corresponding to each image; dictionaries with common words and phrases found in the answers; and some generic image processing software. Examples may be obtained from the ftp server at NIST (see enclosures for details); larger samples for training and testing will be distributed on a series of CD-ROMs.

The test will measure the ability of OCR systems to perform in a "worst case scenario". The images are being digitized from microfilm and may have lower quality than images created from the original paper questionnaires. Also, the 1990 Census questionnaires were designed for key entry data capture of handwriting, and therefore do not contain any design features that might facilitate machine recognition. Other tests using smaller samples of images lifted from the original paper will help to gauge the effect of image quality on OCR performance.

This conference is being organized by the following Committee:
Bob Hammond, Robert Creecy, Norman W. Larsen, Randy M. Klear, and Mark J. Matsko, US Bureau of the Census; Charles L. Wilson, Jon Geist, and R. Allen Wilkinson, National Institute of Standards and Technology; Jonathan J. Hull, Center of Excellence for Document Analysis and Recognition; Thomas P. Vogl, Environmental Research Institute of Michigan; and Christopher J. C. Burges, AT&T Bell Laboratories.

The Committee is chaired by Jon Geist, and the Conference and related activities will be run by NIST for the Committee with R. Allen

Wilkinson serving as the technical liaison for the Conference.

The approximate schedule for the research and Conference follows:

Sample data on ftp server	late June 1993
1st training data CD-ROM	early August 1993
2nd training data CD-ROM	early September 1993
Test data CD-ROM	November 1993
Test results due from participants	November 1993
Conference to announce/discuss results	February 1994
Publish report	June 1994

Seven enclosures provide more information about this research and how to participate in the conference; ENCLOSURE 0 provides an overview.

Sincerely,

Jon Geist

A.1 Enclosure 0: General Information

Enclosure 1 shows five miniforms cropped from 1990 Census Long Forms. The training CD-ROMs will contain between 10,000 and 50,000 of these images and the corresponding ASCII answer (reference) files. The test CD-ROM will be distributed in late October with between 10,000 and 50,000 more images but no reference files. The Conference is scheduled for February of 1994.

The training and testing materials will be distributed as NIST Multiple Image Set (MIS) files in a compressed IHEAD format on separate CDROMs. Each MIS file will contain five images (like on the enclosure). Notice that there are two different form types that are being extracted as miniforms. The most obvious difference between them is the relative location of the large black boxes and the right-most vertical line. However, the location of the answer field boxes relative to the large black boxes is also different.

Participants will be expected to return their classification results (hypothetical classifications) in a NIST MFS file format as illustrated by the d00f00.hyp file shown in Enclosure 2. The hypothetical classifications will be scored against references classifications like those in file d00f00.ref of Enclosure 2. The reference files will be included with the training materials, but not with the test materials. Participants may also return field level rejection or confidence files as indicated by the files named d00f00.rjx and d00f00.con in the SYSTEM_NAME subdirectory tree of Enclosure 2. Character level confidence and reject files will not be used

as discussed in Enclosure 3. More detailed file creation and naming specifications for the hypothesis, rejection, and confidence files will be included with the training materials. Finally, the training and test materials will have identical directory formats to make the transition from training to testing as smooth as possible.

Two measures of field accuracy will be used to score the hypothesis files submitted by Participants. The first of these is the field error fraction, that is, the fraction of the hypothetical fields that differ in any way from the reference fields. The second is a measure of the distance between the hypothetical and reference fields. Both will be plotted as a function of rejection fraction if either confidence or rejection files are submitted with the hypothesis files. Enclosure 3 provides more details about the scoring.

Enclosure 3 also makes some points about the contents of the images in Enclosure 1 that may affect participation and scoring. Beyond those points, however, it should be mentioned that the images in Enclosure 1 are among the best image quality in the hand-print region that have been produced to date. The poorest quality images will be removed from the training and test sets with an automated procedure, but there will be poorer quality images in what remains than shown in Enclosure 1, and poorer quality hand print as well. To get an idea of the range of image and print quality, you may obtain by anonymous ftp a representative sample of the types of images that will be sent for training and for testing from

sequoyah.ncsl.nist.gov, IP 129.6.61.25.

More details can be found in Enclosure 2. This site will also have a whatsnew subdirectory in which important dates and other important information will appear once they become available. Most conference activities will be run using the anonymous ftp site.

Enclosure 4 is the format for an application to participate in the 2nd Conference. Anyone who sends a signed copy of this letter to me before the training data is sent out for writing on CDROM will receive the training materials and test materials when they are sent out. The training data may be ready for writing on the CDROM by July 15. As soon as a firm date is set, it will be posted in the whatsnew subdirectory mentioned above.

The Committee reserves the right to distribute the training and test materials to anyone who returns the form letter after the date specified above, depending upon the availability of these materials. There may also be restrictions on the number of participants and colleagues from a single organization that can actually attend the meeting, and the Committee may request that a single participant from a single organization represent the entire organization and all of its systems.

Notice that the enclosed application format requires the applicant to sign a statement that he or she agrees to abide by the rules of participation stated in Enclosure 5. Finally, Enclosure 6 is a draft of a form for describing your system that will be sent with the test materials; it is to be returned at the same time as your test results. If an applicant fails to provide the information requested on this form (presumably because it is proprietary), that applicant will still be allowed to submit results, and attend the main meeting, but may not be allowed to attend sessions where participants who have provided this type of information describe their systems and their participation in the conference. The decision on this matter will be made on the basis of how many participants provide the requested information and how many do not. In case the number of applicants exceeds the capacity of the meeting facilities, the Committee reserves the right to limit attendance to those participants (and a number of colleagues to be decided) submitting results that exceed a performance threshold chosen to fill the meeting room. This decision will be made at the discretion of the Conference Committee who may, nevertheless, poll the participants for their feelings about this issue.

Comments or suggestions may be sent to me at

geist@magi.ncsl.nist.gov

or

Jon Geist, (301) 590-0932 (FAX).

Please do not suggest that we use any other format other than MIS and IHEAD as changes from these formats are not practical. Also, if there is a large volume of comments, you may not receive a personal reply to your comments, but they will be taken into account in the final plans for the Conference.

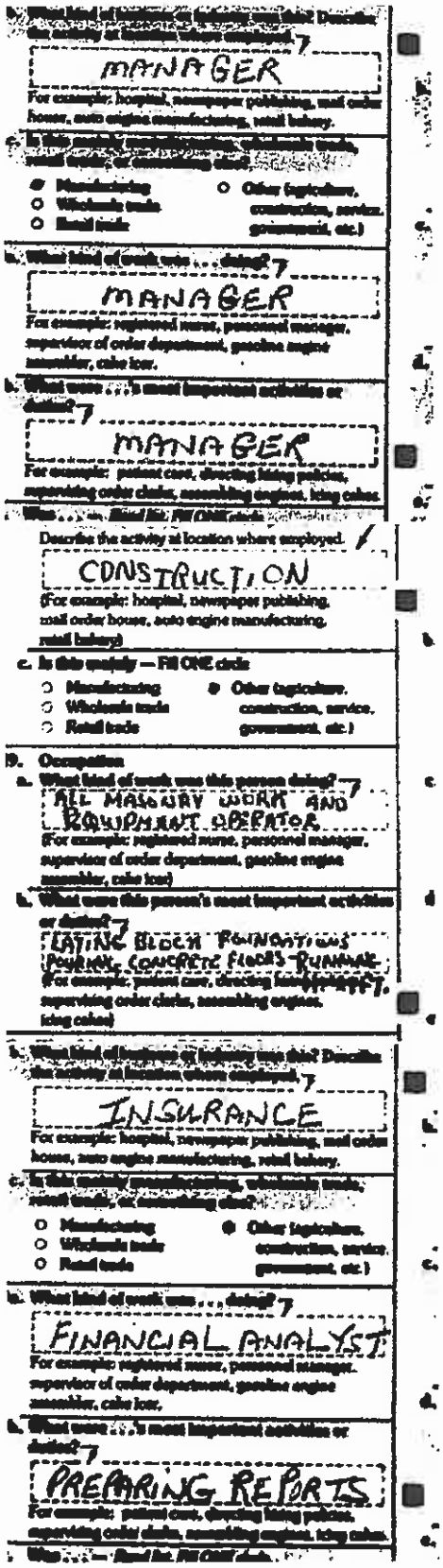
Requests for technical information about the data and other information at the FTP site should be addressed to

urt@magi.ncsl.nist.gov

or

R. Allen Wilkinson, (301) 590-0932 (FAX).

A.2 Enclosure 1: Example of mis file contents



MANAGER

For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, retail bakery.

c. Is this mainly manufacturing, wholesale trade, retail trade, or something else?

- ● Manufacturing
- ○ Wholesale trade
- ○ Retail trade
- ○ Other (agriculture, construction, service, government, etc.)

d. What kind of work was . . . doing?

MANAGER

For example: registered nurse, personnel manager, supervisor of order department, gasoline engine assembler, cake icer.

e. What were . . .'s most important activities or duties?

MANAGER

For example: patient care, directing hiring policies, supervising order clerks, assembling engines, icing cakes.

What . . . — Read the FILL ONE circle

Describe the activity at location where employed.

CONSTRUCTION

(For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, retail bakery)

c. Is this mainly — FILL ONE circle

- ○ Manufacturing
- ○ Wholesale trade
- ○ Retail trade
- ● Other (agriculture, construction, service, government, etc.)

9. Occupation

a. What kind of work was this person doing?

ALL MASONRY WORK AND EQUIPMENT OPERATOR

(For example: registered nurse, personnel manager, supervisor of order department, gasoline engine assembler, cake icer)

b. What were this person's most important activities or duties?

LAYING BLOCK FOUNDATIONS POURING CONCRETE FLOORS RUNNING EQUIPMENT.

(For example: patient care, directing hiring policies, supervising order clerks, assembling engines, icing cakes)

a. What kind of business or industry was this? Describe the activity at location where employed.

INSURANCE

For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, retail bakery.

c. Is this mainly manufacturing, wholesale trade, retail trade, or something else?

- ○ Manufacturing
- ○ Wholesale trade
- ○ Retail trade
- ● Other (agriculture, construction, service, government, etc.)

d. What kind of work was . . . doing?

FINANCIAL ANALYST

For example: registered nurse, personnel manager, supervisor of order department, gasoline engine assembler, cake icer.

e. What were . . .'s most important activities or duties?

PREPARING REPORTS

For example: patient care, directing hiring policies, supervising order clerks, assembling engines, icing cakes.

What . . . — Read the FILL ONE circle

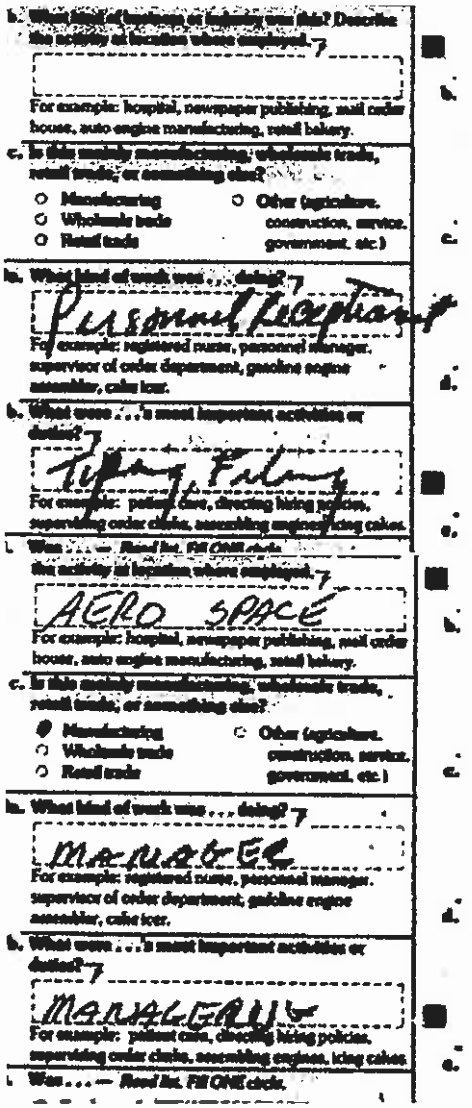


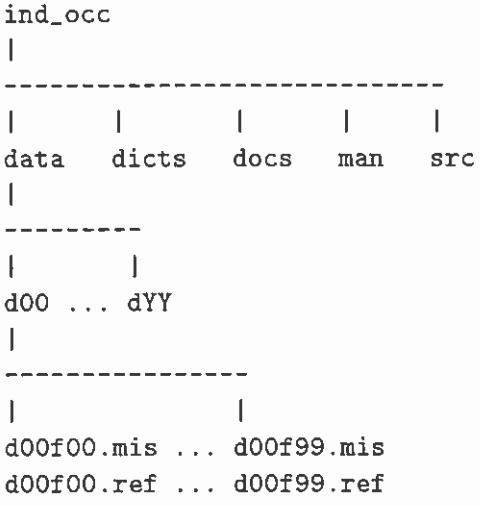
Figure 22: This page contains the last two miniforms in the example of a typical mis file having a total of five miniforms.

A.3 Enclosure 2: Examples of File Contents and Structures

A sample of the type of images that will be used in the testing phase of the Conference can be obtained by anonymous ftp from

sequoyah.ncsl.nist.gov, IP 129.6.61.25.

The images are in the files in the following directory structure:

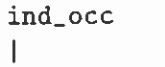


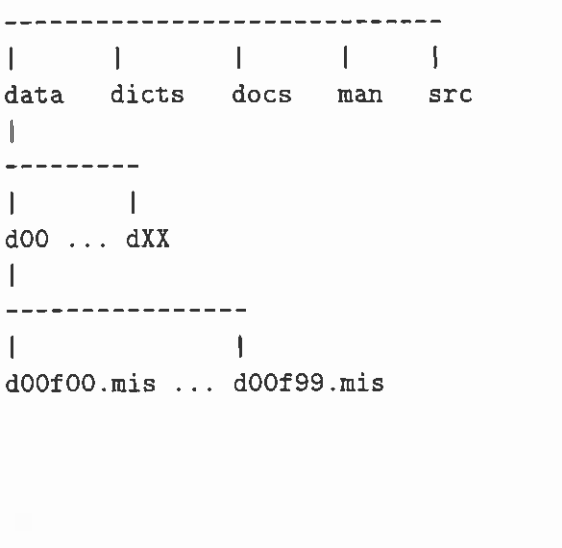
The directory ind_occ is an exact mirroring of what the CD-ROM discs will look like except that there may be more subdirectories dYY. The data directory will contain the image (MIS) files and the reference (REF) files. The dicts directory will contain the dictionaries that are discussed later. The src directory will contain all the source code needed to read the MIS files. This source code has been written and compiled on a SUN workstation using SUN OS 4.1.1 and works on that platform. We can not guarantee that this code will work on any other platform or operating system. The man directory contains the manual pages for the programs and routines supplied in the src directory.

At the time of this mailing, no reference files for the sample directory on the ftp site, except d00f00.ref, are available. The reference files will definitely be supplied with the training MIS files.

The five sample images shown on Enclosure 1 are in mis/d00/d00f00.mis.

Examples of the MIS file and related hypothesis subdirectory structures for the test CDROM and test results are shown below:





XX is a two-place digit that may be different for the training and test data than for the sample data, as was mentioned above in connection with the sample data at the anonymous ftp site.

The contents of d00f00.hyp (on left) and d00f00.ref (on right) are:

```
r00_f00 MANAGER                 r00_f00 MANAGER
r00_f01 MANAGER                 r00_f01 MANAGER
r00_f02 MHAGER                  r00_f02 MANAGER
r01_f00 CDNSTRUCTION            r01_f00 CONSTRUCTION
r01_f01 HHIHHIIH                r01_f01 ALL MASONRY WORK AND EQUIPMENT OPER
r01_f02 HIIHIRIII               r01_f02 LAYING BLOCK FOUNDATIONS POURING CO
r02_f00 INSWRANCE               r02_f00 INSURANCE
r02_f01 FINANUAL ANALYST        r02_f01 FINANCIAL ANALYST
r02_f02 PREMRIW REPDRD          r02_f02 PREPARING REPORTS
r03_f00 BLANK                   r03_f00 BLANK
r03_f01 PHH                     r03_f01 PERSONNEL RECEPTIONIST
r03_f02 THIH                    r03_f02 TYPING FILING
r04_f00 AERO SPME               r04_f00 AERO SPACE
r04_f01 MANSGEC                 r04_f01 MANAGER
r04_f02 MAMALEWG                r04_f02 MANAGERUG
```

There is no significance to the H and I above, except to represent what some imaginary system produces as classifications when the hypothetical character segments are not isolated characters.

A.4 Enclosure 3: Comments about Enclosures 1 and 2

GOALS AND MATERIALS OF THE TEST

The digital images used in the Conference will contain most, if not all ASCII characters, and other non-ASCII characters. The goal of the recognition task will be to convert each image of an upper case or lower case letter in the images into the corresponding upper case ASCII character, to convert the image of each digit into the corresponding ASCII digit, to covert all other characters into ASCII spaces, to replace multiple spaces by a single space, and to report the result as the hypothetical classification.

Refer now to Enclosures 1 and 2, and note that the top three miniforms contain only hand print, but that the quality of the hand print is variable both in character formation and in segmentation.

The bottom two miniforms show answer formats that key punch operators can handle, even though most OCR systems will probably not be able to. This is indicated by the hypotheses made up mostly of H's and I's to illustrate the results of a made-up system that tends to classify anything that's not an ASCII character as either an H or an I. Presumably, the confidences for these hypotheses would be very low or zero, so the adverse effect of these aberrant answers on the OCR error rate (fraction) can be minimized by a good rejection process.

Also, notice that the top field in the last form is empty. The key entry operators were instructed to enter BLANK for the 1980 large sample as shown in the reference file in the case of an empty field, but were not so instructed for the 1990 Census. Instead they were to leave the field empty. In fact, the procedure that we are using to remove the worst quality images from the sample also removes images that have empty fields as a side effect. Nevertheless, you will be instructed to enter BLANK into any blank fields that you encounter as a precaution against a few sneaking through. This is illustrated in the hypothesis file for the last form in Enclosure 1.

Spelling errors by the people filling out the forms and by the key entry operators introduce problems that complicate scoring. These are illustrated by the last field in the last record of the reference file: r04_f02 MANAGERUG. The key entry operators were instructed to type what was printed without attempting to correct spelling or typographical errors, and MANAGERUG is what's in the reference file, even though we might guess that MANAGING is what was meant. However, sometimes the key entry operators will not notice the misspelling, but will just type the word that they recognize. This is illustrated by r03_f02 TYPING FILING,

where the actual writing on the image gives Tiping, Filing. This field also illustrates the fact that all punctuation has been removed from the reference file data in accordance with the goal stated at the beginning of this enclosure.

Some of the incorrect words in the made-up hypothesis file shown in Enclosure 2 can be corrected by a sufficiently powerful dictionary look-up algorithm. Therefore, we will be providing word and phrase dictionaries for use in performing the recognition task. These are available at the ftp site mentioned above in the directory dicts. There are nine dictionaries there:

phrase_0.lng	word_0.lng	phrase_0.sht
phrase_1.lng	word_1.lng	phrase_1.sht
phrase_2.lng	word_2.lng	phrase_2.sht

These were made from a 132,000 sample of the fields f00, f01, and f02 obtained from the 1980 Census. The dictionary phrase_Z.lng contains all of the phrases (after removal of punctuation and double spaces) occurring in field f0Z in the sample, and the dictionary word_Z.lng contains all of the words occurring in that field, while phrase_Z.sht contains all of the phrases that occur more than once. The coverage of the short dictionaries is quite good. Each contains only about 8000 phrases, but covers 70%, 70%, and 60%, respectively, of the fields f00, f01, and f02 in the 132,000 phrases samples for each field. It is expected that the short dictionaries will provide nearly this level of coverage for the 1990 Census data being used for the Conference sample, training, and testing data. The long phrase dictionaries are of the order of 45,000 phrases, and it possible that they will not cover the 1990 Census data much better than the short phrase dictionaries do.

The word dictionaries are about 13,000 words long. About half of the words are either misspellings or abbreviations. We have looked into the possibility of mapping these into correct words or roots, but have not found a fool-proof way of doing this so far. This fact combined with the fact that the key entry operators do not always key what they were supposed to introduces some uncertainties into how to best to use the dictionaries, and the resolution of this problem is left to the participants. Remember, the goal is to reproduce the letters and digits contained in the image, and the dictionaries will contain common misspellings and abbreviations.

A report describing our preliminary study of the problems associated with dictionaries both for correcting the results of OCR and for scoring can be found in /pub/NISTIR/ir_5180.ps at the ftp site listed above. It is in PostScript(C) format; copies can be obtained from Allen Wilkinson at the e-mail and FAX addresses listed above if you don't have access to

Postscript printing capability.

New, improved dictionaries will be provided with the training material. They will have all of the words and phrases in the sample dictionaries, but will also include extra words and phrases. We do not expect the short dictionaries to be much larger, but would not be surprised if the long dictionaries grew substantially.

Since some potential participants may not have dictionary-based correction algorithms available for use with their OCR results, we tentatively plan to allow each participant to request that we run no more than one set of his or her test results through a NIST-developed correction suite. We would then score both sets of results with two different measures of field level accuracy as described below. Typical results for synthesized data designed to simulate NIST participation in the Conference are show below to give an example of what such dictionary correction can do:

SCORING EXAMPLES AND DEFINITIONS

FIELD LEVEL ACCURACY MEASURES FOR SIMULATED OCR CLASSIFICATION DATA

	BEFORE DICTIONARY BASED CORRECTION	AFTER DICTIONARY BASED CORRECTION
field distance fraction	33%	21%
field error fraction	92%	51%

field level rej. fraction	field level error fraction
0.00	0.51
0.10	0.46
0.20	0.40
0.30	0.35
0.40	0.31
0.50	0.27
0.60	0.24
0.70	0.20
0.80	0.14
0.90	0.06

Now we describe what these scores, which are what we plan to use for the Conference, actually mean, and we welcome your comments on this aspect of the plan, which is still being perfected. You will receive the final plan with the training materials, and the decision of the Committee in this regard will become final at that time. Lets start with a few definitions.

A hypothesis classification (hypothesis for short) is an ASCII phrase that has been assigned by a system to an unknown digital image of a hand print phrase. A reference classification (reference for short) is the phrase that the hypothesis phrase will be scored against. Unfortunately, the reference phrase will not always be what you or I would consider the correct phrase. Many images contain misspelled words. The key entry operators were instructed to key what was printed on the form without correcting misspellings. However, since humans recognize words rather than letters when they are reading, the key entry operators sometimes entered the correct version of a word rather than the misspelled version, never noticing the misspelling. Also, many images contain abbreviations. Unfortunately, as mentioned above in connection with dictionaries, we have not been able to devise an automated way to map abbreviations and misspellings onto corrected or expanded words or roots.

Under these conditions, the field error fraction, by itself, might not give a good comparison of the performance of two different systems. Therefore, we will calculate not only the field fraction but also a measure of the distance between the hypothesis and the reference field.

To calculate the field error fraction, we will just compare each hypothesis field with the corresponding reference field, including spaces. If they are identical, we will increment a correct-field hypothesis counter, cf. If not, we will increment a error-field hypothesis counter, ef. We will sum the cf and ef counters over all accepted (not rejected) fields and the field error rate will then be calculated as

field error rate = ef/(cf+ef).

To calculate the distance between a hypothesis and reference field, we will compute an alignment between each hypothesis and the corresponding reference phrase that minimizes the Levenstein distance [1-5] between the two phrases. In calculating the Levenstein distance, we plan to use 3, 1,

and 5, as the penalties for letter substitution, insertion, and deletion errors, respectively. Finally, we will use the alignment of the hypothesis and reference phrase to calculate

error rate = (s+i+d)/(c+s+i+d),

where

s = # of substitution errors
i = # of insertion errors
d = # of deletion errors
c = # of correct characters

are summed over all accepted fields.

We will calculate both the field error fraction and the field distance fraction as a function of the field level rejection fraction. We will not use any character level rejection fractions. The latter do not seem to be useful as final system outputs even though they might be very useful in obtaining the final system output. This is illustrated below for an image that says

TIPING FILING

Suppose the hypothesis were

TIPMG FILMQ

with a field level confidence of 0.72, and with the following confidences for the individual letters:

T 0.853
I 0.573
P 0.993
M 0.678
G 0.921

F 0.950
I 0.976
L 0.892
M 0.734
Q 0.621

If character level rejection were used, then with a rejection fraction of 0.00, the hypothesis would be

TIPMG FILMQ

but with a rejection fraction of 0.40, the hypothesis might be

T P G FIL

This does not seem to be useful for any application.

On the other hand, with a field level rejection of 0.72, either the entire hypothesis along with all of its letters is accepted and included in the set of hypotheses to be scored for both field error and distance, or else it is rejected and withheld from the set to be scored, depending upon the rejection threshold.

This example can also be used to illustrate a potential problem with the field error fraction. Suppose the hypotheses from three different systems for the image that said TIPING FILING were

ENCLOSURE 3, PAGE 5

TIPMG FILMQ

TYPING FILING

and

TIPING FILING

and suppose that the reference phrase were TYPING FILING because the key entry operator did not notice the misspelling. Then the system giving the correct classification TIPING FILING would get the same bad score, ef = ef + 1, for the correct phrase as did the system giving the almost unreadable TIPMG FILMQ, while the system giving the incorrect phrase TYPING FILING would get a good score of cf = cf +1 for what is actually an incorrect classification. Since this type of error is rare, it may not be a significant source of error, but correlation of the field error fraction with the field distance fraction for the various systems should help to point out potential problems if there are any, or show that there are none.

[1] H. G. Zwakenberg, Inexact Alphanumeric Comparisons, The C Users Journal, 127 (May 1991).

[2] R. Valdes, Finding String Distance, Dr. Dobb's Journal, 56 (April 1992), and references therein.

[3] R. A. Wagner and M. J. Fischer, The String-to-String Correction Problem, J. ACM 21, 168 (1974), and reprinted in S. N. Srihari, Tutorial: Computer Text Recognition and Error Correction, IEEE Computer Press (1985).

[4] M. D. Garris and S. A. Janet, NIST Scoring Package User's Guide Release 1.0, NISTIR 4950 (October 1992).

[5] M. D. Garris, Methods for Evaluating the Performance of Systems Intended to Recognize Characters from Image Data Scanned from Forms, NISTIR 5129, (February 1993).

A.5 Enclosure 4: Form of Letter to Request Participation

YOUR RETURN ADDRESS AND AFFILIATION IF ANY

Jon Geist
NIST/225/B063
Gaithersburg, MD 20899

Dear Dr. Geist,

I hereby request that you include me as a participant in the 2nd Census OCR Systems Conference. My information pertinent to participation is given below:

YOUR MAILING ADDRESS INCLUDING YOUR NAME

YOUR ACTUAL ADDRESS IF DIFFERENT OR IF MAILING ADDRESS IS A PO BOX

YOUR VOICE PHONE NUMBER

YOUR FAX PHONE NUMBER, IF ANY

YOUR E-MAIL ADDRESS, IF ANY

I have read, understood, and agree to abide by the rules of participation dated 93/06/16.

Sincerely,

YOUR SIGNATURE

A.6 Enclosure 5: Rules of Participation in 2nd Census OCR Systems Conference

1) Participants shall not human classify or human correct any results. If Participants human check any results as a sanity or blunder check, they shall report this fact and any resulting changes when returning their results.

2) Participants shall return their classification results within two weeks after receiving the test materials from NIST in the format and media requested by NIST.

3) Participants shall return one filled out summary sheet per entry with all obligatory responses completed.

4) Attendance at the Conference may be limited to Participants chosen by the Committee based on the space available to accommodate Participants.

5) Attendance at Conference session(s) where Participants openly discuss their procedures and results may be restricted to those Participants who have agreed to participate openly as demonstrated by the answers that they provide on the form illustrated in Enclosure 6 that is to be returned at the same time as the test results.

6) The Committee will make the final decision if any unforeseen questions arise, and the only recourse (other than continued participation) that is open to Participants not satisfied with any of these decisions is withdrawal from further participation.

7) All Participant's summary scores (error vs. rejection data, etc.) will be given to the Participants at the Conference and will be published in the Conference results. Participant's own scores or all scores might be sent to the Participants in advance of the Conference at the discretion of the Committee, but if they are sent to one Participant, they will be sent to all Participants.

8) Anyone to whom NIST sends the training materials will be listed as participating in the Conference, and if they fail to return their test results, this fact will be stated in the Conference report.

9) The Conference report will contain a disclaimer that is similar, if not identical to the following:

The U.S. Bureau of the Census (Census) and the National Institute of Standards and Technology (NIST) sponsored this Conference as part of ongoing research into machine recognition of hand-print. Participants may have submitted the results of experimental or developmental systems for scoring even if they have commercial products in the market place, and the efforts of the participants in conducting the tests were not proctored in any way. While some test results from this Conference may appear in marketing literature, potential buyers must beware! Census and NIST can make only one recommendation to potential buyers: use your own application- specific data to thoroughly test the performance of any system or component in a realistic setting.

A.7 Enclosure 6: Example of Questionnaire to be Filled Out and Returned for each Result Submitted for Scoring

FORM TO BE RETURNED FOR EACH TEST RESULT SUBMITTED FOR SCORING

OBLIGATORY RESPONSES:

SYSTEM NAME:

PARTICIPANT NAME:

VOLUNTARY RESPONSES:

1. CHARACTER SEGMENTATION (PLEASE CHECK APPROPRIATE BOXES.):
   - DONE BEFORE CHARACTER RECOGNITION ..... [ ]
   - DONE ITERATIVELY WITH CHARACTER RECOGNITION ..... [ ]
     - CHECKING DICTIONARY AS PART OF ITERATIVE PROCESS . [ ]
   - DONE SIMULTANEOUSLY WITH CHARACTER RECOGNITION .... [ ]
   - BLOB COLORING ..... [ ]
   - SPATIAL HISTOGRAMS ..... [ ]
   - LOCAL MINIMA AND MAXIMA ..... [ ]
   - INTENTIONAL OVER SEGMENTATION ..... [ ]
   - STROKE RECONSTRUCTION ..... [ ]
   - VARIABLE SCALE NETWORKS ..... [ ]
   - CENTRAL OBJECT ..... [ ]
   - TIME DOMAIN NEURAL NETWORK ..... [ ]
   - OTHER: ..... [ ]_____

2. CHARACTER RECOGNITION

2.1 PREPROCESSING (PLEASE CHECK APPROPRIATE BOXES.):
   - CONVERSION TO GREY SCALE ..... [ ]
   - HEIGHT/WIDTH NORMALIZATION WHILE PRESERVING SHAPE [ ]
   - SEPARATE NORMALIZATIONS FOR HEIGHT AND WIDTH ..... [ ]
   - SLANT NORMALIZATION ..... [ ]
   - ROTATION ..... [ ]
   - LOCAL SUPPORT (GABOR, WAVELET, ETC.) ..... [ ]
   - FOURIER OR SIMILAR TRANSFORM ..... [ ]
   - OTHER: ..... [ ]_____

2.2 SEPARATE FEATURE EXTRACTION AND CLASSIFICATION: [ ] YES
   IF YES GO TO QUESTION 2.3, ELSE GO TO QUESTION 2.5.

2.3 FEATURE EXTRACTION (PLEASE CHECK APPROPRIATE BOXES.):
   - *ADAPTIVE LEARNING ..... [ ]
   - +SUPERVISED ..... [ ]

TIME DOMAIN NEURAL NETWORK ..... [ ]
RECEPTOR FIELDS ..... [ ]
OTHER SUPERVISED: ..... [ ]_____
+SELF-ORGANIZED ..... [ ]
KOHONEN MAPS ..... [ ]
NEO-COGNITRON ..... [ ]
OTHER SELF-ORGANIZED: ..... [ ]_____
+OTHER ADAPTIVE: ..... [ ]_____
*RULE-BASED ..... [ ]
+LINEARIZING TRANSFORMS ..... [ ]
LINE FIT ..... [ ]
POLYNOMIAL ..... [ ]
OTHER LINEARIZING TRANSFORM: ..... [ ]_____
+CONVOLUTION/CORRELATION ..... [ ]
-TRANSFORMS ..... [ ]
HAND CODED ..... [ ]
GABOR ..... [ ]
OTHER TRANSFORMS: ..... [ ]_____
-TEMPLATES ..... [ ]
-OTHER CONVOLUTION/ETC: ..... [ ]_____
+MODEL ..... [ ]
STROKES ..... [ ]
SHAPES ..... [ ]
HOLES ..... [ ]
CAVITIES ..... [ ]
MORPHOLOGICAL ..... [ ]
OTHER MODEL: ..... [ ]_____
+STATISTICAL ..... [ ]
PRINCIPAL COMPONENT ANALYSIS (K-L TRANSFORM) ... [ ]
HISTOGRAM ..... [ ]
OTHER STATISTICAL: ..... [ ]_____
+OTHER RULE-BASED: ..... [ ]_____

2.4 CLASSIFICATION (PLEASE CHECK APPROPRIATE BOXES)

*ADAPTIVE LEARNING ..... [ ]
+SUPERVISED ..... [ ]
MULTI-LAYER PERCEPTRON ..... [ ]
LEARNED VECTOR QUANTIZATION ..... [ ]
REDUCED COULOMB ENERGY ..... [ ]
AFFINE TRANSFORMATION ..... [ ]
OTHER SUPERVISED: ..... [ ]_____
+SELF-ORGANIZED ..... [ ]
CASCADED NEURAL NETWORK ..... [ ]
LOOK-UP TABLE ..... [ ]
PROBABILITY NEURAL NETWORK ..... [ ]
OTHER SELF-ORGANIZED: ..... [ ]_____
+OTHER ADAPTIVE: ..... [ ]_____

*RULE-BASED ..... [ ]
+GEOMETRIC ..... [ ]
NEAREST NEIGHBOR ..... [ ]
K-NEAREST NEIGHBOR ..... [ ]
PNN ..... [ ]
OTHER GEOMETRIC: ..... [ ]_____
+STATISTICAL ..... [ ]
PROBABILITY ..... [ ]
QDF ..... [ ]
POLYNOMIAL ..... [ ]
OTHER STATISTICAL: ..... [ ]_____
+OTHER RULE-BASED: ..... [ ]_____

2.5 HYBRID FEATURE EXTRACTION AND CLASSIFICATION: PLEASE GIVE A DESCRIPTIVE NAME FOR YOUR APPROACH USING TERMS FROM QUESTIONS 2.3 AND 2.4 WHERE APPROPRIATE.

2.6 TRAINING

*CHARACTERS IN SPECIAL DATA BASE 1 ..... [ ]
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]
LOWER CASE ..... [ ]
DIGITS ..... [ ]
*CHARACTERS IN SPECIAL DATA BASE 3 ..... [ ]
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]
LOWER CASE ..... [ ]
DIGITS ..... [ ]
*CHARACTERS IN SPECIAL DATA BASE 7 ..... [ ]
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]
LOWER CASE ..... [ ]
DIGITS ..... [ ]
*CHARACTERS IN SPECIAL DATA BASE 11 ..... [ ]
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]
LOWER CASE ..... [ ]
DIGITS ..... [ ]
*CHARACTERS IN SPECIAL DATA BASE 12 ..... [ ]
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]
LOWER CASE ..... [ ]
DIGITS ..... [ ]
*OTHER CHARACTER SETS..... [ ]_____
+APPROXIMATE TOTAL NUMBER USED ..... [ ]
UPPER CASE ..... [ ]

LOWER CASE ..... [ ]
DIGITS ..... [ ]

3. DICTIONARY-BASED CORRECTION (PLEASE CHECK APPROPRIATE BOXES.):
*NOT DONE (REQUEST NIST CORRECTION) ..... [ ]
*NOT DONE (REQUEST NO FURTHER CORRECTION) ..... [ ]
*DONE AFTER CHARACTER RECOGNITION ..... [ ]
+FIRST (OR ONLY) PASS THROUGH A DICTIONARY ..... [ ]
-USED WORD DICTIONARIES ..... [ ]
NIST SUPPLIED ..... [ ]
OTHER ..... [ ]
-USED PHRASE DICTIONARIES ..... [ ]
NIST SUPPLIED ..... [ ]
LONG ..... [ ]
SHORT ..... [ ]
OTHER ..... [ ]
-OBJECTS IN DICTIONARY CODED AS LETTERS ..... [ ]
-OBJECTS IN DICTIONARY CODED AS OTHER ..... [ ]
-SEARCH ENTIRE DICTIONARY ..... [ ]
-SEARCH SUBSET OF DICTIONARY ..... [ ]
HASHED OR INDEXED SEARCH OF DICTIONARY ..... [ ]
OTHER ..... [ ]
-EXACT MATCH REQUIRED ..... [ ]
-STATISTICAL DISTANCE MEASURE MINIMIZED ..... [ ]
-OTHER DISTANCE MEASURE MINIMIZED ..... [ ]
+SECOND PASS THROUGH A DICTIONARY ..... [ ]
-USED WORD DICTIONARIES ..... [ ]
NIST SUPPLIED ..... [ ]
OTHER ..... [ ]
-USED PHRASE DICTIONARIES ..... [ ]
NIST SUPPLIED ..... [ ]
LONG ..... [ ]
SHORT ..... [ ]
OTHER ..... [ ]
-OBJECTS IN DICTIONARY CODED AS LETTERS ..... [ ]
-OBJECTS IN DICTIONARY CODED AS OTHER ..... [ ]
-SEARCH ENTIRE DICTIONARY ..... [ ]
-SEARCH SUBSET OF DICTIONARY ..... [ ]
HASHED OR INDEXED SEARCH OF DICTIONARY ..... [ ]
OTHER ..... [ ]
-EXACT MATCH REQUIRED ..... [ ]
-STATISTICAL DISTANCE MEASURE MINIMIZED ..... [ ]
-OTHER DISTANCE MEASURE MINIMIZED ..... [ ]
+MORE THAN TWO PASSES THROUGH DICTIONARIES ..... [ ]
*DONE ITERATIVELY WITH CHARACTER RECOGNITION ..... [ ]
OBJECTS IN DICTIONARY CODED AS LETTERS ..... [ ]
OBJECTS IN DICTIONARY CODED AS OTHER ..... [ ]

SEARCH ENTIRE DICTIONARY ..... [ ]
SEARCH SUBSET OF DICTIONARY ..... [ ]
HASHED OR INDEXED SEARCH OF DICTIONARY ..... [ ]
EXACT MATCH REQUIRED ..... [ ]
STATISTICAL DISTANCE MEASURE MINIMIZED ..... [ ]
OTHER DISTANCE MEASURE MINIMIZED ..... [ ]_____
*NOT EXACTLY A DICTIONARY SEARCH, BUT SAME PURPOSE . [ ]
HIDDEN MARKOV MODEL ..... [ ]
OTHER ..... [ ]_____

4. OTHER CONTEXT BASED CORRECTION ..... [ ]
LETTERS BY SAME WRITER ..... [ ]
WORDS OR PHRASES BY SAME WRITER ..... [ ]
OTHER: ..... [ ]_____

5. PLEASE ATTACH A LIST OF REFERENCES TO YOUR PERTINENT PUBLICATIONS.

6. PLEASE ATTACH A LIST OF REFERENCES TO YOUR GENERAL TECHNIQUES, IF NOT ALREADY DESCRIBED IN AN ATTACHED LIST OF REFERENCES.

7. IF THE ABOVE QUESTIONS DO NOT CAPTURE THE ESSENCE OF YOUR SYSTEM, PLEASE PROVIDE AN ATTACHMENT THAT DOES.

8. IF OTHER THAN THE DIGITS, THE UPPER CASE LETTERS, AND THE LOWER CASE LETTERS, PLEASE STATE THE FULL CHARACTER SET THAT YOUR SYSTEM RECOGNIZES.

B The Instructions For Participants

Dear Participant in the 2nd Census OCR Systems Conference

The following will be included with Special Database 13, which will contain the test materials for the conference:

Dear Participant in the 2nd Census OCR Systems Conference

Please find enclosed Special Database 13, which constitutes the test materials for the Conference, and a DOS formatted 3.5" disk with a revised version of chkfiles.

Also, note that there are duplicate entries in the dictionaries on Special Databases 12 and 13. This is not a major problem; its only effect is to somewhat increases the time needed to search the dictionaries. Even so, we apologize for this error. This error can be easily corrected, particularly if you were intending to copy the dictionaries to a hard disk anyway to shorten the search time. For instance, on a UNIX system, you can make a subdirectory new_dicts on a writeable medium, run

```
sort -u /cd/dicts/phrase_1.lng > new_dicts/phrase_1.lng
```

etc., for each dictionary of interest, where /cd is the mount point for the CD-ROM, and carry out all dictionary searches on the contents of new_dicts.

All of the key points of the instructions for returning results to NIST are given below, including minor corrections from the version that was sent to you on Thursday, November 25. These instructions are reviewed using imaginary NIST entries as examples. You will use your system name wherever the NIST system name is used.

B.1 Example 1:

The first example is NIST's first entry for the images scanned from paper. It includes both hypothesis files (.hyp) and confidence files (.con) in the following directory structure:

```
SYSTEM NAME: NIST
TEST DATA SET: directory data4 of Special Database 13

NIST_0.P
|
---------
|       |
d00 ... d05
|
----------------
|              |
d00f00.hyp ... d00f99.hyp
d00f00.con ... d00f99.con
```

The hypothesis files should have been created by running the images in the mis files in data4 of Special Database 13 through an OCR system and then through a dictionary-based correction system or through an OCR system that carries out dictionary-based correction during the OCR process. If your system produces raw OCR results for individual characters, without any dictionary-based correction either during or following the OCR process, then go to EXAMPLE 4 before reading this example. EXAMPLE 4 should probably be read by all participants for a small point about strategy, even though many will want to skip EXAMPLE 2 and EXAMPLE 3.

The hypothesis files should have the same format as the reference files (.ref) on Special Database 12 and as illustrated below:

```
d00f00.hyp                    d00f00.con
-------------------------     ----------------------
r00_f01 MANAGER               r00_f01 0.937722
r00_f02 MANAGER               r00_f02 0.905212
r00_f03 MANAGER               r00_f03 1.123456e-2
r01_f01 CONSTRUCTON           r01_f01 0.734562
r01_f02 INSURANCE SALES       r01_f02 0.534900
r01_f03 INSURANCE SALES       r01_f03 0.994562
r02_f01 INSURANCE             r02_f01 0.865297
r02_f02 FINANCIAL ANALYST     r02_f02 0.369254
r02_f03 PREPARED REPORT       r02_f03 0.273099
r03_f01 BLANK                 r03_f01 1.000000
r03_f02 TYPING                r03_f02 0.667345
```

```
r03_f03 TYPING                  r03_f03 0.934121
r04_f01 AERO SPACE              r04_f01 0.485903
r04_f02 MANAGEMENT              r04_f02 0.123963
r04_f03 MANAGING                r04_f03 0.528631
```

The LINE_IDs in the hypothesis and confidence files must be identical to those on the corresponding lines of the .mis files in Special Database 13. There must be one and only one space between a LINE_ID and the hypothesis or confidence that follows it.

The hypothesis must consist only of the ASCII digits, upper case letters, and spaces character, and may not start or end with a space character. The hypotheses will be scored by aligning them, including spaces, with the corresponding references using the Levenstein distance algorithm and scoring each deletion, insertion, and substitution as an error, and all aligned characters as correct.

Only one confidence that applies to each hypothesis is allowed on the corresponding line of the confidence file. The confidence must be a number between 0.0 and 1.0 that can be read by the c atof (more precisely, the stdtod) function. The hypotheses will be ordered using the confidences and scored as a function of rejected fraction, where the hypotheses with the lowest confidences will be rejected first.

The new_line character (UNIX) or the ASCII carriage return and line feed characters (DOS) must follow each hypothesis and confidence (without any intervening spaces) to terminate each line.

The format of the hypothesis and confidence files in the directory tree should be checked by running the program chkfiles that can be compiled from chkfiles.c that is provided in the src subdirectory on Special Database 13 or from the enclosed DOS formatted 3.5" disk. The two versions have somewhat different user interfaces. A man page giving instructions for use is provided in the man subdirectory on each medium.

Any format errors that are discovered the hypothesis and confidence files should be corrected by modifying the program that generates them rather than by hand.

The entire directory must be in either UNIX or DOS format.

A hypothesis file must be provided for each mis file in data4 on Special Database 13. The use of confidence files is optional, but you must provide either all of them or none of them. We strongly prefer confidence files to reject files, so only consider the next example if your system cannot produce confidence files.

B.2 Example 2:

The second example is NIST's second entry for the images scanned from paper. It includes both hypothesis files (.hyp) and reject files (.rjX), X = 0, 1, ..., 9, in the following directory structure:

```
SYSTEM NAME: NIST
TEST DATA SET: directory data4 of Special Database 13

NIST_1.P
|
---------
|       |
d00 ... d05
|
----------------
|              |
d00f00.hyp ... d00f99.hyp
d00f00.rj0 ... d00f99.rj0
d00f00.rj1 ... d00f99.rj1
           ...
d00f00.rj9 ... d00f99.rj9
```

Everything is the same as in the first example, except the reject files look like:

```
d00f00.rj0
-----------------------
r00_f01 0
r00_f02 0
r00_f03 1
r01_f01 0
r01_f02 0
r01_f03 0
r02_f01 0
r02_f02 0
r02_f03 0
r03_f01 1
r03_f02 0
r03_f03 0
r04_f01 0
r04_f02 0
r04_f03 0
```

Only one rejection code that applies to each hypothesis is allowed on the corresponding line of the rejection file. The reject code must be an ASCII

zero or an ASCII one. A zero means that the corresponding hypothesis is to be accepted and scored; a one means that the corresponding hypothesis is to be rejected and not scored. Be careful: this is exactly the opposite of the interpretation of zero and one as used with confidence files. (We apologize for this poor choice of convention.) If you do provide rejection files rather than confidences, please try to provide at least one in which a little less than half of the rejection codes are 1 and a little more than half are zero giving a rejected fraction or rejection rate of about 0.50.

You may provide as many as ten sets of rejection files, but the benefit of having one rejection file giving something near to but less than 50% rejection is much larger than that from having a lot of sets of rejection files, none of which produce close to (and preferably, less than) 50% rejection.

The new_line character (UNIX) or the ASCII carriage return and line feed characters (DOS) must follow each rejection code (without any intervening spaces) to terminate each line.

The format of at least one set of reject files in the directory tree should be checked by running the program chkfiles described in connection with the first example. If you compile it from Special Database 13, it will be necessary to globally change the extension from .rjX, for one of X = 0, 1, ..., or 9, to .rej. (We apologize for this mistake. This is why we have included a corrected version on the enclosed 3.25" DOS format disk. In any case, since we have made this error, you may leave the set you check with the extension .rej after you have finished checking it.) On the other hand, if you compile chkfiles from the source code on the enclosed 3.5" disk, you will be able to put the extension names to be checked on the command line.

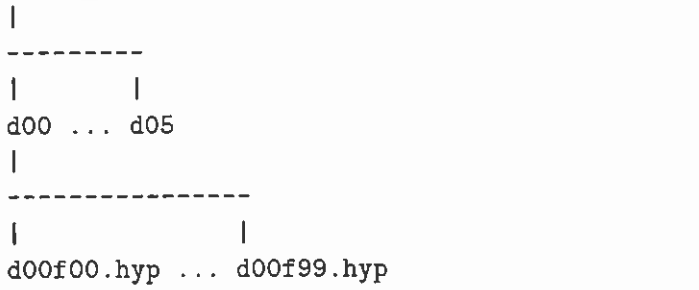
Any format errors that are discovered should be corrected by modifying the program that generates the hypothesis or reject files rather than by hand.

As pointed out above, confidence files are our first choice. Our second choice is both confidence files and rejection files, but we will score using only the confidence files unless you can give us a pretty good reason to also score using the rejection data. Our third choice is a set of rejection files, at least one of which produces a rejection fraction approximately equal to, but less than to 0.5. Our fourth choice is a set of rejection files that does not have this property. Our last choice is neither confidence nor reject files, but we will use it as our next example just so we can give one example of the directory structure for the images scanned from microfilm that are in data3 on Special Database 13.

B.3 Example 3:

The third example is NIST's first entry for the images scanned from microfilm. It includes only hypothesis (.hyp) files in the following directory structure:

SYSTEM NAME: NIST
TEST DATA SET: directory data3 of Special Database 13



Everything is the same as in the first example, except there are neither confidence nor rejection files.

B.4 Example 4:

This example is of interest if you wish to return raw OCR results that have never been subjected to dictionary-based correction, but the short comment on strategy might be of interest to all participants.

If you return results as described below, we will compare the field error and distance for your raw OCR results with those for any other raw OCR results that we receive, including those from NIST.

We will also pass your raw OCR results through our dictionary-based correction algorithm and compare the field error and distance rates with those for the other dictionary-based results.

The fourth example is NIST's third entry for the images scanned from paper, but it differs from EXAMPLE 1 in that the hypotheses were produced with raw OCR that outputs isolated characters rather than a dictionary-based correction program or a dictionary-assisted OCR system that outputs words or phrases from a dictionary rather than isolated characters. This example includes both hypothesis files (.hyp) and confidence files (.con) in the following directory structure:

```
SYSTEM NAME: NIST
TEST DATA SET: directory data4 of Special Database 13

NIST_9.P
|
---------
|       |
d00 ... d05
|
----------------
|              |
d00f00.hyp ... d00f99.hyp
d00f00.con ... d00f99.con
```

The formats are the same as in EXAMPLE 1, except that the 9 identifies the directory as containing raw OCR results. An example of the files is shown below:

```
d00f00.hyp                    d00f00.con
--------------------------    ----------------------
r00_f01 MAXGER                r00_f01 0.937722
r00_f02 MANAGER               r00_f02 0.905212
r00_f03 MANIIGER              r00_f03 0.347631
r01_f01 CDNSTRUCTON           r01_f01 0.734562
```

```
r01_f02 IBSVANCESALES          r01_f02 0.534900
r01_f03 HSORAND FCES           r01_f03 0.000000
r02_f01 INSUTANCE              r02_f01 0.865297
r02_f02 FINAFIALHNALYST        r02_f02 0.369254
r02_f03 FREFARDREURT           r02_f03 0.000000
r03_f01 BLANK                  r03_f01 1.000000
r03_f02 TYPING                 r03_f02 0.667345
r03_f03 TYPXG                  r03_f03 0.934121
r04_f01 AEROSPACE              r04_f01 1.000000
r04_f02 MANAGEMENT             r04_f02 0.904562
r04_f03 MAIIAGWG               r04_f03 0.528631
```

The hypotheses are shown without any spaces between the words. You may submit them either way, but unless your system is quite good at locating spaces correctly, it will probably hurt rather than help our dictionary-based algorithm, although it may help rather than hurt your raw OCR score.

Our dictionary-based algorithm will ignore all confidence values except those that are zero. These will be used as an indication that the raw OCR results are so bad that dictionary-based correction is hopeless and the best strategy is to try to minimize the damage. A made up example of what our dictionary-based algorithm might do with the above hypothesis and confidence files is shown below:

```
d00f00.hyp                     d00f00.con
-------------------------      ----------------------
r00_f01 MANAGER                r00_f01 0.873451
r00_f02 MANAGER                r00_f02 0.979342
r00_f03 MANIIGER               r00_f03 0.883290
r01_f01 CONSTRUCTION           r01_f01 0.983289
r01_f02 INSURANCE SALES        r01_f02 0.673098
r01_f03                        r01_f03 0.000000
r02_f01 INSURANCE              r02_f01 0.763291
r02_f02 FINANCIAL ANALYST      r02_f02 0.459827
r02_f03                        r02_f03 0.000000
r03_f01 BLANK                  r03_f01 1.000000
r03_f02 TYPING                 r03_f02 0.964195
r03_f03 TYPING                 r03_f03 0.782458
r04_f01 AEROSPACE              r04_f01 1.000000
r04_f02 MANAGEMENT             r04_f02 0.983245
r04_f03 MANAGING               r04_f03 0.423486
```

The important point is that the two raw OCR hypotheses having confidence zero have been replaced by blank lines, while the remaining hypotheses have been replaced by phrases taken from a dictionary of expected phrases. A blank line consists of a new_line character (UNIX) or an ASCII carriage return and line feed character (DOS) immediately following the ASCII space

character that follows the LINE_ID.

Replacing the raw OCR by a blank is a near optimum strategy to minimize the field distance when you haven't a clue even to how many characters were in the original image. Guessing one or two ASCII Es might be a little better; I haven't studied this issue.

B.5 A Summary of the Instructions Sent Out November 16 in Light of the Above

1) Generate from zero to five different sets of your best hypotheses (with confidence or rejection data, if possible) for the test images in data4, and label them NAME_0.P through NAME_4.P, where NAME is the name assigned to your system on the notice of acceptance in the Conference. Test their format with chkfiles.

1) Generate from zero to five different sets of your best hypotheses (with confidence or rejection data, if possible) for the test images in data3, and label them NAME_0.M through NAME_4.M. Test their format with chkfiles, and correct the program that generated them if incorrect formats are found.

3) Generate at most one set of raw (no dictionary based correction) OCR hypotheses for the images in data4, and label them NAME_9.P. Test its format with chkfiles, and correct the program that generated it if incorrect formats are found.

4) Generate at most one set of raw (no dictionary based correction) OCR hypotheses for the images in data3, and label them NAME_9.M. Test its format with chkfiles, and correct the program that generated it if incorrect formats are found.

5a) If your OCR system produces its results on a UNIX-based system, after having created the NAME_X.Y file directories, type from the directory in which each is located:

```
tar -cvf NAME_0_P.tar ./NAME_0.P
compress NAME_0_P.tar
```

and so forth for each set of results. Then copy the resulting files to either a DOS or UNIX formatted 3.25" floppy disk as described in 5a1), 5a2), or 5a3) below.

5a1) For instance, if you have a 3.25" drive mounted on your UNIX machine as /pcfs and you have a DOS formatted disk in that drive, you just type

```
cp NAME_0_P.tar.Z /pcfs/NAME0P.Z
```

The name change is required to accommodate the DOS name and extension conventions. Actually, we prefer this format because we expect some participants will have no choice but to provide us with DOS formatted disks as described next.

5a2) For instance, if you have no 3.25" drive mounted on your UNIX machine, but can ftp (or kermit, etc.) ASCII files from your UNIX machine to a DOS machine, then carry out the name change described in 5a1) above, and then copy the tarred and compressed file to a DOS formatted disk in the usual way. For instance, if the 3.25" disk drive is b:, you type

```
copy NAMEOP.Z b:NAMEOP.Z
```

5a3) For instance, if you have a 3.25" drive mounted on your UNIX machine as /pcfs and you have a UNIX formatted disk in that drive, you just type

```
cp NAME_O_P.tar.Z /pcfs/NAME_O_P.tar.Z
```

5b) If your OCR system produces its results on a DOS-based system, and you have no way to transfer them to a UNIX machine for tarring and compression as in 5a) above, install the pkzip package that we provided to you at your specific request, and after having created the NAME_X.Y file directories, type from the directory in which each file is located:

```
pkzip -rP NAMEOP.zip ./NAME_O.P/*.*
```

and so forth for each complete directory tree containing the results. Then copy the resulting files to either a DOS or UNIX formatted 3.25" floppy disk in the usual way. For instance, if the 3.25" disk drive is b:, you type

```
copy NAMEOP.Z b:NAMEOP.Z
```

6) give floppy disks to FedEx on or before the 15th of December for shipment to us at a FedEx address to be provided. It is OK to return results on more than one disk, for instance to get a very early return dates for one of your entries.

Also, (again) if you plan to return results in pkzip format, please let me know so that we can provide you with a copy of pkzip. If worse comes to worse the data for one paper or microfilm test can each be separated into three directories per disk and returned that way.

C System Summaries For On-Time Submissions

Stanley Janet and Jon Geist

This appendix contains system summaries for all results that were received on time. For each organization that supplied them, copies of viewgraphs describing that organization's system are followed by graphs of the field distance rate, field error rate, field distance rejection efficiency, and field error rejection efficiency for each system. Note that the same system name will apply to results from microfilm and from paper whenever both were provided.

The viewgraphs contained in the system summaries were discussed by the participants at the Second Conference meeting in February of 1994. The participants did not know that they would be asked to submit their viewgraphs for publication, and were not instructed to produce the viewgraphs in a format that would reproduce nicely for this document. Nevertheless, they most graciously agreed to provide the viewgraphs when it became clear that reproduction of the completed questionnaires from Appendix A.7 would create a number of misleading ideas about some systems.

Some participants turned in originals of their viewgraphs, others copies. Therefore, the starting image quality was not always as good as might be desired. The images were scanned and some crude image processing was carried out on some images where dot patterns (pseudo-gray scale) in the images were not well captured by the scanning process. As a result of these circumstances, which were beyond the participant's control, some of the images are not of ideal quality. Also, some are difficult to understand without the accompanying commentary. Nevertheless, they give a much more accurate impression of the various systems than the questionnaires would have. Most of the participants have published descriptions of their systems as they evolved, and some will publish descriptions of their experiences and results in the Second Conference test. The viewgraphs are not meant to replace these publications, but to whet the readers' appetite for them.

The format for the systems summaries used the first viewgraph as the introduction to the system summary where possible, and used a made-up introductory page otherwise. The viewgraphs then follow in an order approved by the participant(s) who developed the system. The graphs giving the field error and distance rates and efficiencies for each system follow the viewgraphs for that system.

The field error and distance rates plotted in the system summaries were defined in Chapter 6. The field error rejection efficiency is defined as

$$E_{fe}(r_f) = \frac{F_e(r_f) - F_e(r_f + \Delta_f)}{F_c(r_f) - F_c(r_f + \Delta_f) + F_e(r_f) - F_e(r_f + \Delta_f)} \tag{7}$$

where $F_e(r_f) - F_e(r_f + \Delta_f)$ is the number of incorrect fields that are rejected when the field rejection rate is increased from r_f to $r_f + \Delta_f$, and where $F_c(r_f) - F_c(r_f + \Delta_f)$ is the number of correct fields that are rejected when the rejection rate is increased from r_f to $r_f + \Delta_f$.

The field distance rejection efficiency is defined as

$$E_{fd}(r_f) = \frac{C_\epsilon(r_f) - C_\epsilon(r_f + \Delta_f)}{C_c(r_f) - C_c(r_f + \Delta_f) + C_c(r_f) - C_c(r_f + \Delta_f)} \tag{8}$$

where $C_c(r_f) - C_\epsilon(r_f + \Delta_f)$ is the number of incorrect characters (deletions, insertions, and substitutions) in the alignments that are rejected when the field rejection rate is increased from r_f to $r_f + \Delta_f$, and where $C_c(r_f) - C_c(r_f + \Delta_f)$ is the number of correct characters in the alignments

that are rejected when the rejection rate is increased from r_f to $r_f + \Delta_f$.

Both the field error and the field distance rejection efficiencies are bounded above by one and below by zero. When they are one, all of the fields or characters being rejected are incorrect; when they are zero, all of the fields or characters being rejected are correct. The quantity Δ_f was set to 0.02 in the graphs of the field distance and error rejection efficiencies to be consistent with the sampling interval used in the graphs of the field distance and error rates. These efficiencies are a measure of how efficiently the rejection process removes fields with errors while retaining fields without error at any given rejection rate. It is desirable to have the highest efficiency possible at the lowest rejection rates. This can be seen by comparing the NIST and University of Bologna field distance rates and efficiencies. The NIST field distance is greater than that of the University of Bologna at zero rejection rate, but falls below it at greater rejection rates. The difference is that the rejection efficiency at low rejection rates is somewhat greater for the NIST system. This more than makes up for the lower efficiency of the NIST system above 40% rejection rate. In fact, the average rejection efficiency for the NIST system is somewhat less than that for the University of Bologna system.

NIST Conference

2nd Census OCR Systems Conference

CEDAR methodology and results

Venu Govindaraju

Center of Excellence for Document Analysis and Recognition [CEDAR]

SUNY at Buffalo

Feb 15, 1994

Outline

Outline of Presentation

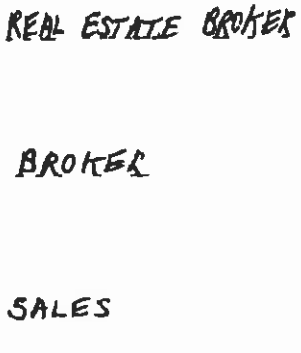
- Census/ NIST Task
- Methodology
- Field Extraction Algorithm
- Word Recognition Engine
- Lexicon Analysis
- Control Strategies: CEDAR_0,1,2
- Performance Analysis
- Future Improvements

Methodology

- Global form removal was not used
  - ➤ pages not registered uniformly
    (registration squares had touching material from other side of questionnaire)
- Local analysis of page to find boxes
  - ➤ detect dashes to find fields
  - ➤ remove dashes to find text
- Analyze lexicon to determine recognition strategy
- Apply methods developed for postal word recognition

Field Extraction Algorithm

1. Generate Connected Components
2. Detect Connected Components that look like dashes (Based on height and width information).
3. Cluster these dashes into horizontal lines.
4. Find location of each of box using horizontal lines.
5. Extract all connected components that lie within boxes.
6. Remove any remaining isolated dashes (horizontal or vertical) near the border of each box.



Word recognizers

CEDAR Word Recognition Engine

- Combination of two algorithms
- Algorithm 1
  - Recognition driven segmentation
  - Image based features for character recognition
    - ➢ Gradient features; Structural features; Concavity features
    - ➢ Lexicon introduced at the postprocessing stage
- Algorithm 2
  - Hard segmentation based on non-OCR image processing
  - Boundary features for segment (character: 1...4 segments)
    - ➢ Curvature and histogram information
    - ➢ Lexicon driven grouping of segments into characters

Methodology

Word Recognition Engine Performance

on 2500 postal word images (212 ppi)

Lexicon:	10		100		1000	
	Top Choice	Error	Top Choice	Error	Top Choice	Error
No threshold	97%	3%	92%	8%	80%	20%
Threshold	92%	0.5%	79%	0.6%	62%	1.8%
Speed (Sparc 10/30)	2.5 seconds		5 seconds		8 seconds	

Methodology

Comparison of USPS Address Reading & Census Questionnaire Reading

	USPS Address Reading	Census Form Reading
Form structure	Syntax loosely constrained	Known
Lexicon completeness	90%	60-75%
Lexicon size	1...1000 (avg: 20)	upto 60,000
Word types	Unconstrained (mostly cursive)	Discrete printing
Noise	<5% have underlines	Boxes always present

- CEDAR Word recognition engine developed for Address Reading Task
- No "special" tuning done to the tools for the purpose of this test

Lexicon Analysis

Census / NIST lexicon characteristics

1. field specific
2. large (10,000 - 60,000 entries)
3. incomplete
4. Each field has corresponding LONG and SHORT lexicons

Lexicon :	LONG	SHORT
Size	~ 60,000	~ 10,000
Nature of Entry	Verbatim responses from 1990	Entries occurring twice in LONG
Misspellings present ?	Yes	Yes

5. Word-lexicon := words that constitute the entries in the phrase lexicon. (~ 6000 entries)

Lexicon Analysis

Tradeoffs between long and short lexicons

LEXICON :	LONG	SHORT
Hit Rate	High	Low
Confusion Rate	More	Less
Computational Speed	Low	High

Lexicon Analysis

Preprocessing :

- We chose to work with SHORT lexicons.
- Lexicon Reduction :
  - ❶ entries that contain misspellings (deliberately present)
  - ❷ duplicate entries
- Lexicon Expansion:
  - training truths were appended to lexicons.

Lexicons after preprocessing :

lexicon	*approx size (entries)*	*approx hit rate (for training truths)*
phrase_short.[*i*]	8,000	63 %
phrase_short.comb	20,000	66 %
phrase_long.[*i*]	40,000	73 %

i = 1,2,3

Lexicon Analysis

Field and length-wise analysis of hit rate in short phrase lexicon

FIELDS	*FIELD 1*	*FIELD 2*	*FIELD 3*	*OVERALL*
ONE WORD	90	93	96	93 (36)
TWO WORD	63	65	55	61 (44)
THREE + WORDS	23	20	8	15 (20)
OVERALL	64	71	56	

OBSERVATIONS

❶ Single word phrases constitute 36 % of all responses and have a high hit rate (> 90 %).

❷ Long phrases constitute 20 % of all responses and have a low hit rate (< 23 %).

CONCLUSIONS

① Lexicon-driven approach will fail for "long" phrases, owing to low hit rate.

② Long phrases are relatively rare.

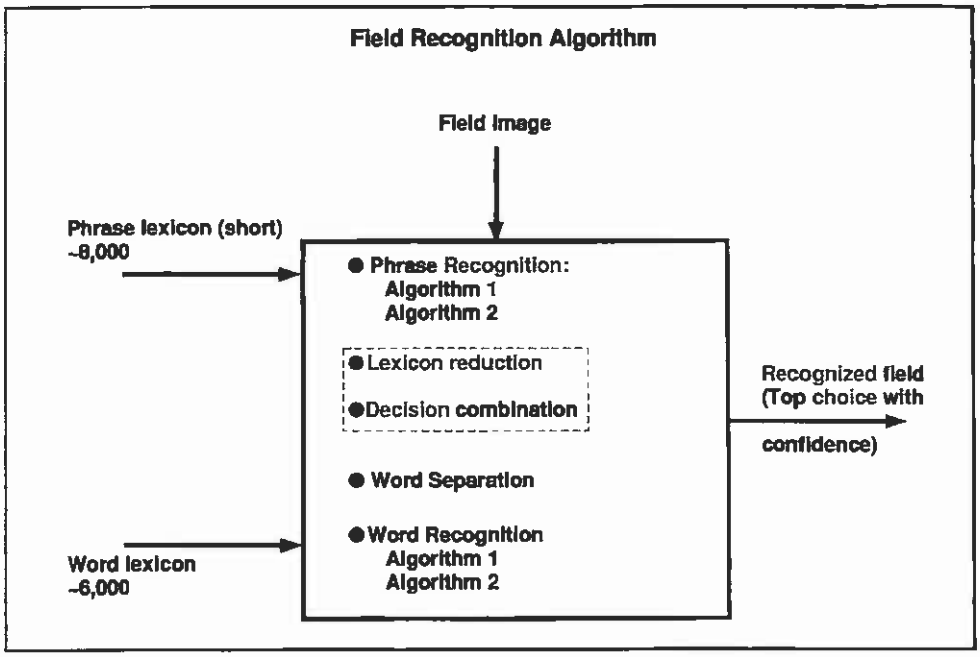
③ Lexicon-driven recognition strategy is viable if long phrases can be rejected.

④ Alternative : break phrases into words; identify each word separately.

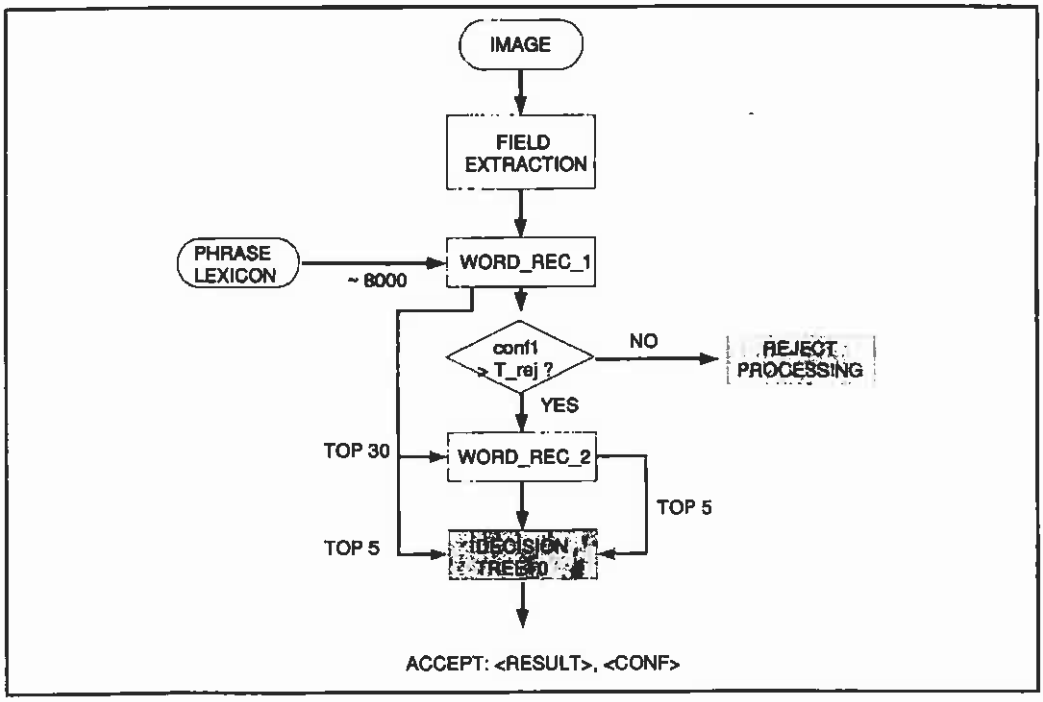
NIST Strategies

Strategy :	*Whole Phrase Recognition*	*Segment-then-recognize*	*Recognize-then-segment*
Segmentation	Treat entire image as single word	Split image into "words"	Use raw OCR of entire image
Recognition	Identify nearest phrase in phrase lexicon	Recognize each word individually using word lexicon	Find best sequence of words from word lexicon
Lexicon used	Phrase lexicon	Word lexicon	Word lexicon
Speed	Fast	Fast	Slow
Performance	Limited by phrase hit rate	Limited by word break determination accuracy and word hit rate	Limited by word hit rate
Postprocessing	Not called for	Essential	Essential

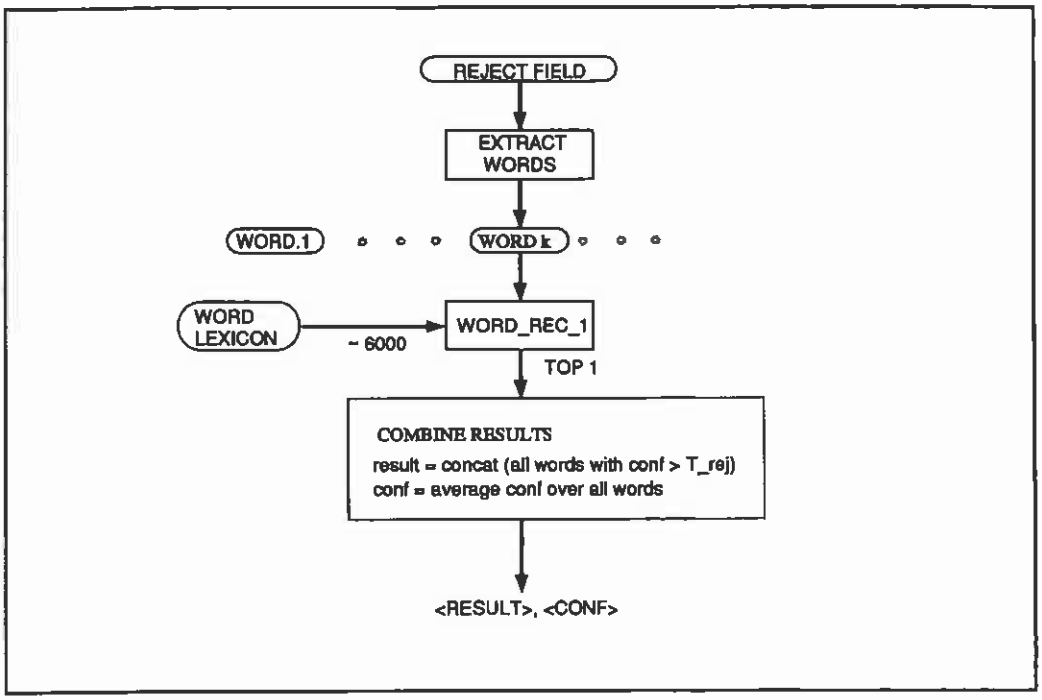
Field Recognition



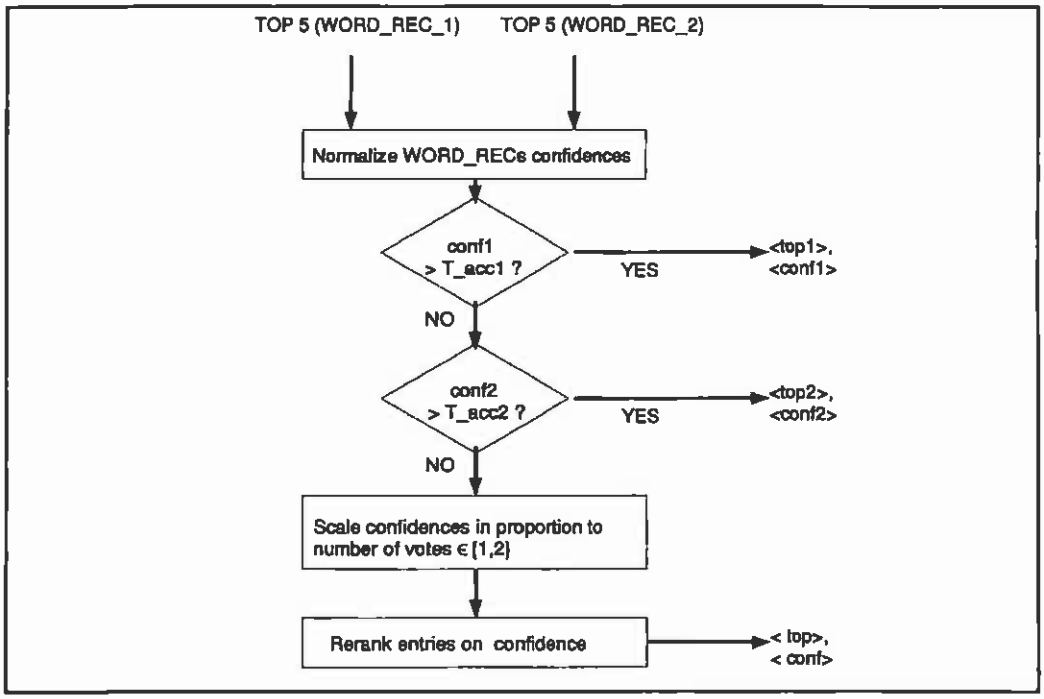
CEDAR_0 Control Flow



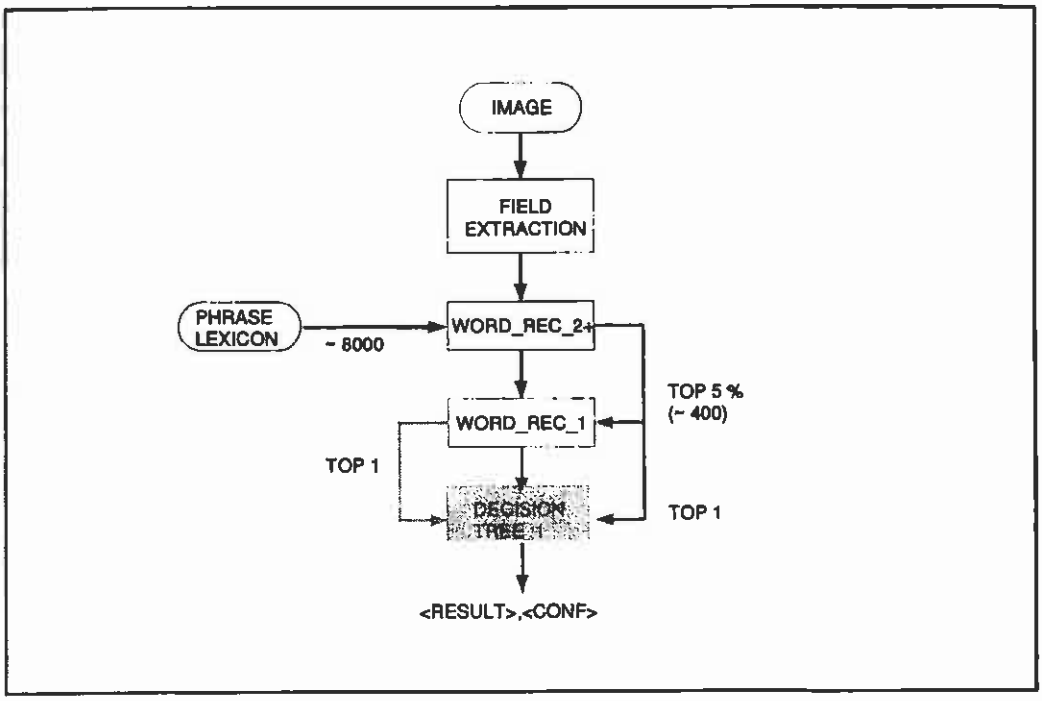
CEDAR_0 Reject Processing



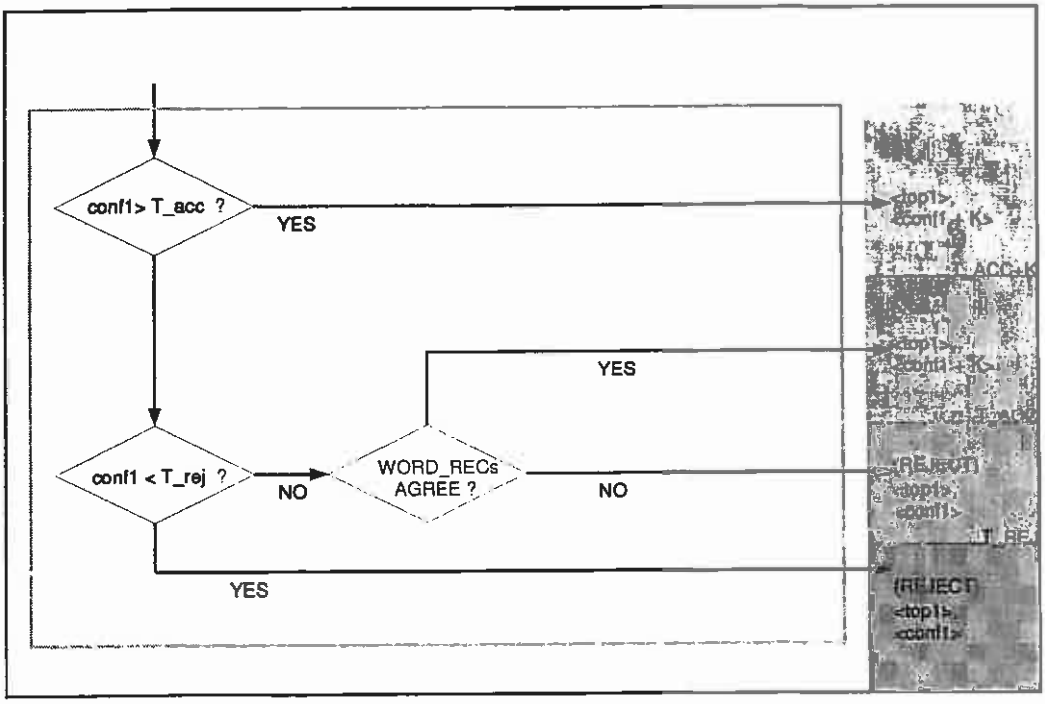
Decision Tree - 0



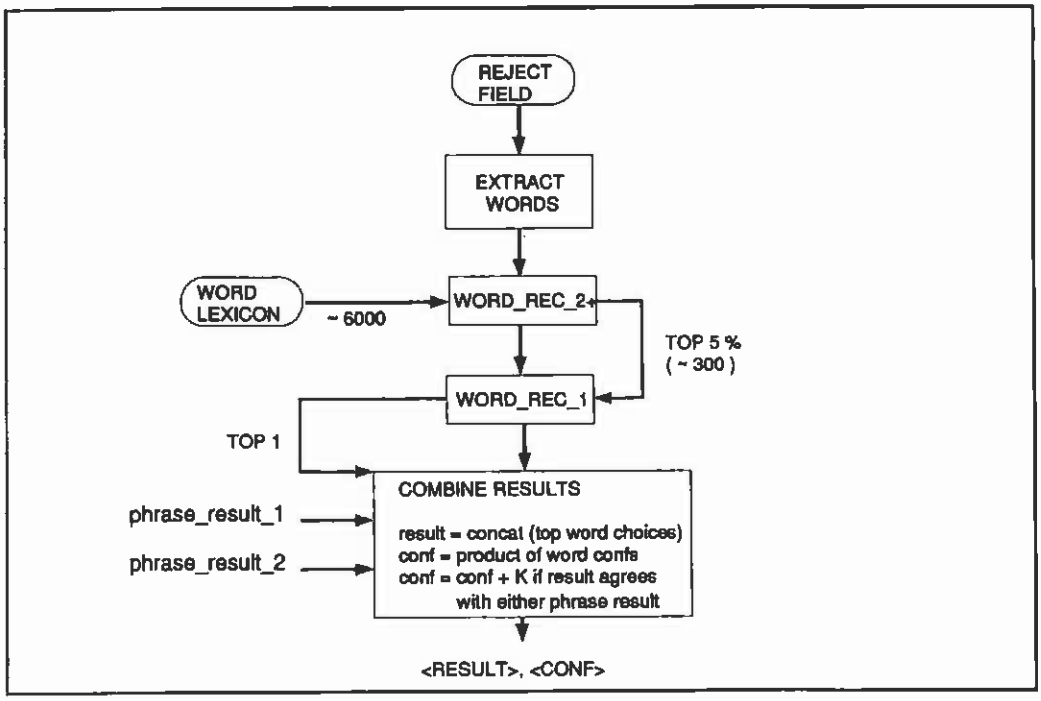
CEDAR_1 Control Flow



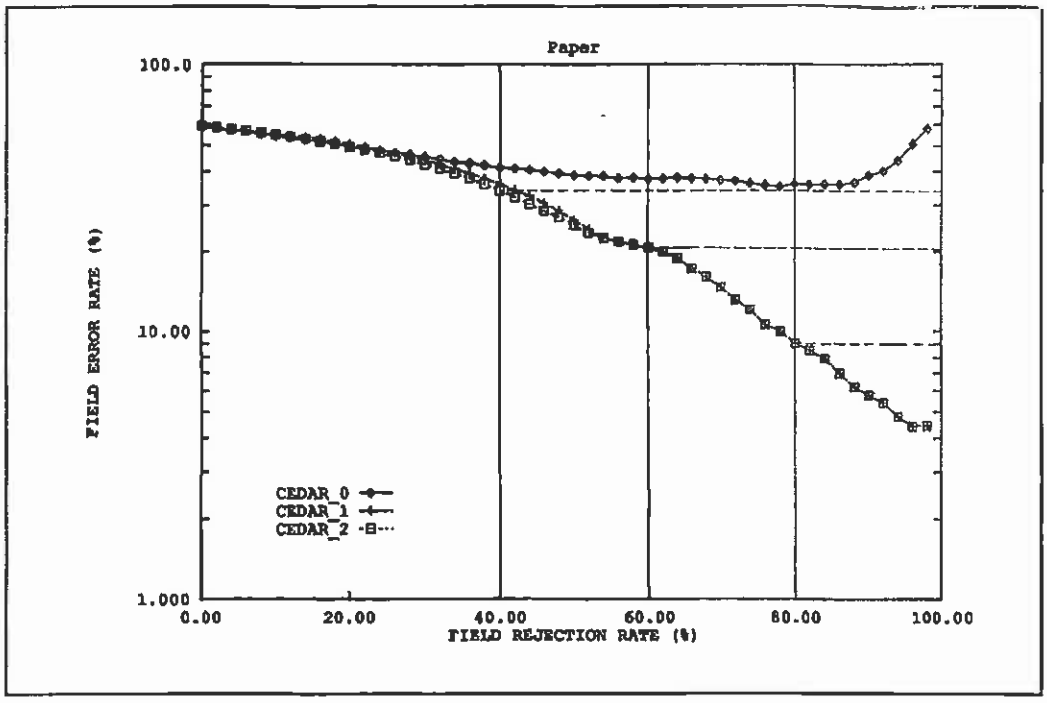
Decision Tree_1



CEDAR_2 Reject Processing

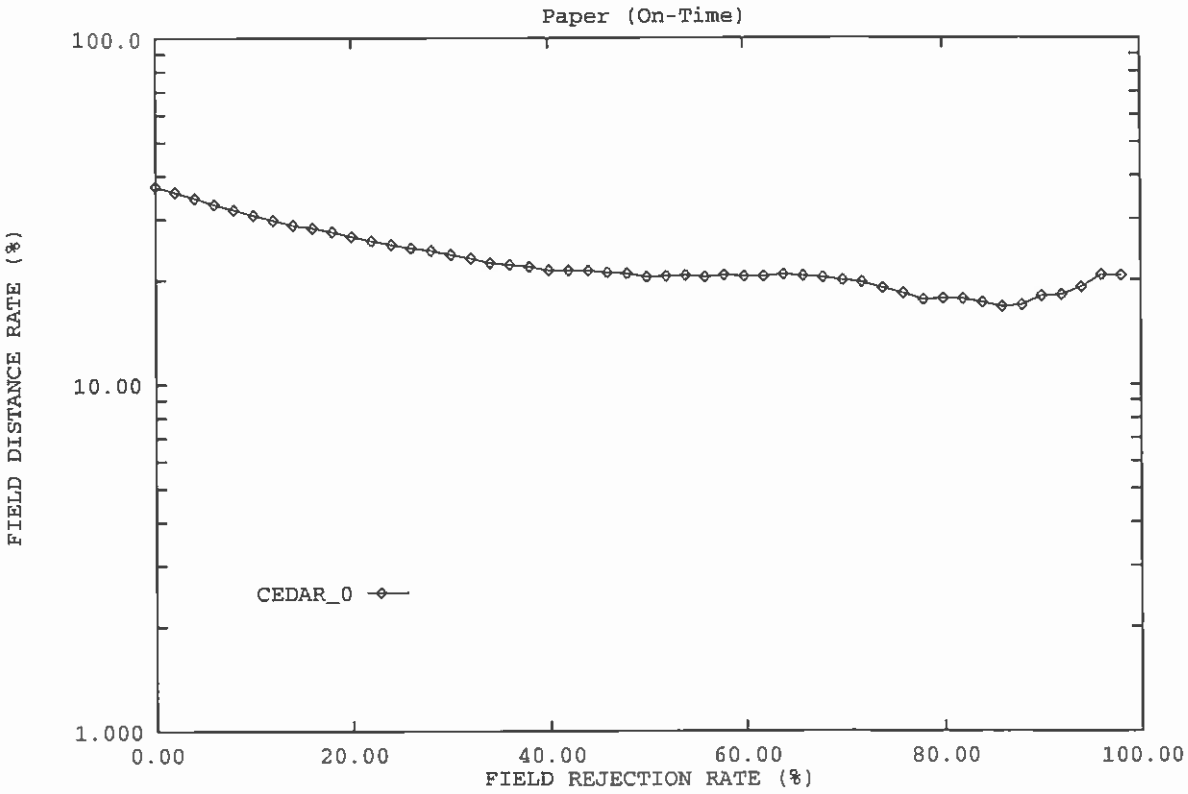


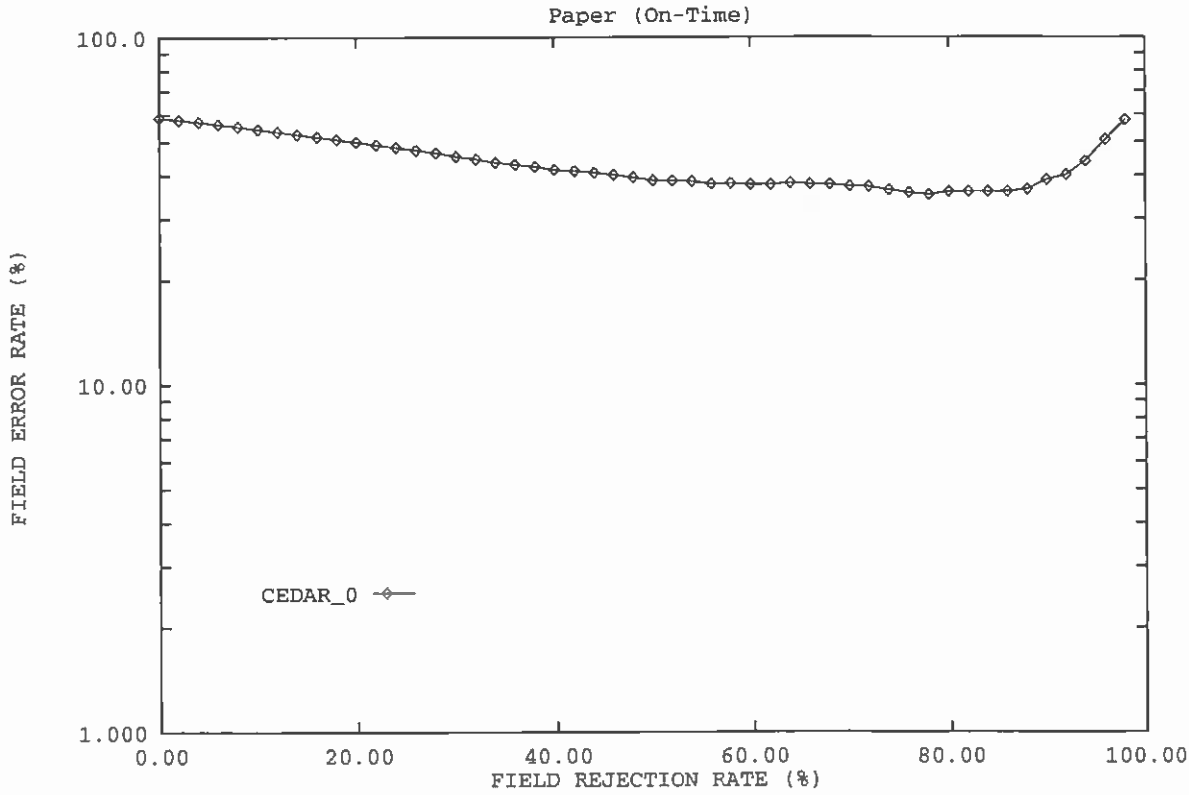
Performance Analysis

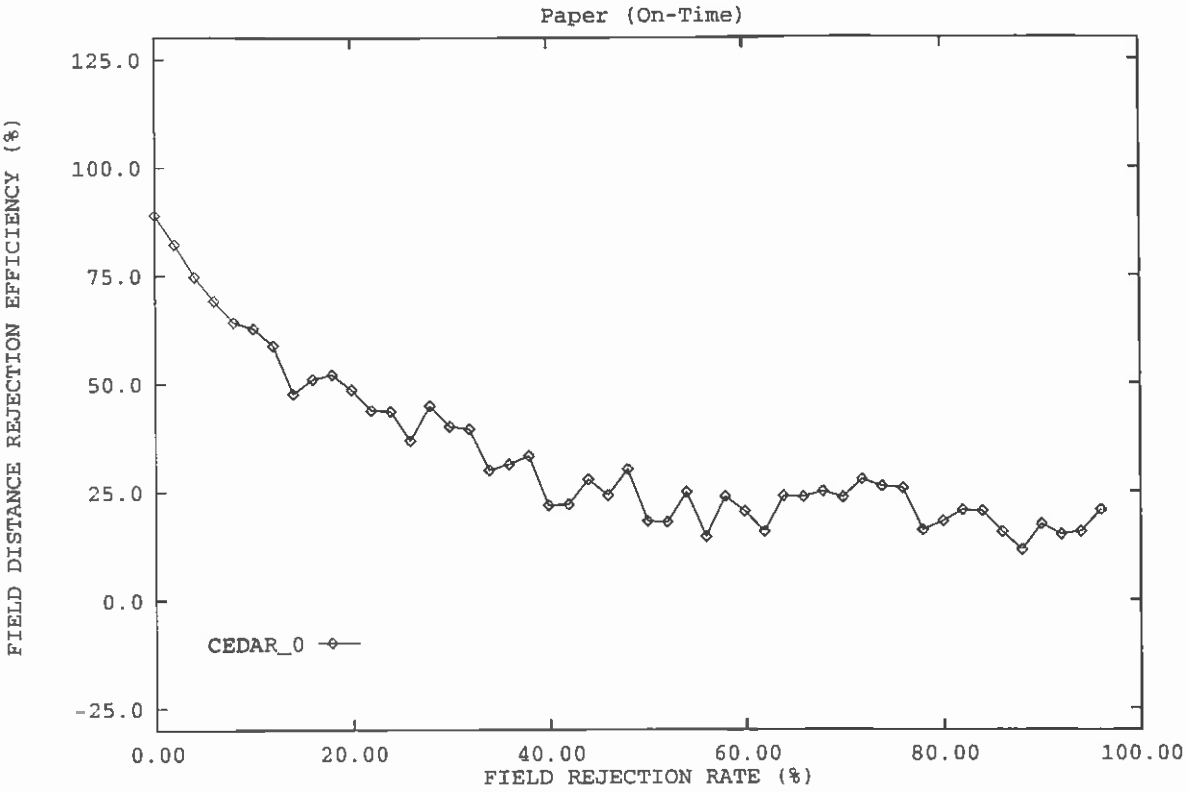


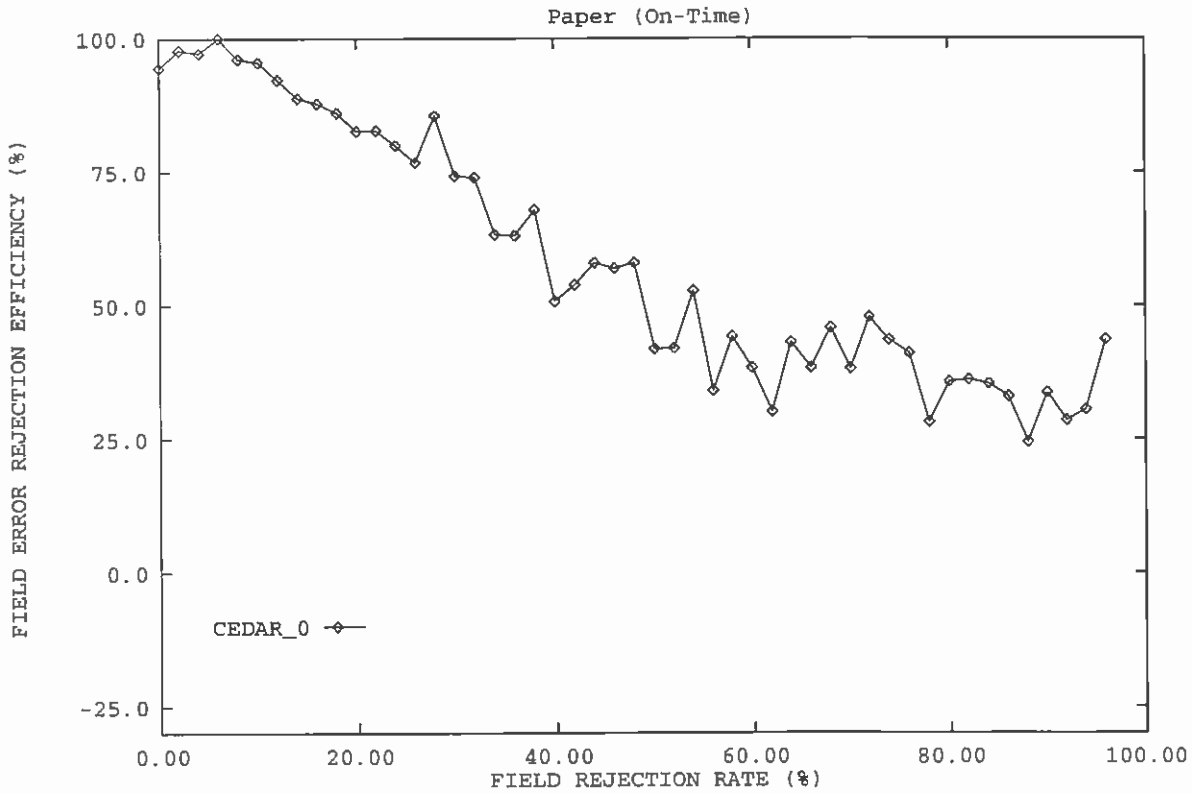
Future Work

1. Improve field extraction and preprocessing
2. Improve word extraction
3. Train classifiers on discrete / touching characters from census forms
4. Apply word collocation postprocessing
5. Apply "recognize-then-segment" approach
6. Improve recognition algorithm combination


Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CEDAR_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CEDAR_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CEDAR_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CEDAR_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

NIST 2
Read Performance Test

CGK Computer Gesellschaft Konstanz mbH
Department Product Development
Max-Stromeyer-Str. 116
78467 Konstanz
Germany

CGK NIST -
Forms - Processing

Line Extraction Process

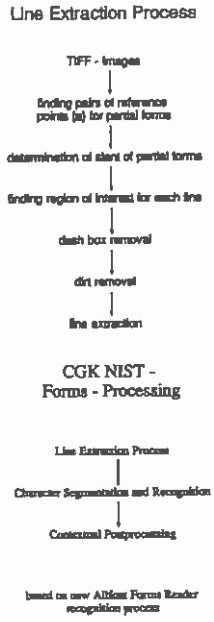
↓

Character Segmentation and Recognition

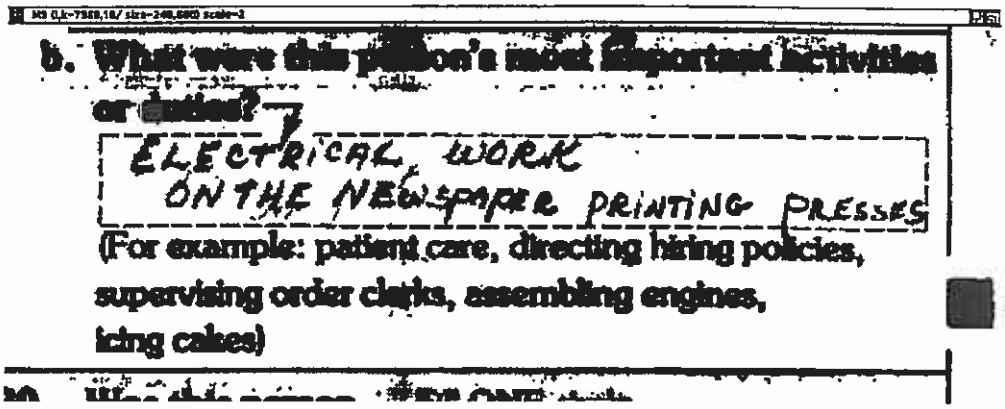
↓

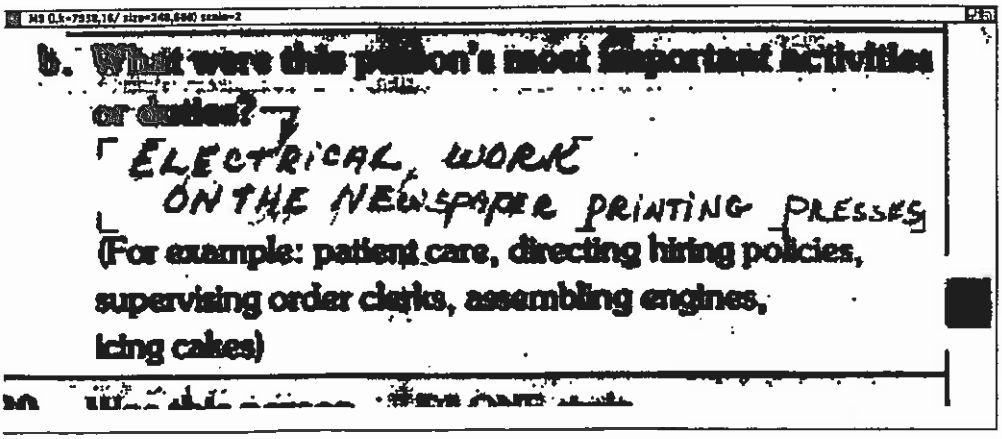
Contextual Postprocessing

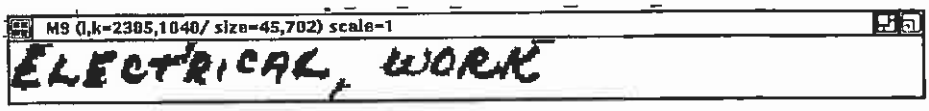
based on new Allfont Forms Reader
recognition process

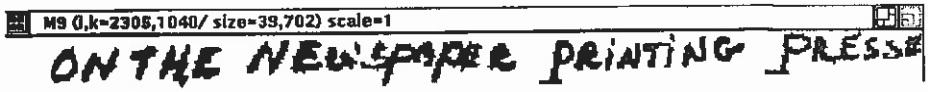


example of line finding and extraction






M9 (J,k=2385,1040/ size=45,702) scale=1
ELECTRICAL WORK


M9 (J,k=2305,1040/ size=39,702) scale=1
ON THE NEWSPAPER PRINTING PRESSE

Character Recognition Process

Line Slant Normalization

↓

Algorithm to split the line into segments

↓

Character count estimation of the word segments

↓

Finding segmentation column candidates by searching minima and maxima in upper and lower contour of the line segments

↓

Finding split pathes along the segmentation columns for the extraction of overlapping and connected character candidates

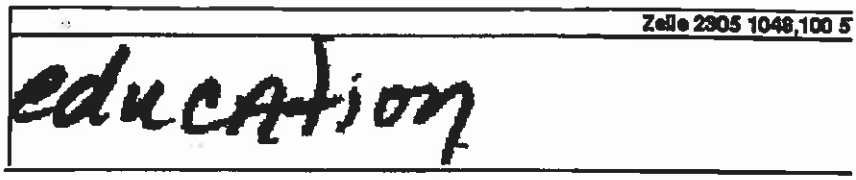
↓

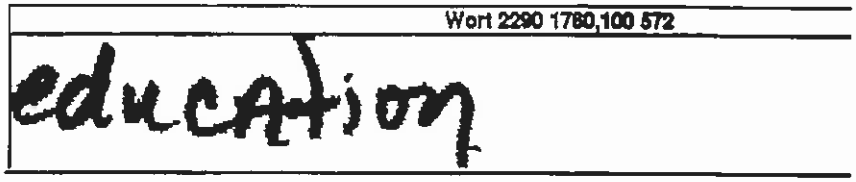
Statistical classification of the character candidates

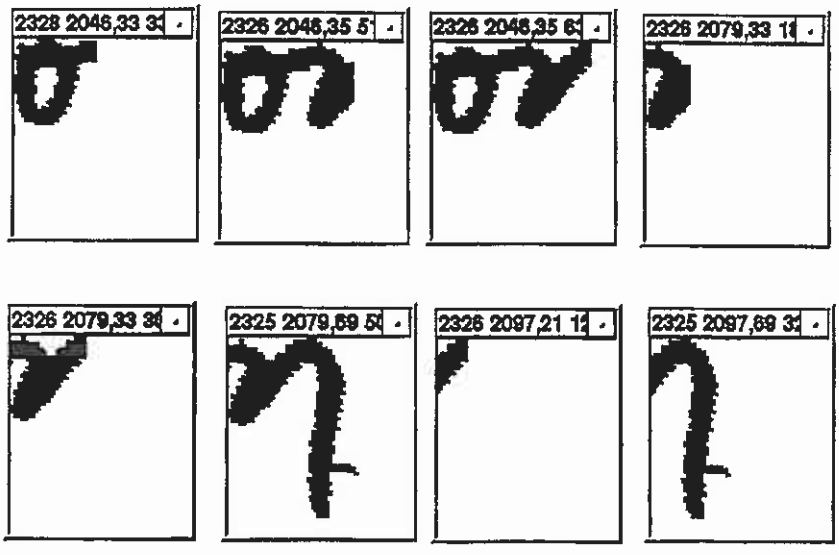
↓

Graph search algorithm to find the result string with the best character candidates

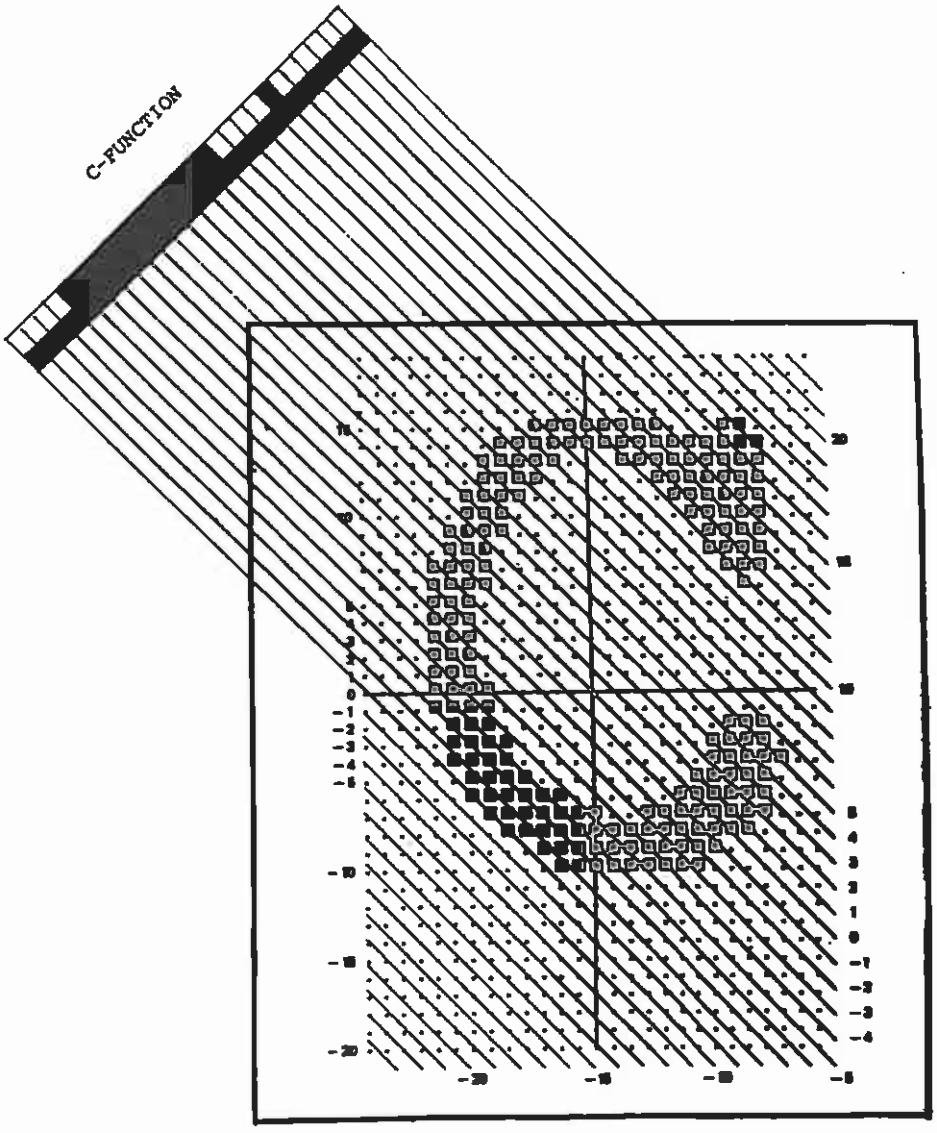
example of slant normalization and seperation proposals



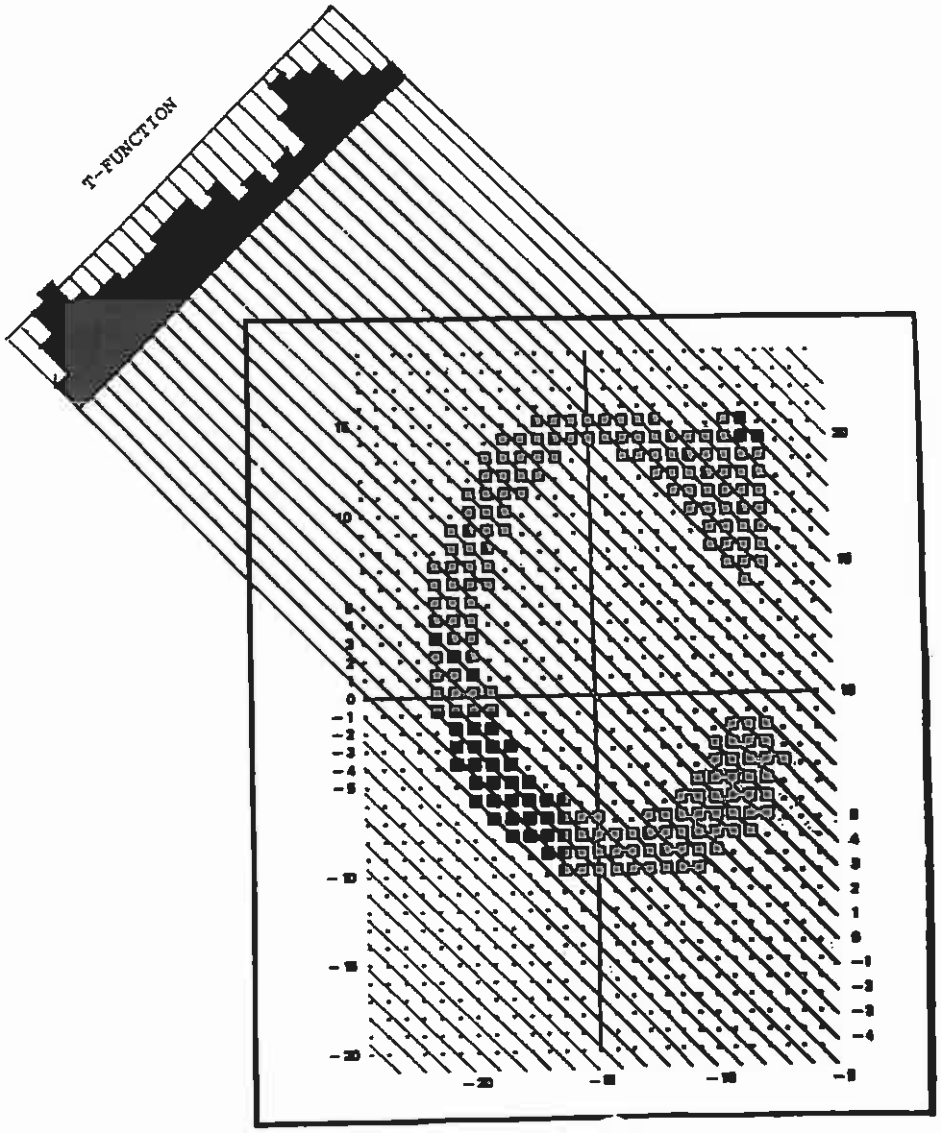




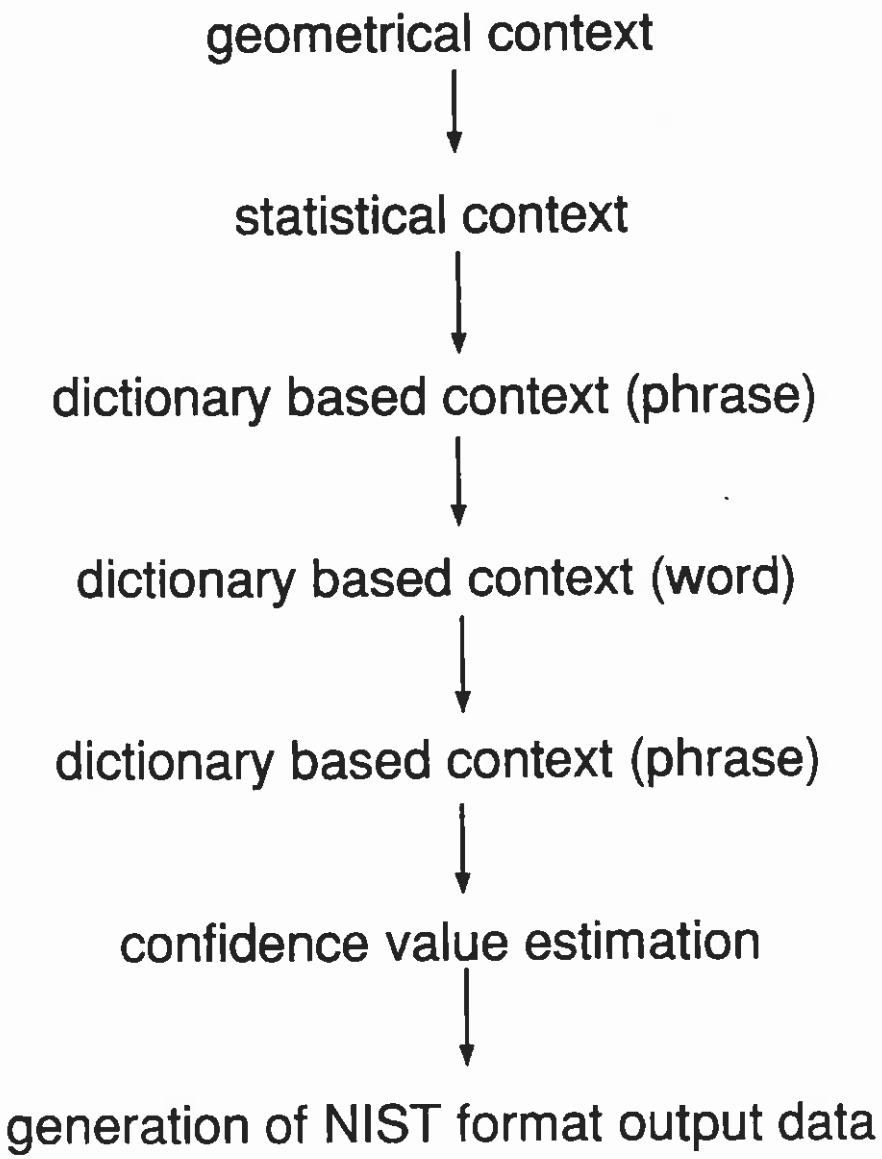
FIRST STEP IN GLOBAL FEATURE EXTRACTION:



FIRST STEP IN GLOBAL FEATURE EXTRACTION:



Contextual Postprocessing



Nist 2 - Read Performance Test

What else could be done:

1. Enhancements in line extraction

2. Enhancements in character segmentation by using more intelligent seperation candidate detection

3. **most important**: interaction between character recognition and seperation and dictionary based post-processing

4. processing of multiple line fields

5. Enhancements in confidence value generation

Nist 2 - Read Performance Test

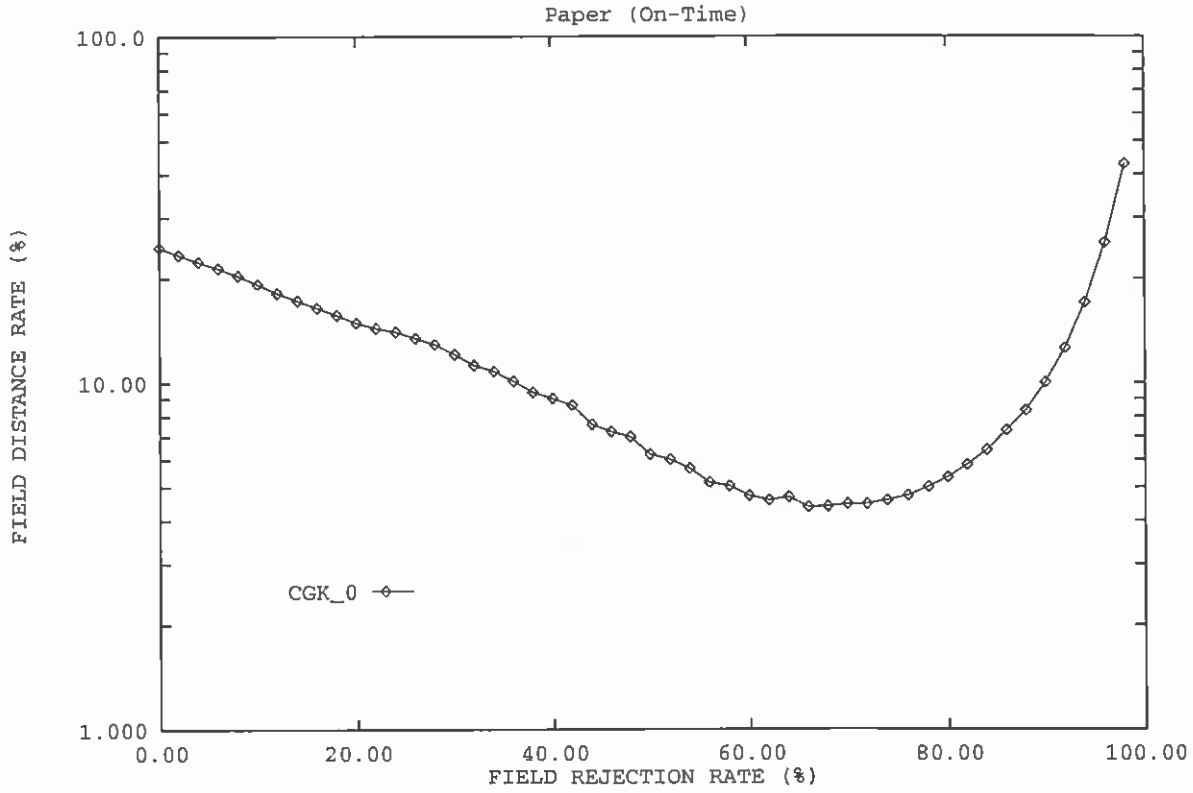
Problems:

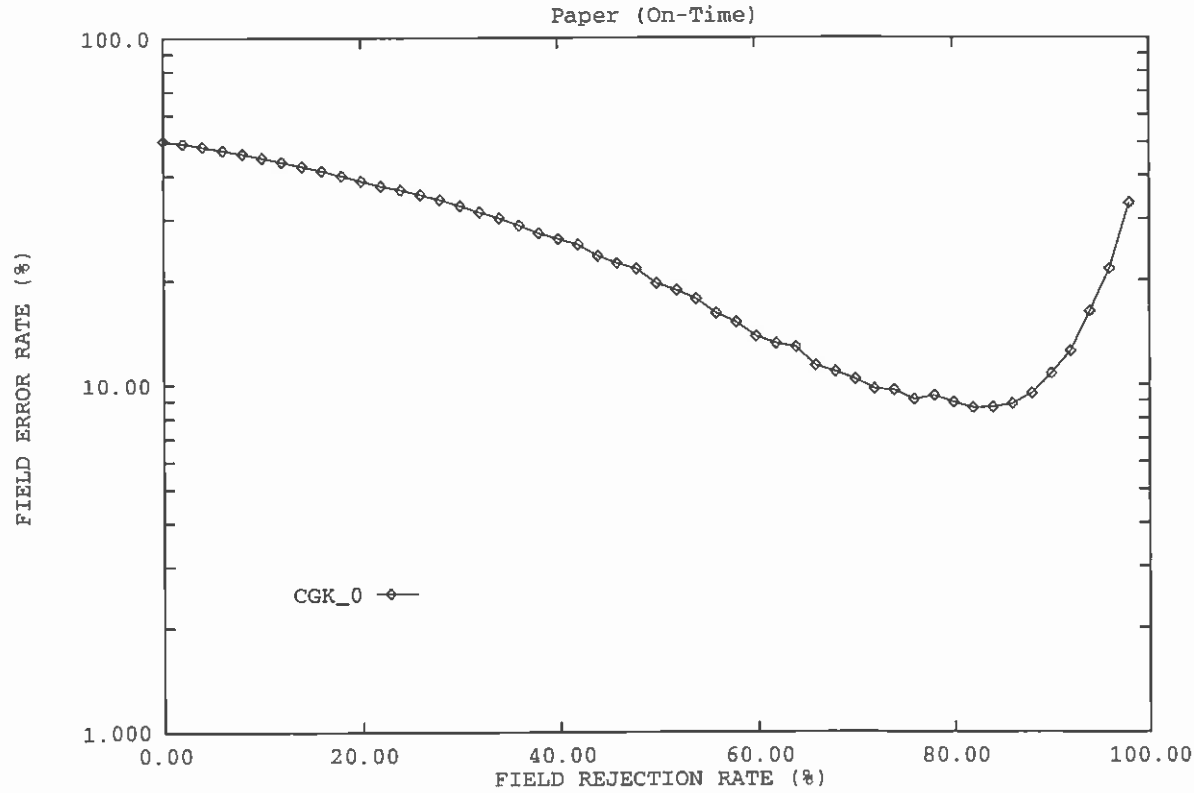
1. Scan and Image Quality

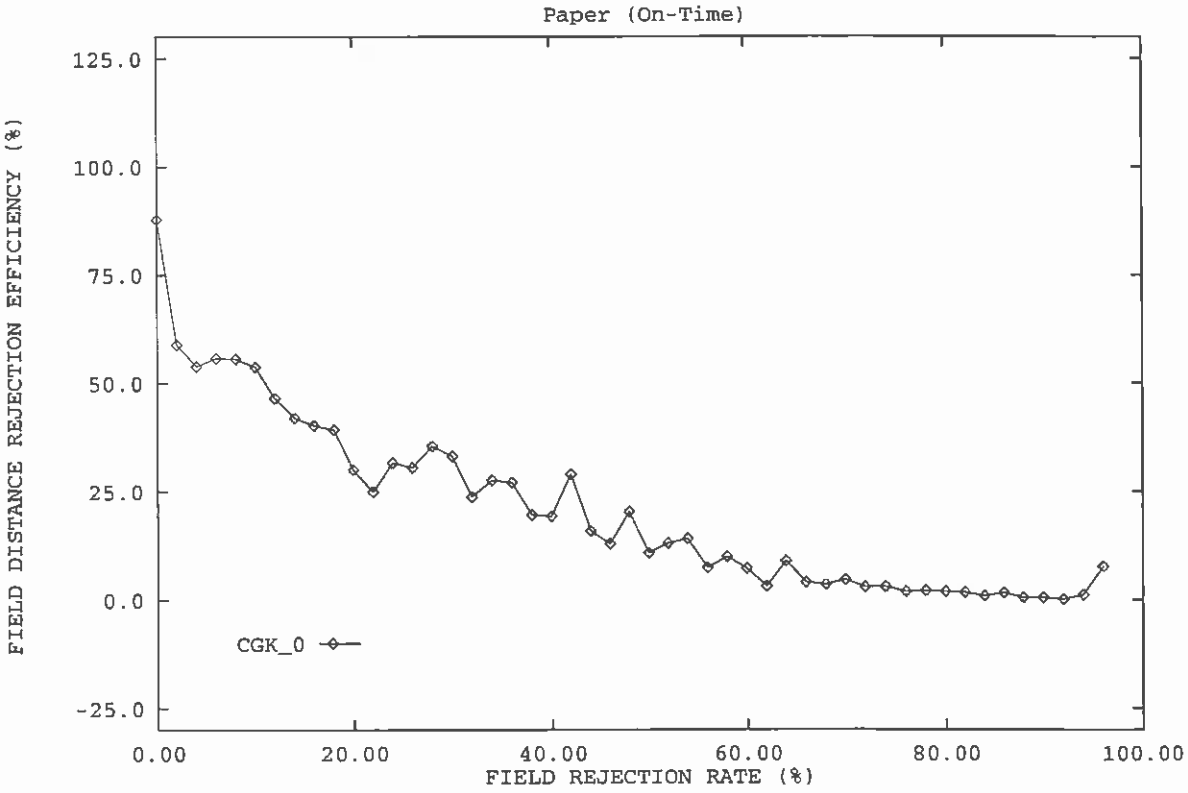
2. Page design (5 not aligned forms)

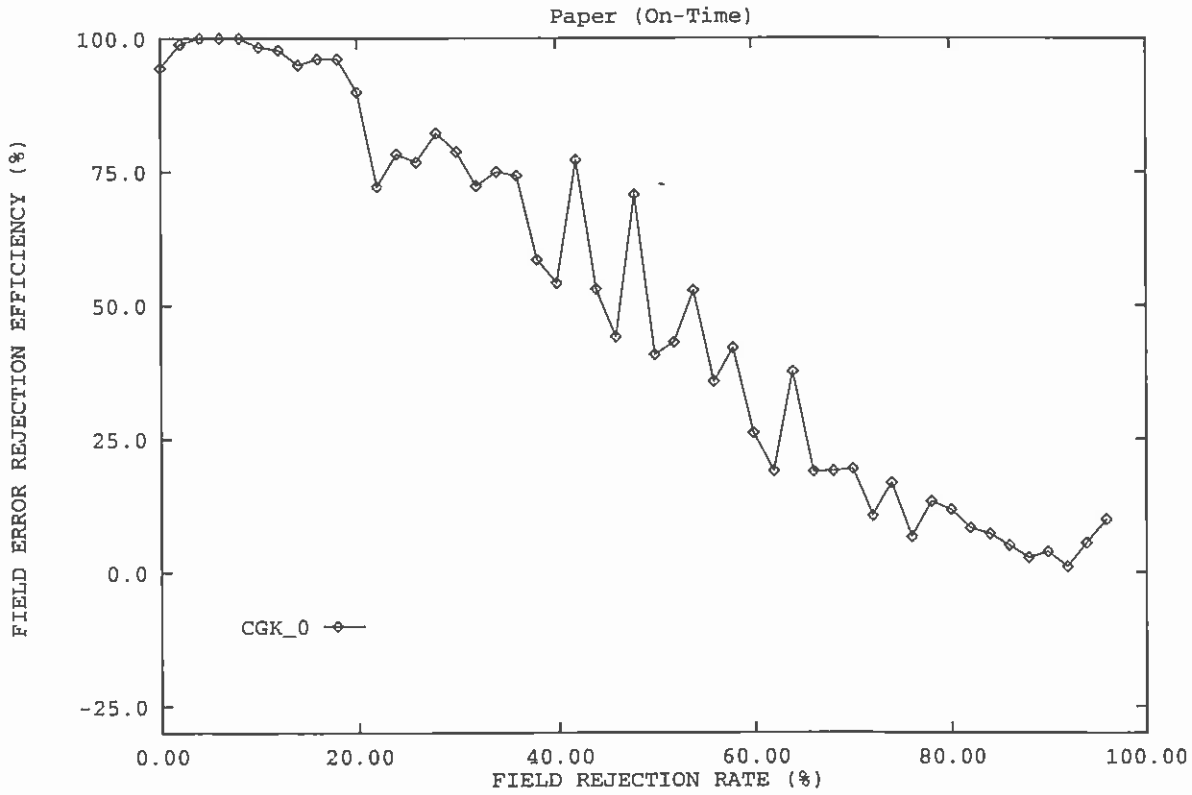
3. Forms design not suitable

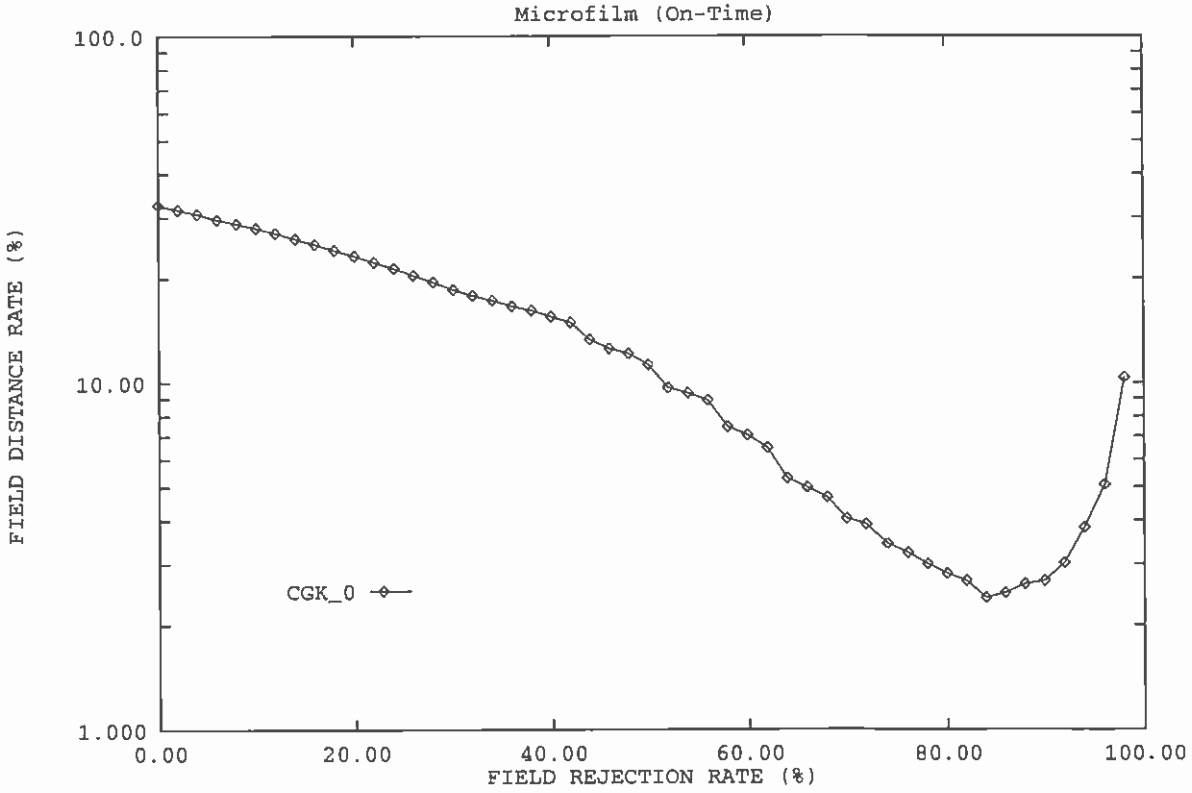
4. Dictionary quality: How to deal with misspellings and abbreviations

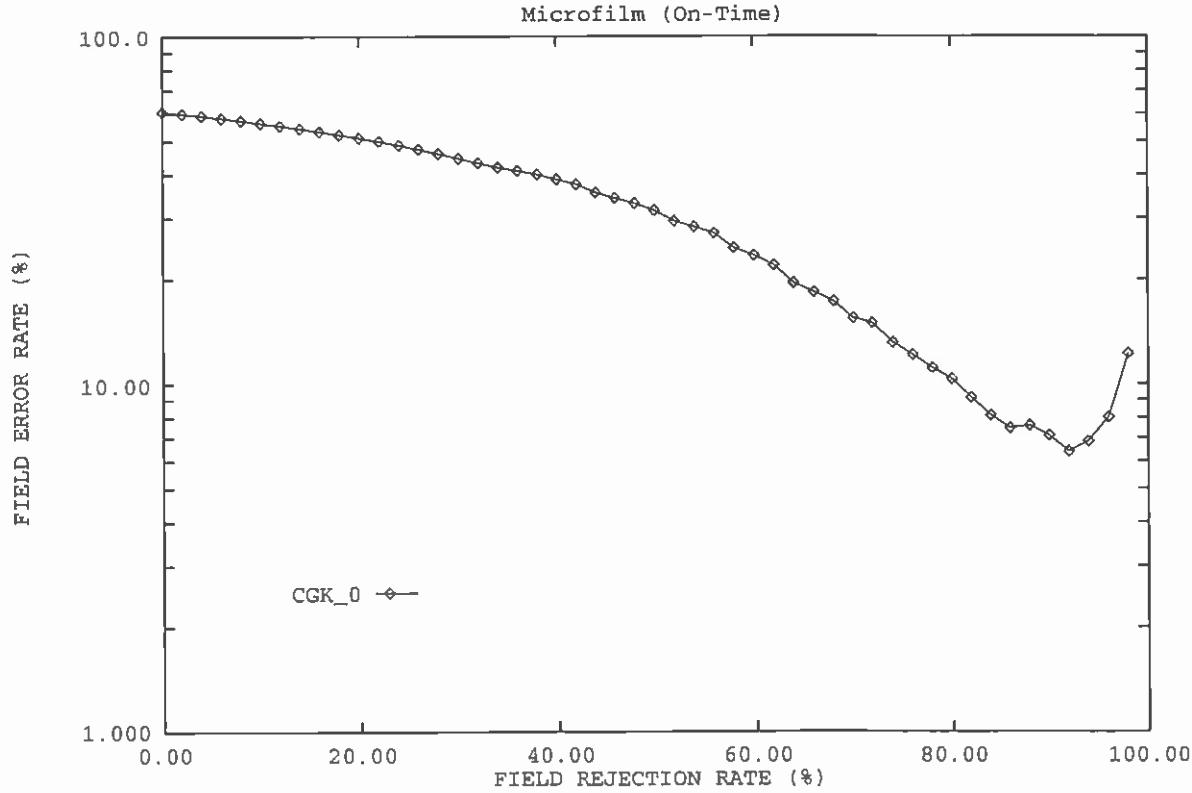

Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

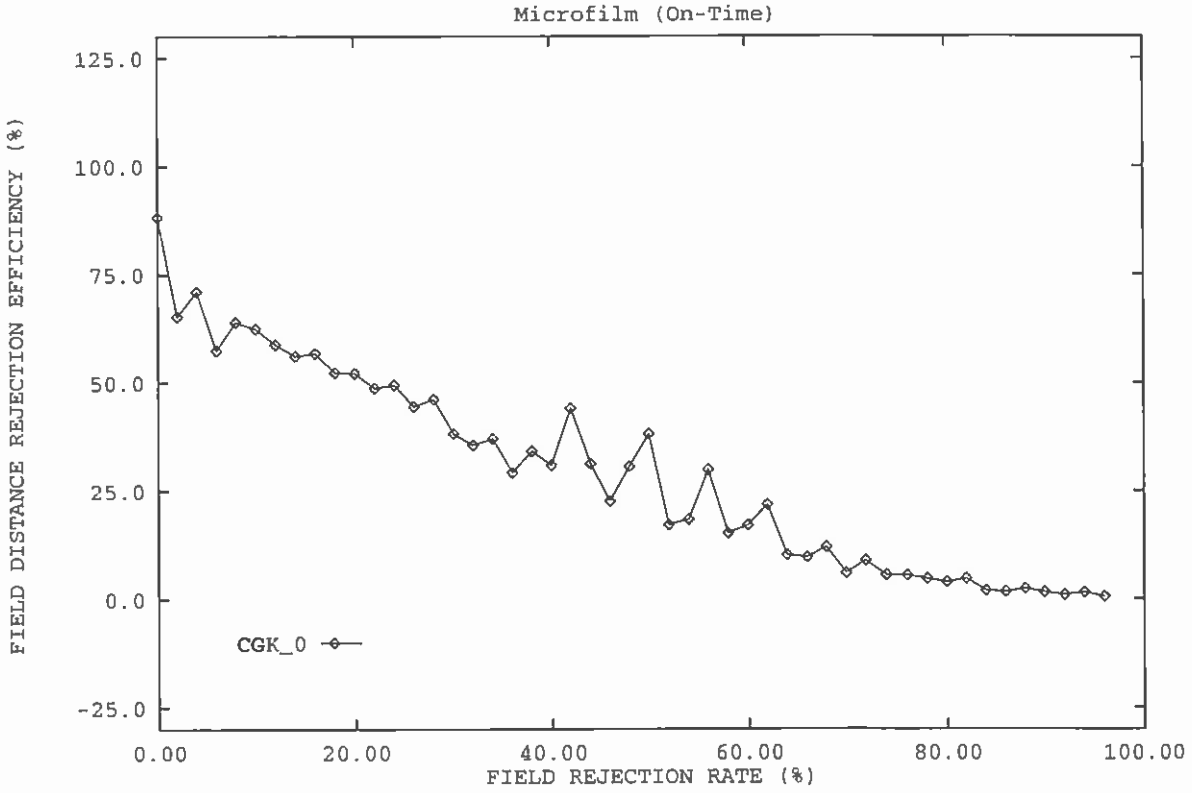

Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

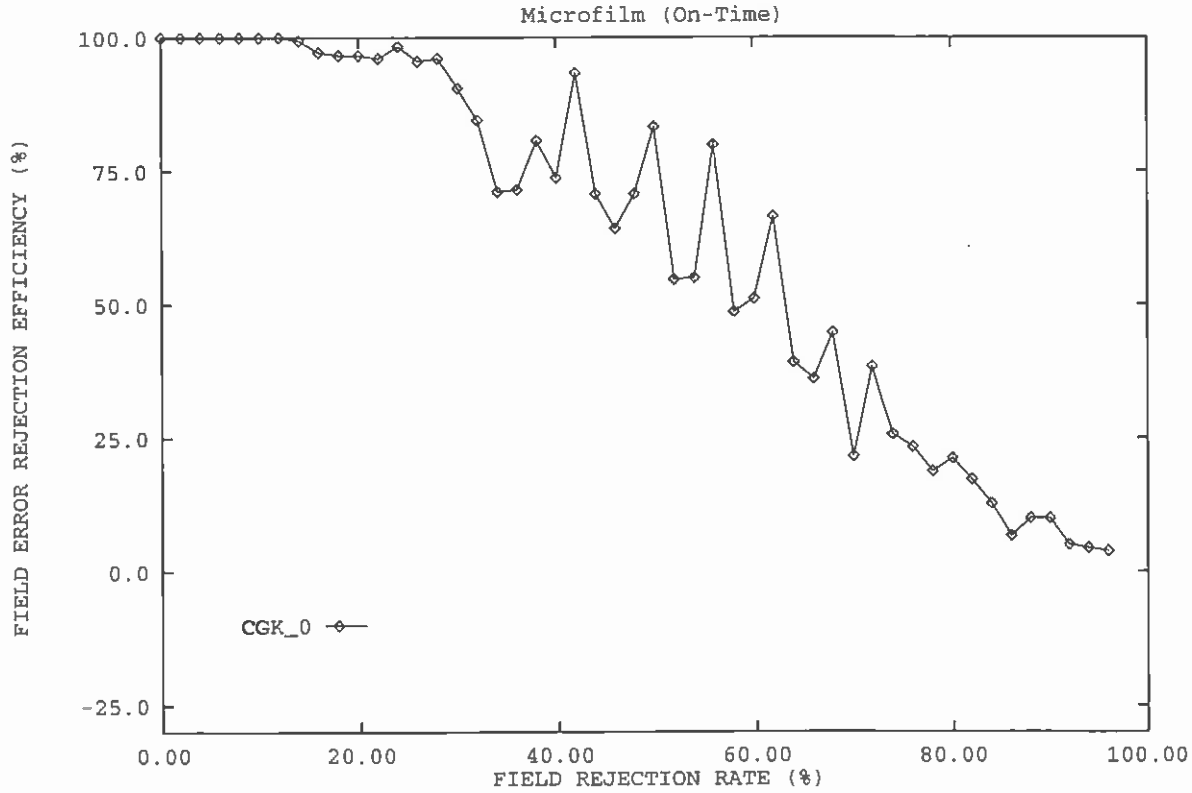

Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

Census Form Recognition System

Environmental Research Institute of Michigan
ERIM

Andy Gillies
Dan Hepp
Rich Rovner
Peggy Ganzberger
Mark Hamilton

2nd Census OCR Systems Conference

Background

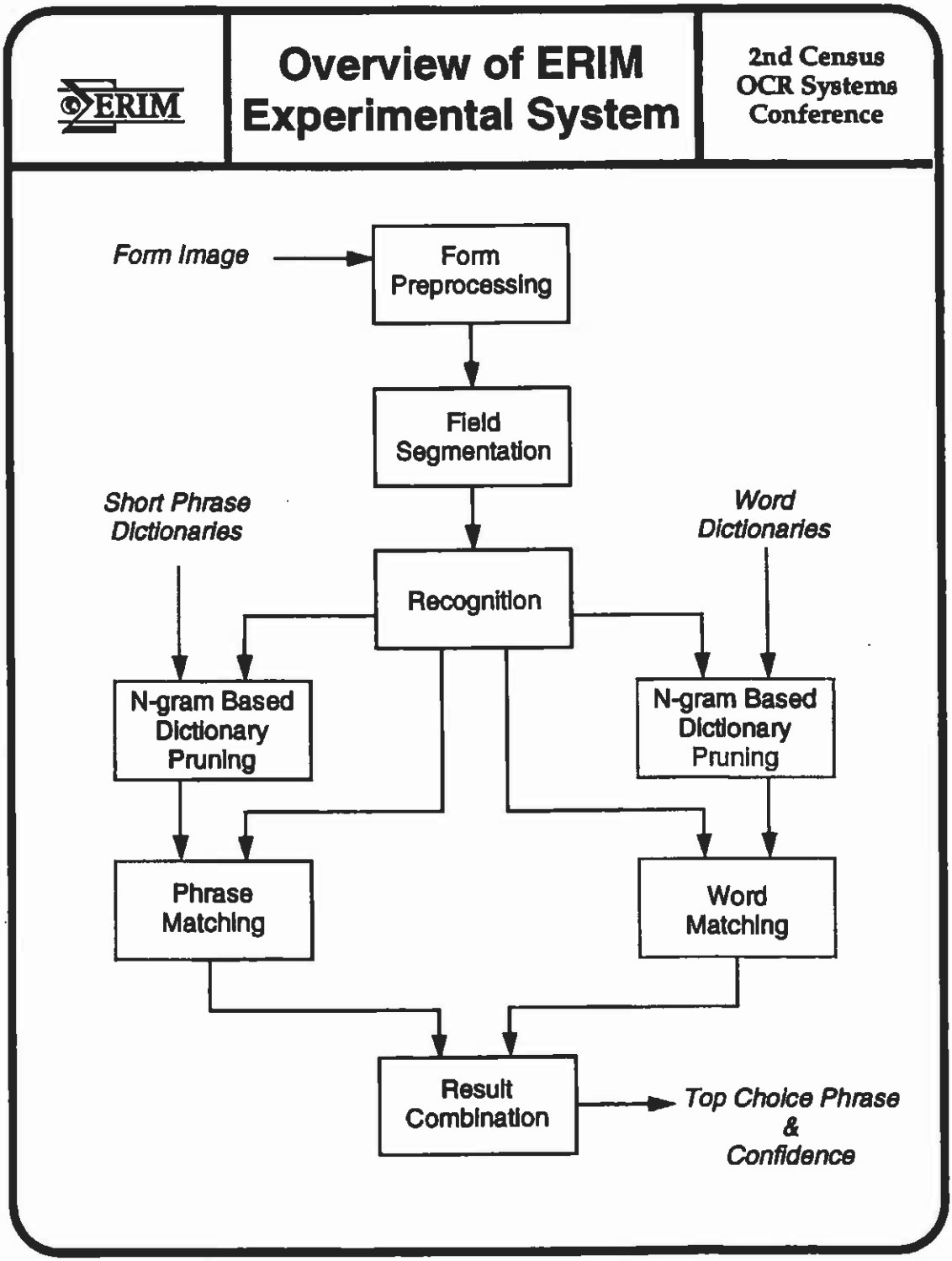
- **May 1992 - 1st Census OCR Systems Conference**
  - Isolated character recognition
  - Training data
    - 2100 writers (full-time census workers)
    - 223,125 digits / 44,951 uppercase / 45,313 lowercase
  - Test data
    - 500 writers (high school students)
    - 58,646 digits / 11,941 uppercase / 12,000 lowercase
  - Results per character 96% digits / 95% uppercase / 86% lowercase
- **February 1994 - 2nd Census OCR Systems Conference**
  - Field extraction and recognition
  - Training data - occupation fields from 1990 Census Form
    - Microfilm 50 x 100 x 5 x 3 = 75,000 fields
    - Paper 12 x 100 x 5 x 3 = 18,000 fields
  - Test data
    - Microfilm 6 x 100 x 5 x 3 = 9,000 fields
    - Paper 6 x 100 x 5 x 3 = 9,000 fields
  - Results per field: 34% accept with 8% error microfilm / 42% accept with 3.5% error paper

ERIM

Form Image Preprocessing

2nd Census OCR Systems Conference

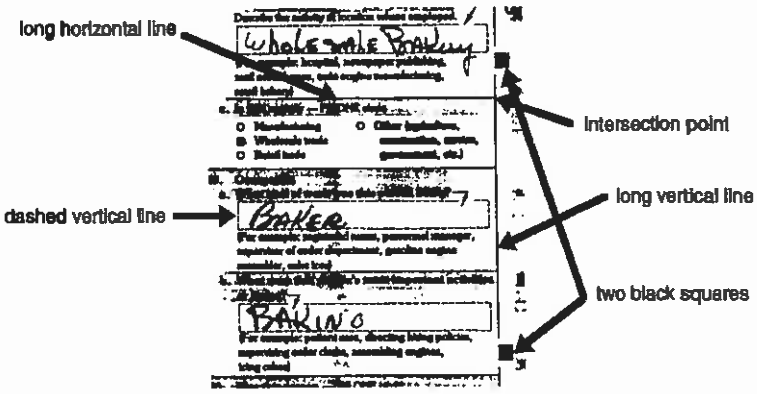
- **Registration of form image to known coordinates**
  - form type classification
  - cue detection
  - image resampling
  - region of interest extraction
- **Extraction of field contents**
  - background removal
  - stroke repair


ERIM
Overview of ERIM Experimental System
2nd Census OCR Systems Conference
Form Image
Form Preprocessing
Field Segmentation
Short Phrase Dictionaries
Word Dictionaries
Recognition
N-gram Based Dictionary Pruning
N-gram Based Dictionary Pruning
Phrase Matching
Word Matching
Result Combination
Top Choice Phrase & Confidence

ERIM

Cue Detection

2nd Census OCR Systems Conference

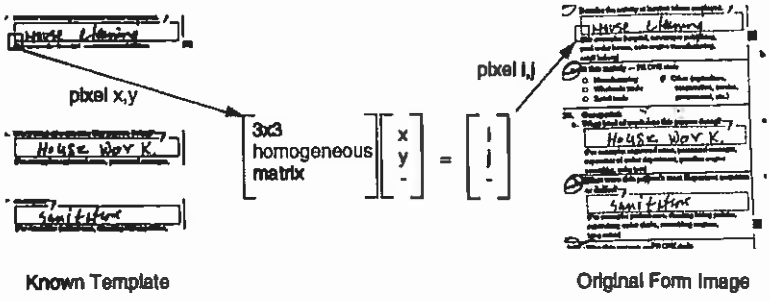


- **morphological processing detects:**
  - rotation
  - scale
  - translation

ERIM

Image Resampling

2nd Census OCR Systems Conference



- **resample from known template to form image**
- **region of interest only**

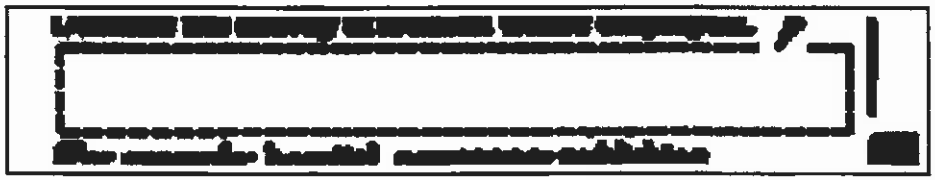
ERIM

Extraction of Field Contents

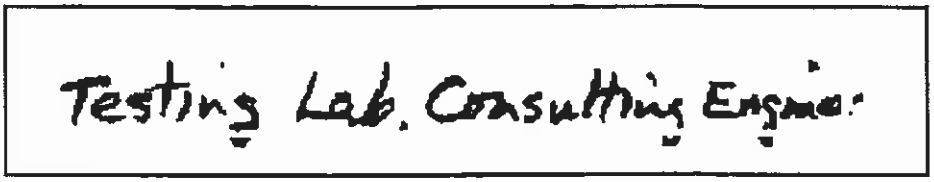
2nd Census OCR Systems Conference

Original Field Image

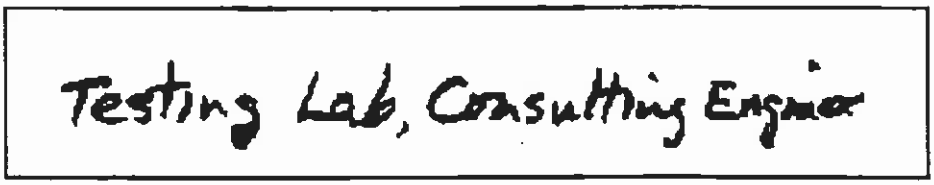
Field Mask



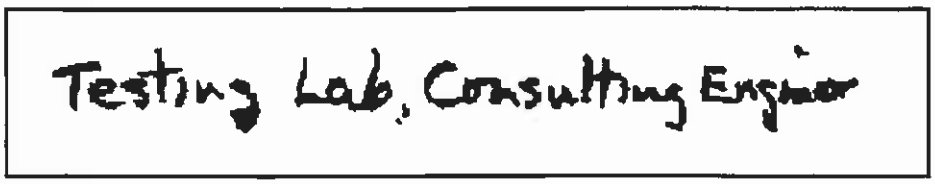
Remove Mask Pixels

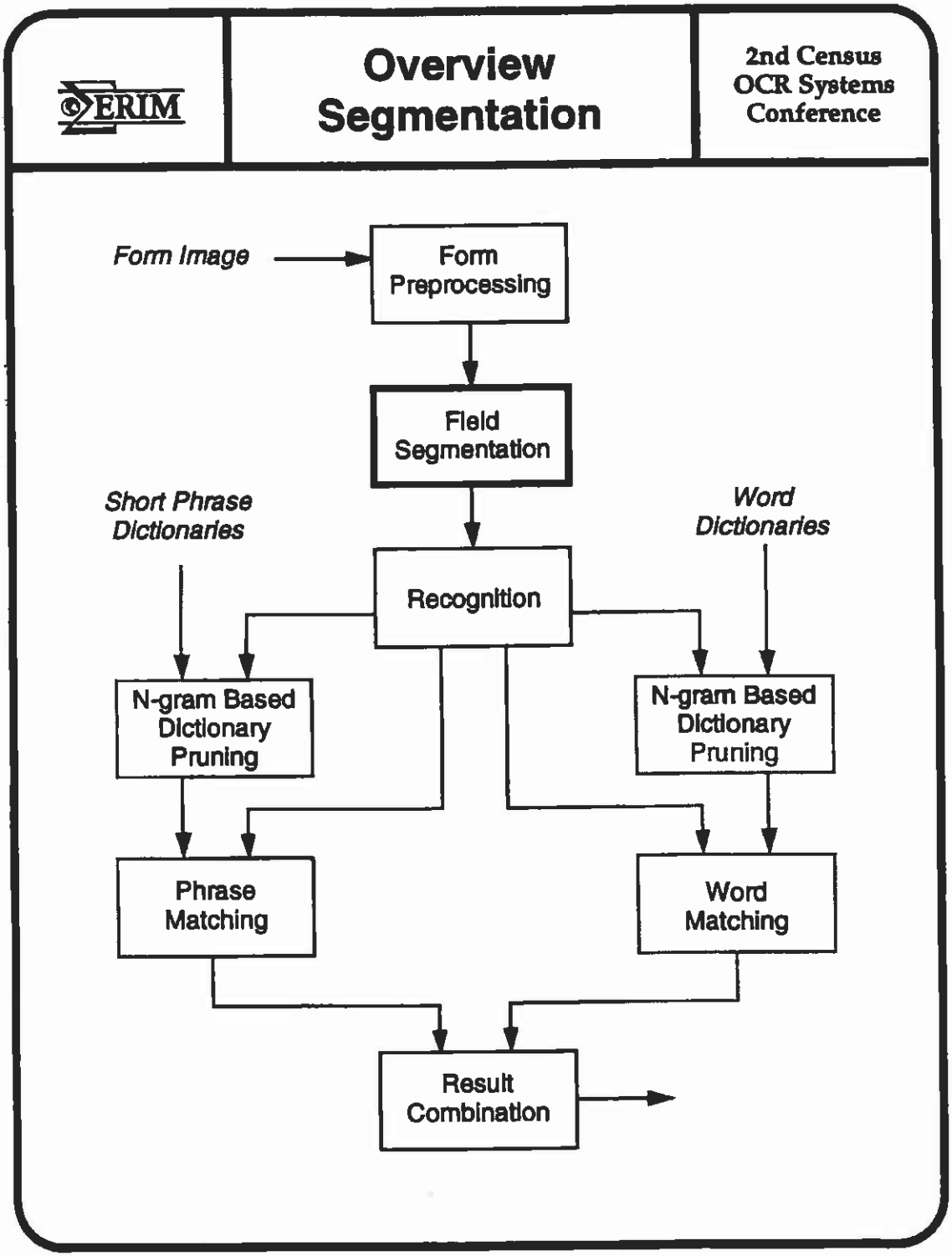


Repair Strokes



Filter & Correct Slant




ERIM
Overview
Segmentation
2nd Census
OCR Systems
Conference
Form Image
Form
Preprocessing
Field
Segmentation
Short Phrase
Dictionaries
Word
Dictionaries
Recognition
N-gram Based
Dictionary
Pruning
N-gram Based
Dictionary
Pruning
Phrase
Matching
Word
Matching
Result
Combination

ERIM

Field Segmentation

2nd Census OCR Systems Conference

- Goal is oversegmentation of characters
- Based on outer contours of word images
- Uses local minima and maxima as segmentation points
- Result is primitive segments

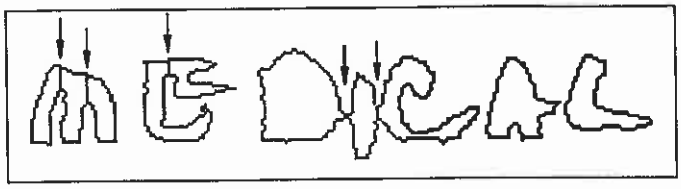
ERIM

Field Segmentation

2nd Census OCR Systems Conference

Contours of Field Image

Segmentation Points



Resulting Primitive Segments

ERIM

Unions

2nd Census OCR Systems Conference

- A union is a character hypothesis
- Formed by joining from 1 to 4 primitive segments
- Rule-based sanity checks
- Character Recognition is performed for each union

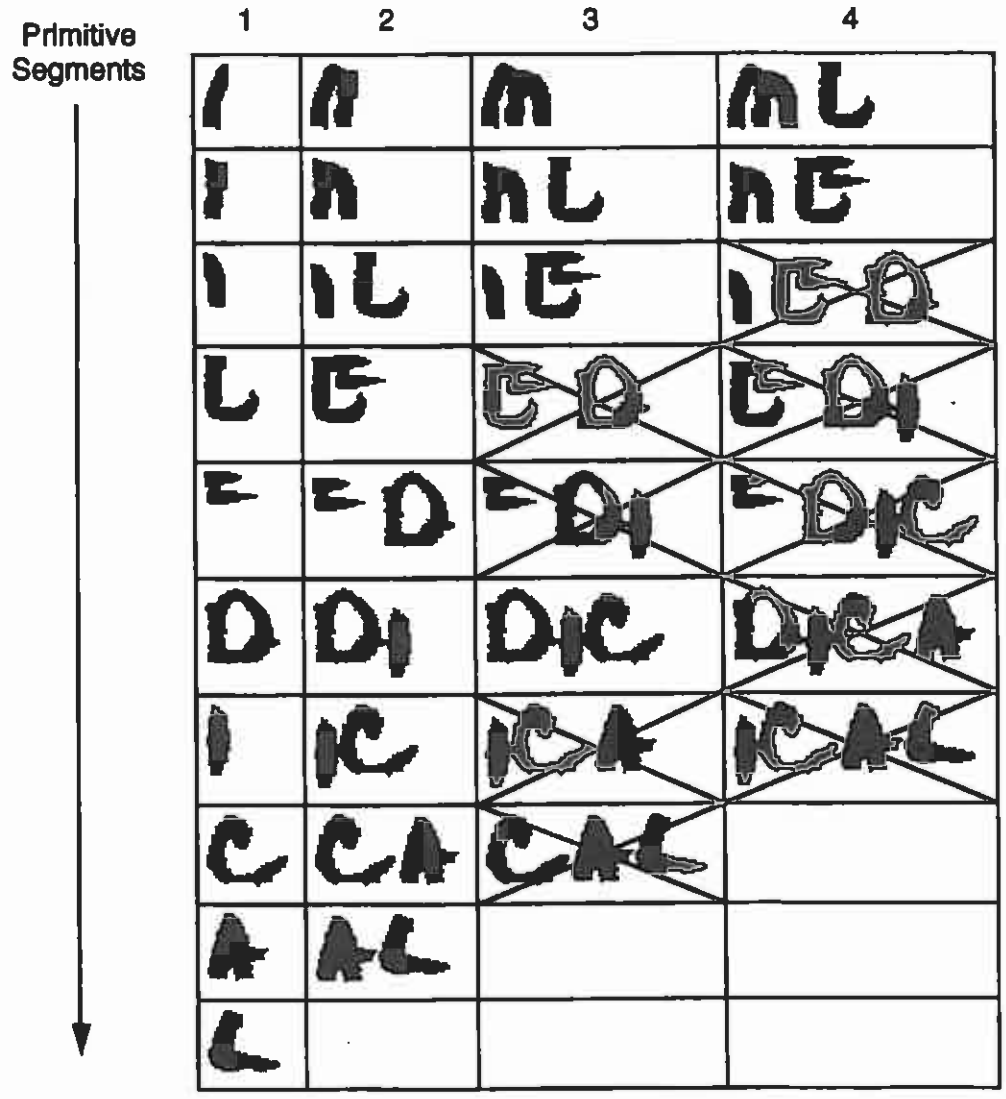
ERIM

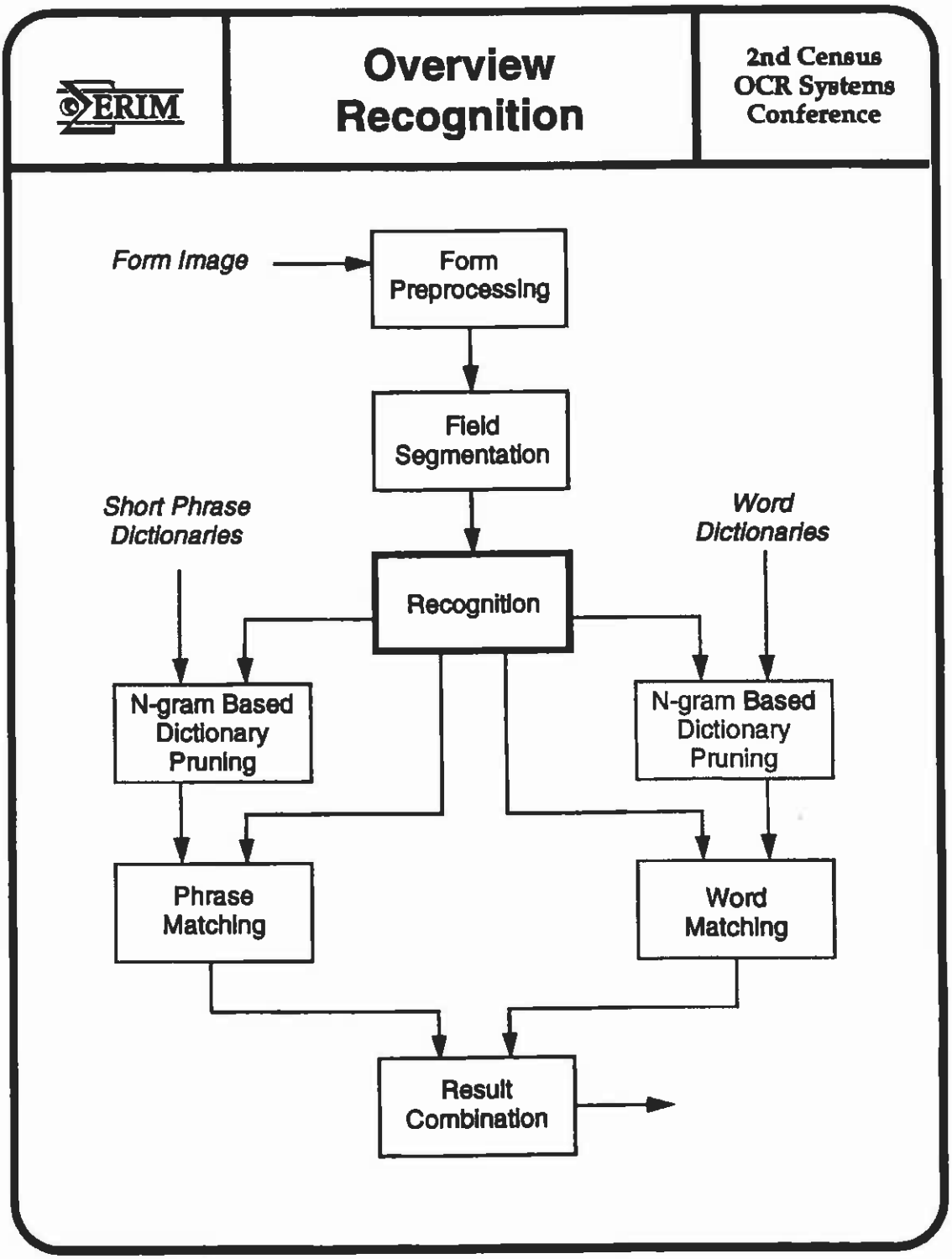
Unions

2nd Census OCR Systems Conference

Union Size
(Number of Primitive Segments)

Primitive Segments




ERIM
Overview
Recognition
2nd Census
OCR Systems
Conference
Form Image
Form
Preprocessing
Field
Segmentation
Short Phrase
Dictionaries
Word
Dictionaries
Recognition
N-gram Based
Dictionary
Pruning
N-gram Based
Dictionary
Pruning
Phrase
Matching
Word
Matching
Result
Combination

ERIM

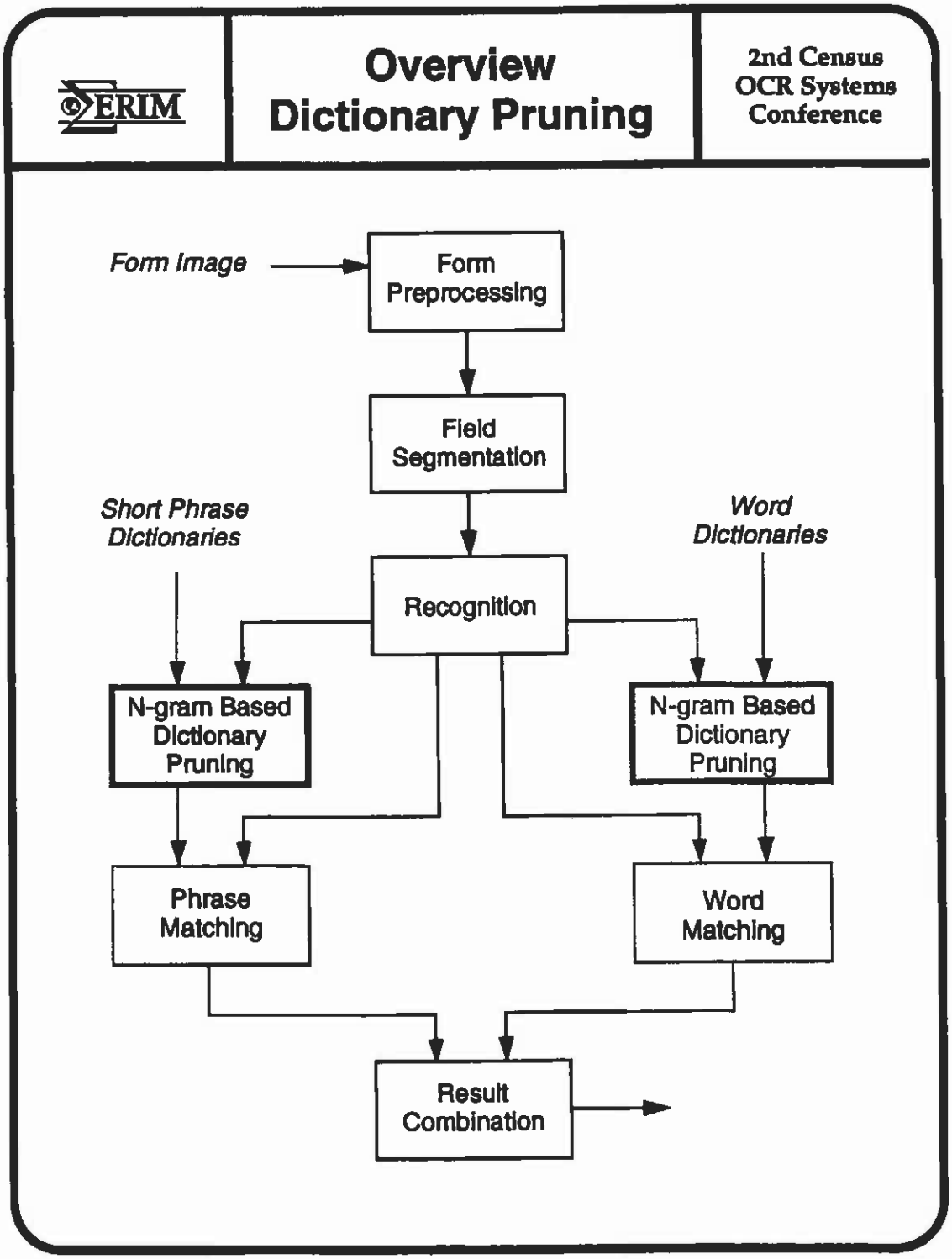
Character Recognition

2nd Census OCR Systems Conference

- **Neural networks**
  - multi-layer perceptron with backpropagation training
  - uppercase, lowercase, digits
  - 2 feature sets
- **Network training (~2.9M presentations)**
  - uppercase transition 83%
  - lowercase transition 78%
  - upper-case bar 86%
  - lower-case bar 80%

Training Data

	upper case	lower case	digits
Census Forms	21,076	18,203	0
USPS	13,414	18,731	~8,000
Total	34,490	36,934	~8,000


ERIM
Overview
Dictionary Pruning
2nd Census
OCR Systems
Conference
Form Image
Form
Preprocessing
Field
Segmentation
Short Phrase
Dictionaries
Word
Dictionaries
Recognition
N-gram Based
Dictionary
Pruning
N-gram Based
Dictionary
Pruning
Phrase
Matching
Word
Matching
Result
Combination

ERIM

Phrase Dictionary Pruning

2nd Census OCR Systems Conference

- **Extract n-grams from recognition results**
- **Use short phrase dictionary**
  - ~8,000 phrases (depending on field)
  - coverage of paper training fields ~62%
- **Prune phrase dictionary**
  - ~1,000 phrases
  - correct phrase removed small fraction of the time

ERIM

Phrase Lexicons Coverage Analysis

2nd Census OCR Systems Conference

		field 1	field 2	field 3	overall
short (ftp site)	size	8,173	8,475	7,788	N/A
	coverage (%)	62	70	55	62
short (CD-ROM)	size	8,340	8,633	7,956	N/A
	coverage (%)	63	72	55	63
long (CD-ROM)	size	50,224	49,853	66,318	N/A
	coverage (%)	74	80	68	74

ERIM

Word Lexicons Coverage Analysis

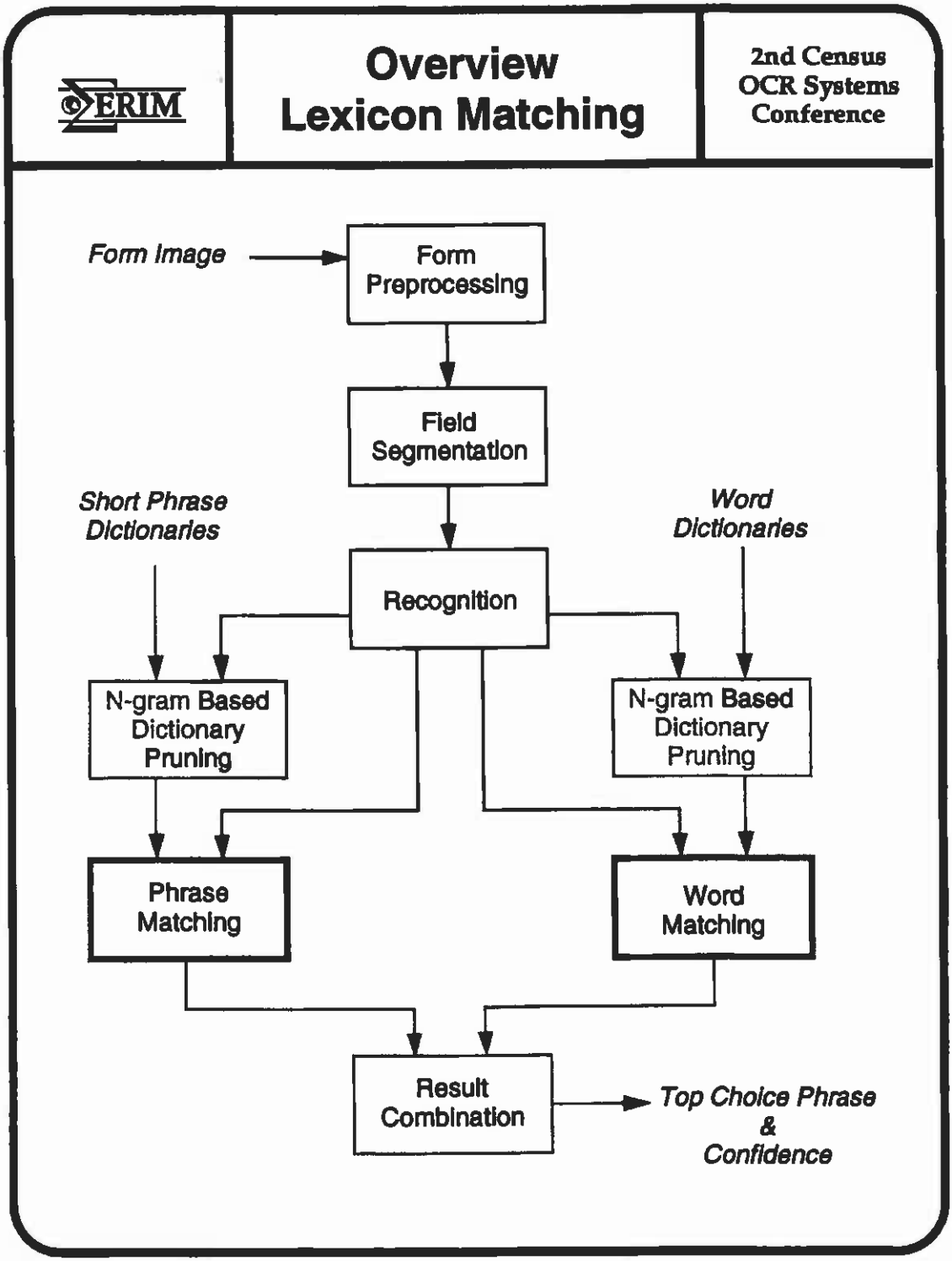
2nd Census OCR Systems Conference

Short Word Dictionaries (from FTP site)

	field 1	field 2	field 3	overall
lexicon size	4,744	4,777	5,902	N/A
1 word phrases	30	37	27	31
2 word phrases	41	41	41	41
3 word phrases	13	9	14	12
1 U 2	71	79	67	72
1 U 2 U 3	84	88	82	85
word *	87	90	87	88

Long Word Dictionaries (from CD-ROM)

	field 1	field 2	field 3	overall
lexicon size	14,545	14,503	17,239	N/A
1 word phrases	31	39	27	32
2 word phrases	45	45	44	45
3 word phrases	15	11	17	14
1 U 2	76	83	72	77
1 U 2 U 3	92	94	89	92
word *	96	97	96	96


ERIM
Overview
Lexicon Matching
2nd Census
OCR Systems
Conference
Form Image
Form
Preprocessing
Field
Segmentation
Short Phrase
Dictionaries
Word
Dictionaries
Recognition
N-gram Based
Dictionary
Pruning
N-gram Based
Dictionary
Pruning
Phrase
Matching
Word
Matching
Result
Combination
Top Choice Phrase
&
Confidence

ERIM

Lexicon Matching (for words and phrases)

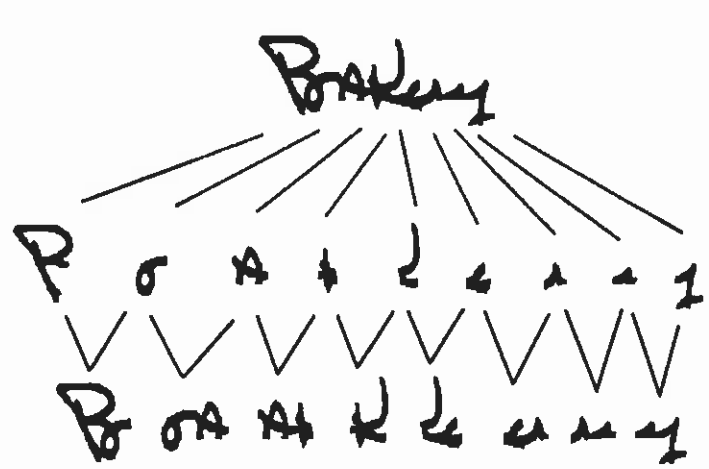
2nd Census OCR Systems Conference

- Inputs: recognition results for *unions* + lexicon
- Use Viterbi match (dynamic programming) to find best match for each lexicon word
- Confidence is sum of recognition results of characters used in match

ERIM

Word and Phrase Matching

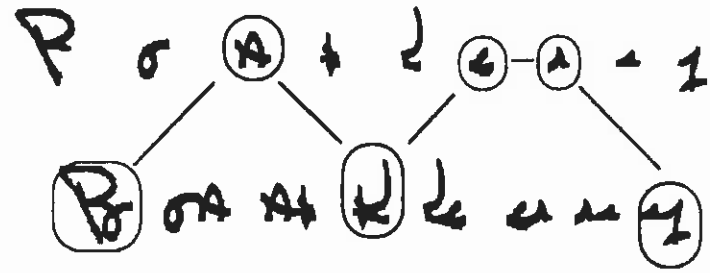
2nd Census OCR Systems Conference



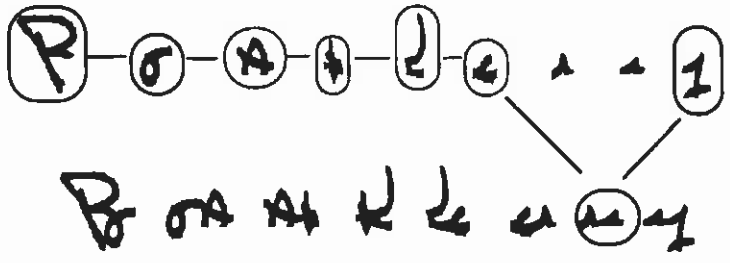
Original Image

Primitive Segments

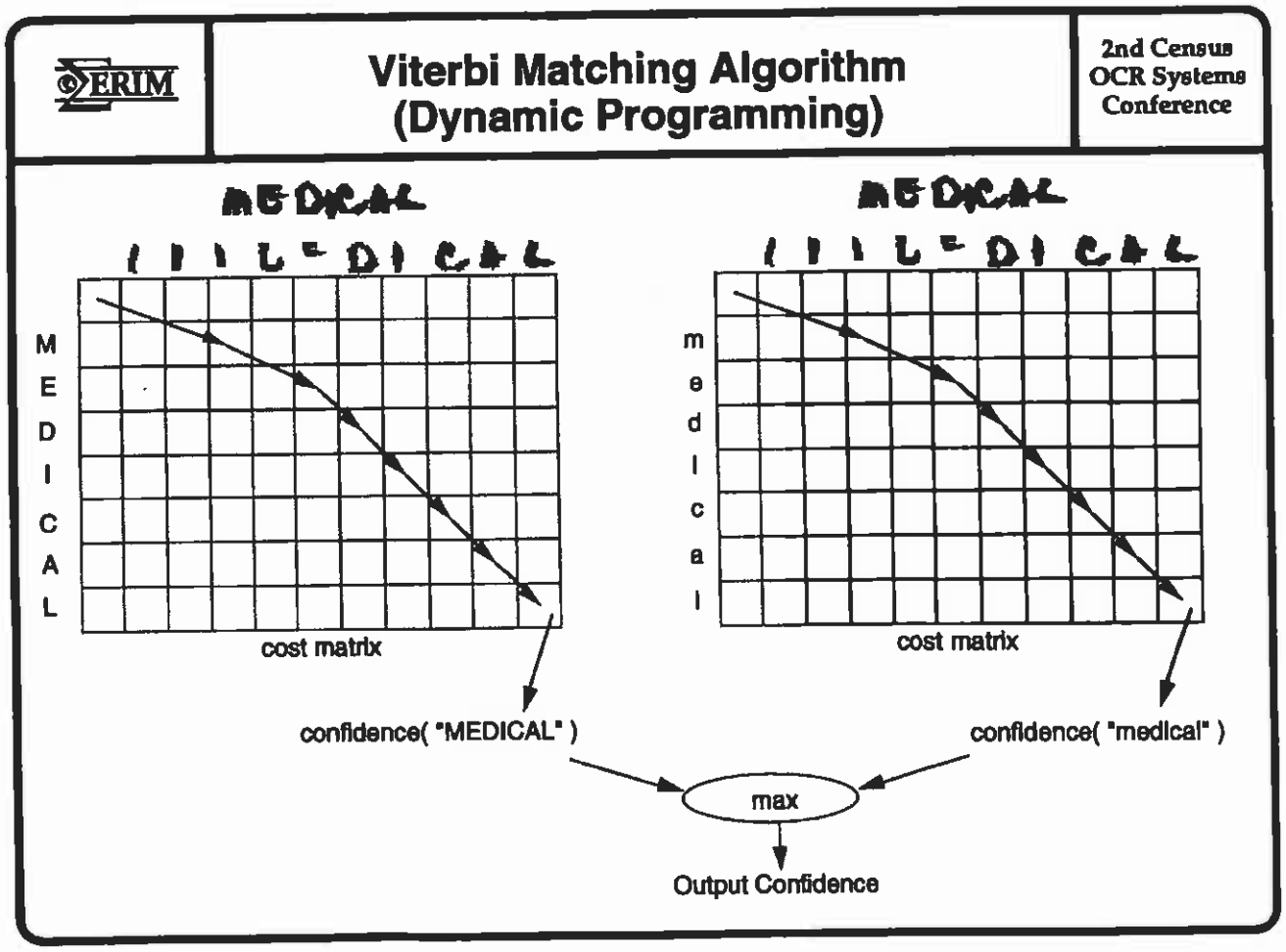
Unions of two Primitive Segments

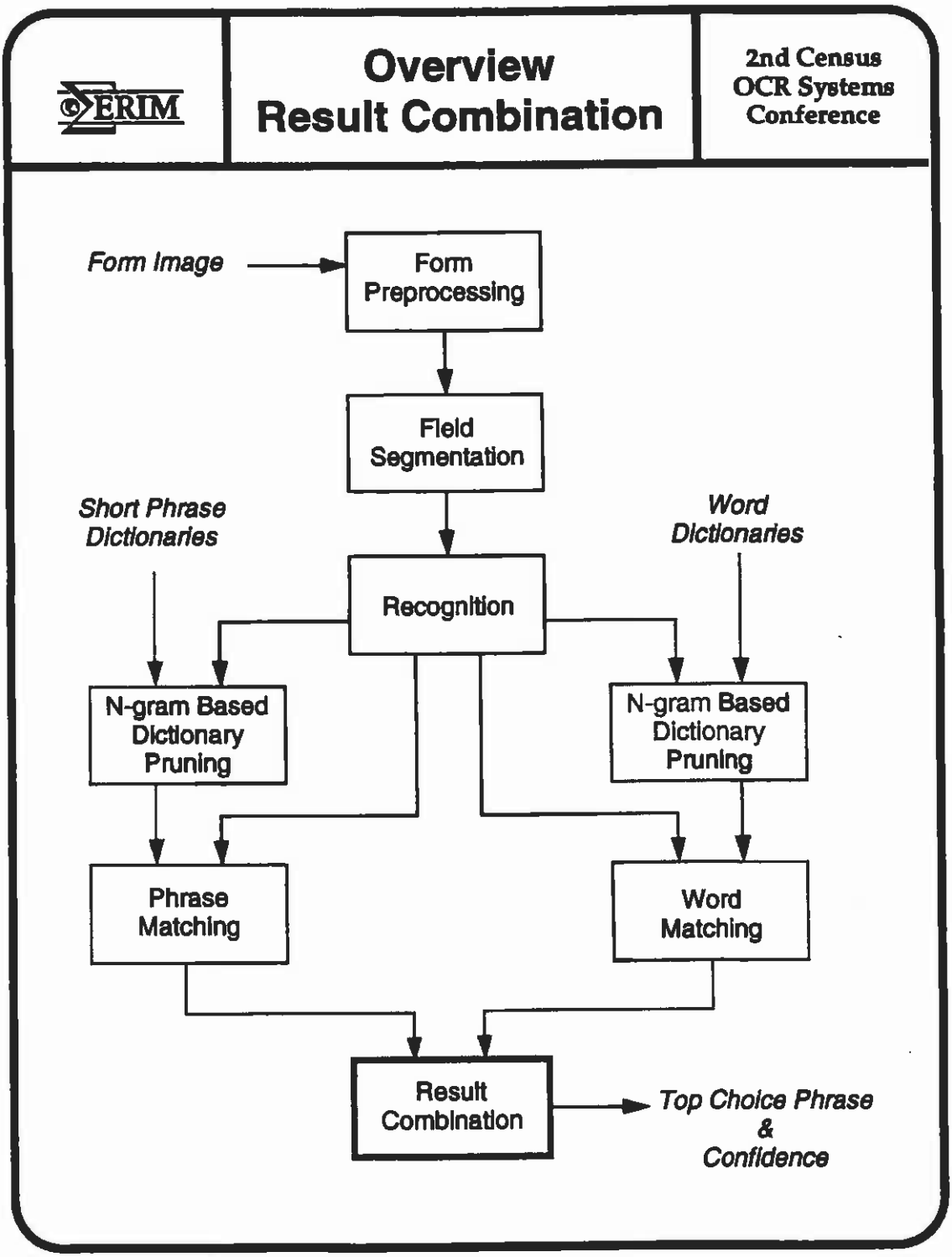


Best Match to Word "BAKERY"



Best Match to Word "POACHING"


ERIM
Viterbi Matching Algorithm
(Dynamic Programming)
2nd Census
OCR Systems
Conference
MEDICAL
MEDICAL
M
E
D
I
C
A
L
cost matrix
m
e
d
i
c
a
l
cost matrix
confidence("MEDICAL")
confidence("medical")
max
Output Confidence


ERIM
Overview
Result Combination
2nd Census
OCR Systems
Conference
Form Image
Form
Preprocessing
Field
Segmentation
Short Phrase
Dictionaries
Word
Dictionaries
Recognition
N-gram Based
Dictionary
Pruning
N-gram Based
Dictionary
Pruning
Phrase
Matching
Word
Matching
Result
Combination
Top Choice Phrase
&
Confidence

ERIM

Result Combination

2nd Census OCR Systems Conference

- **Word match scores decremented to compare with phrase match scores**
  - Constant β chosen heuristically
- **Divide results into three non-overlapping groups**
  - 1) Two-word top choice phrase = two-word top choice word match
  - 2) Three-word top choice phrase = three-word top choice word match
  - 3) All other cases
- **Reassign confidence values so that all of group 1 have higher confidence than group 2 which in turn have higher confidence than group 3**
- **Within groups, old confidence value determines order**

ERIM

Processing Times

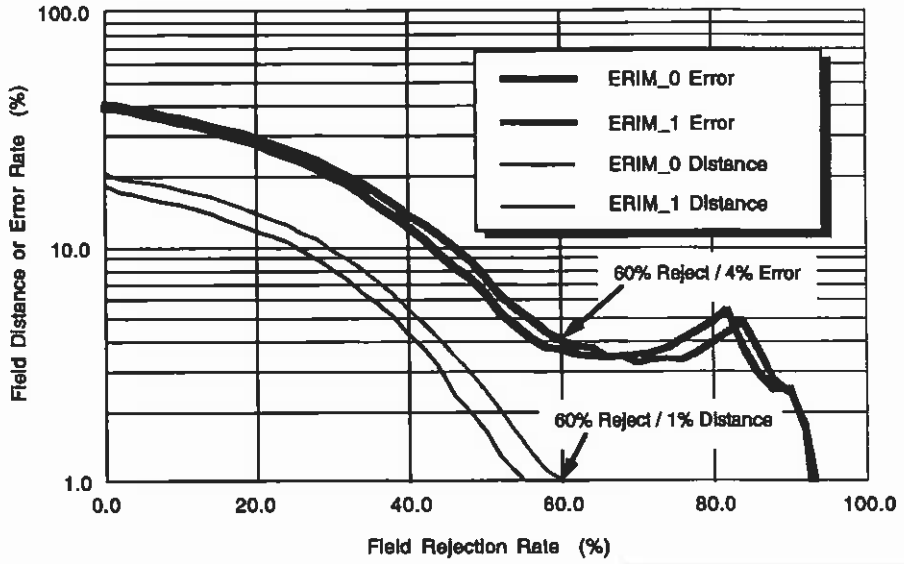
2nd Census OCR Systems Conference

	Preprocessing	Recognition ERIM_0	Recognition ERIM_1	Total ERIM_0	Total ERIM_1
Paper	4.6	27.1	17.9	31.6	22.5
Microfilm	26.6	20.5	16.4	47.1	43.0

All times in seconds per field on a SUN Sparcstation2

ERIM System Results (Paper)

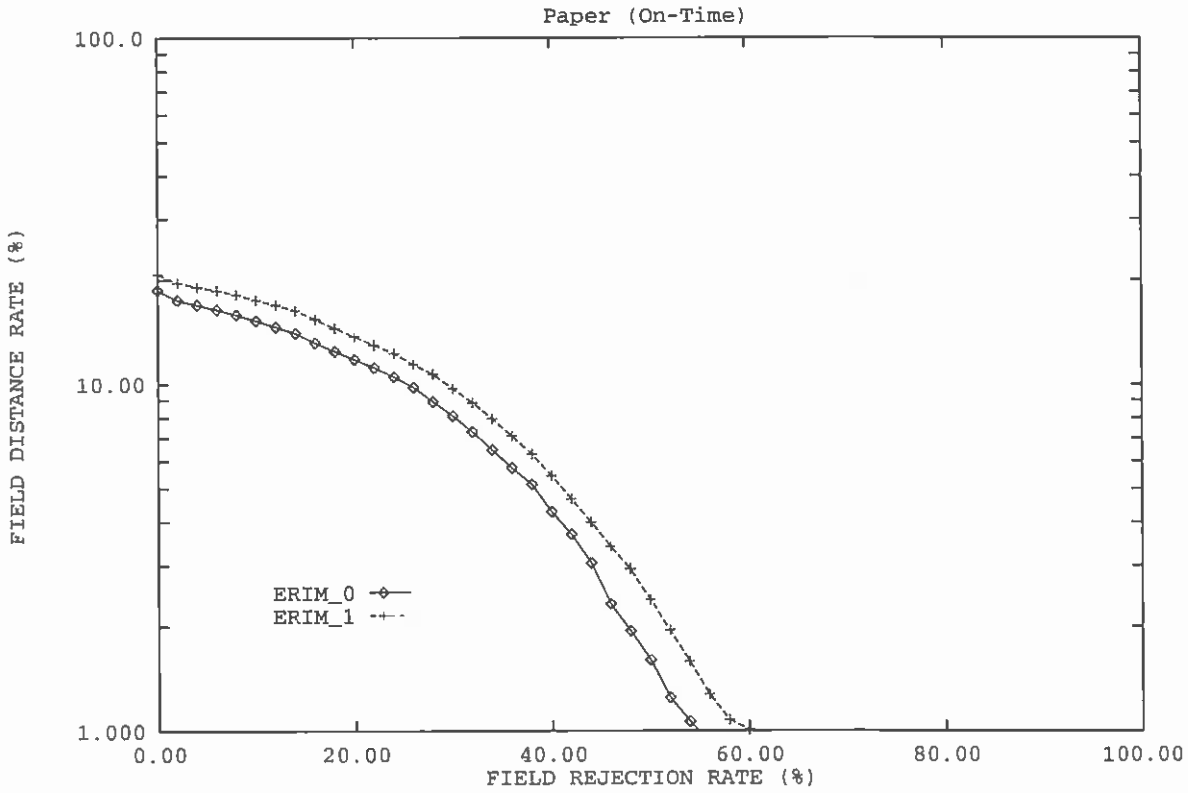
2nd Census OCR Systems Conference

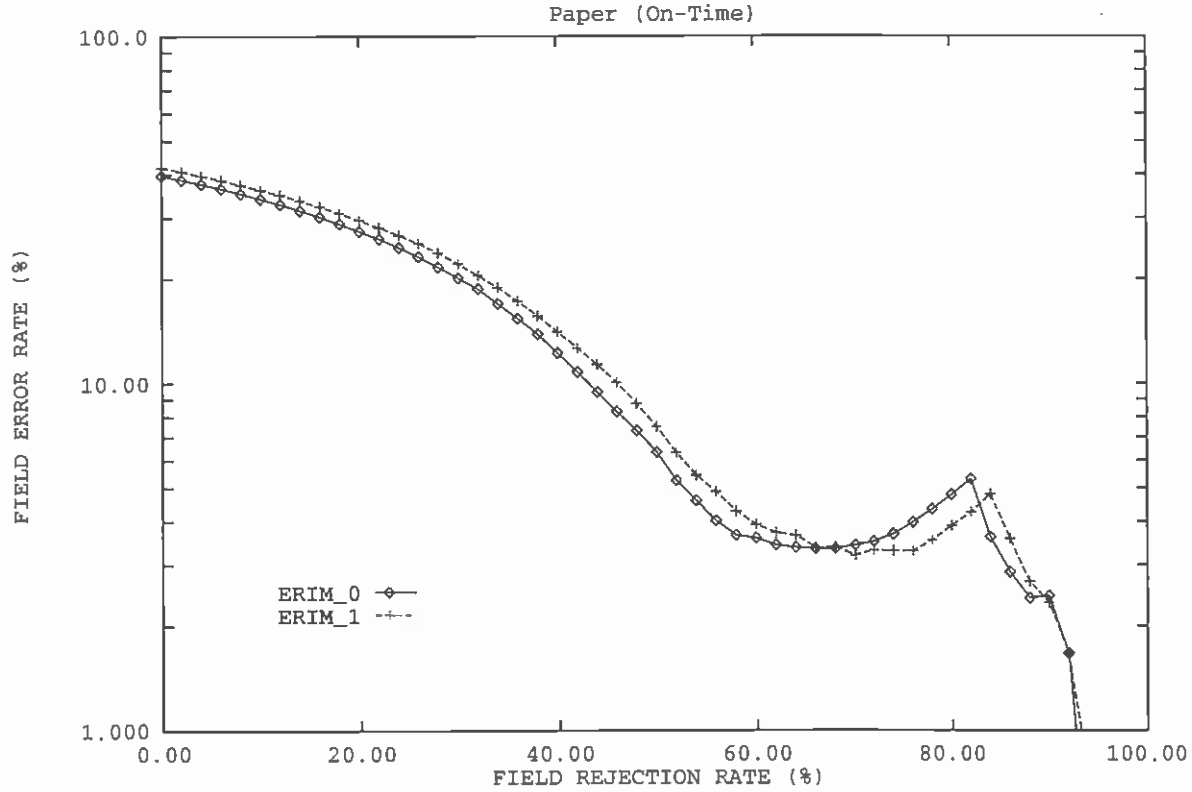


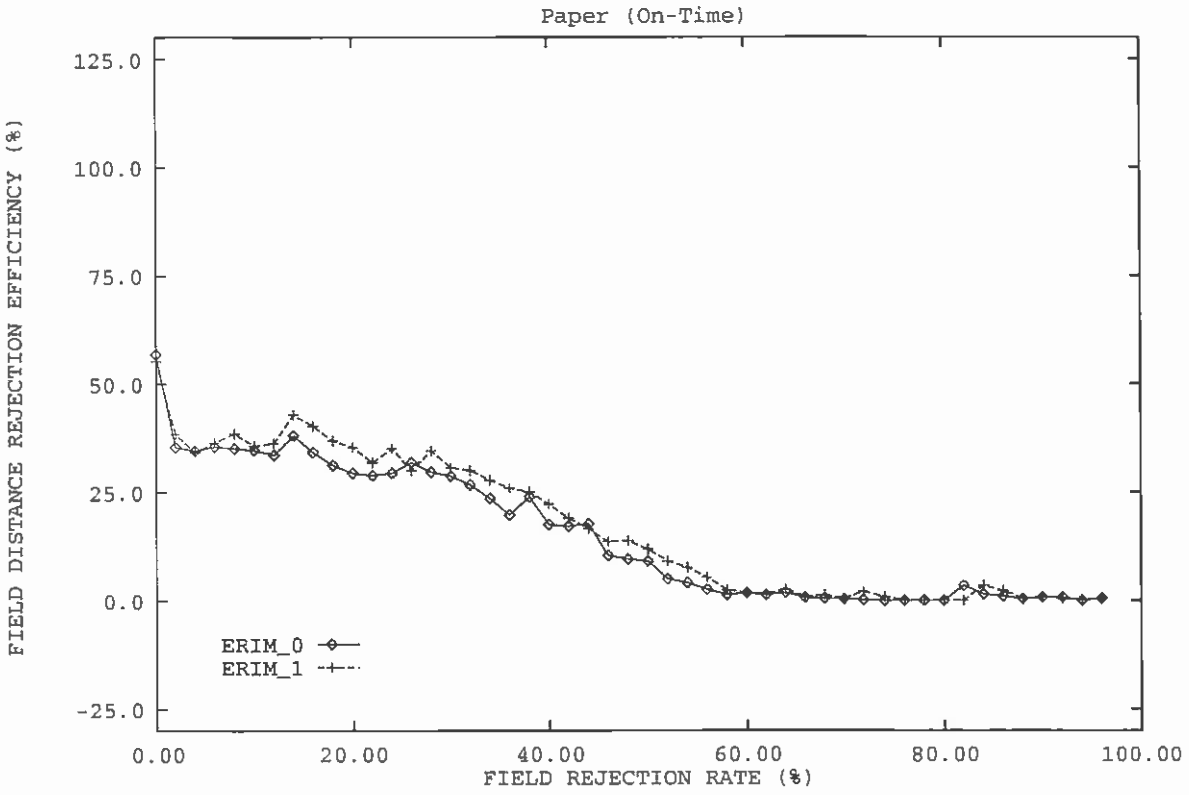
Conclusions

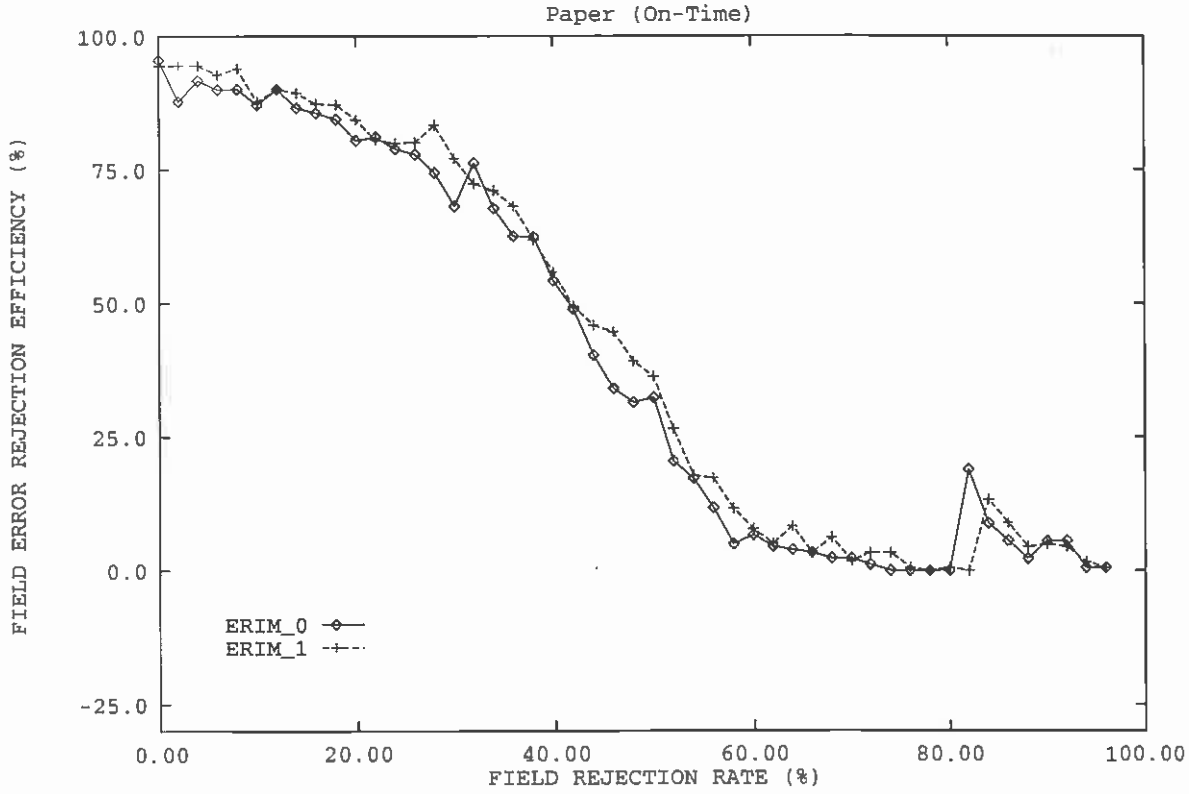
2nd Census OCR Systems Conference

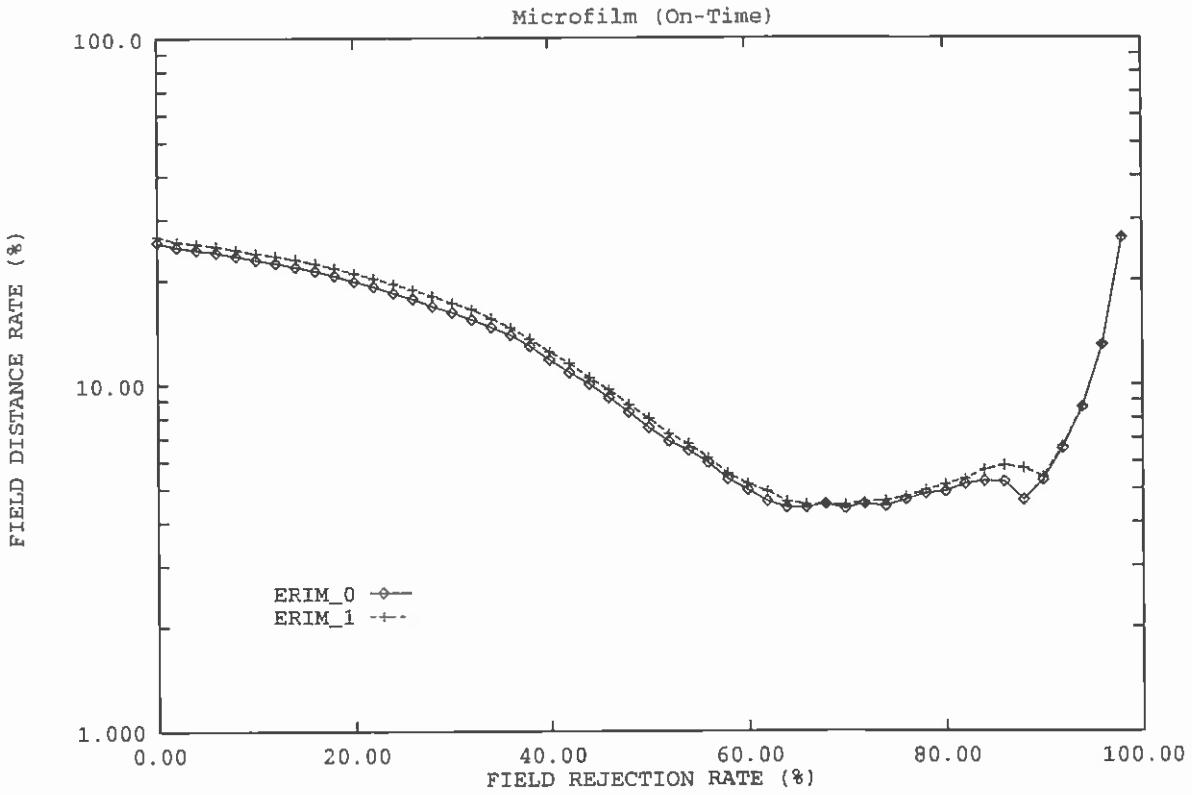
- The technology for OCR assisted census forms processing is available today
- Solution can be made using off-the-shelf hardware
- Immediate cost savings are substantial
- Future cost savings will become available as technology improves

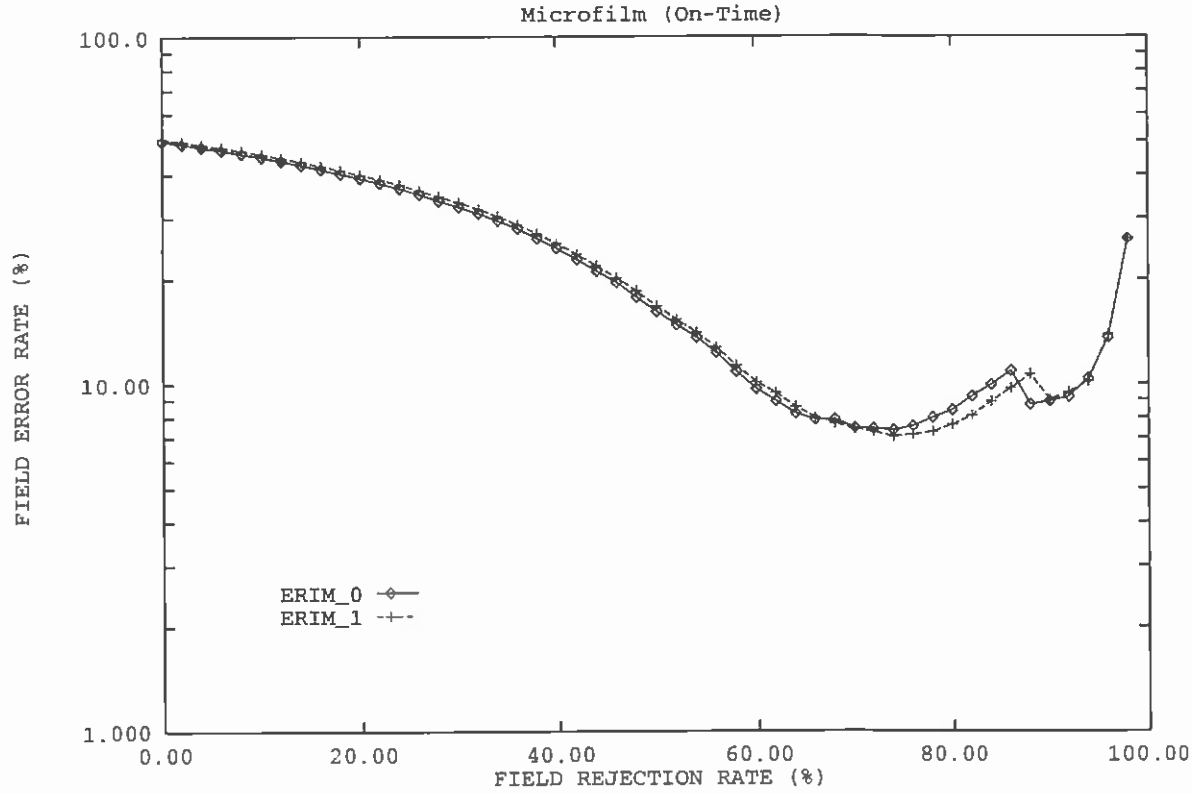

Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

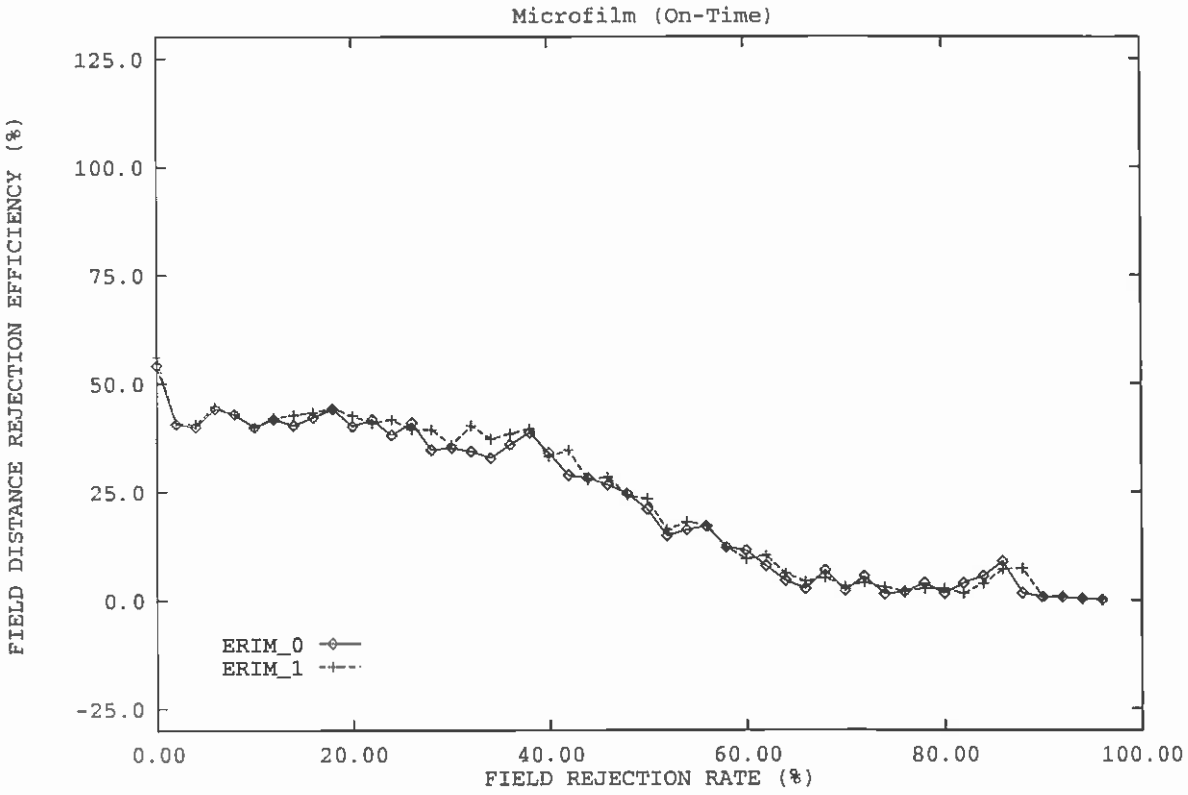

Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

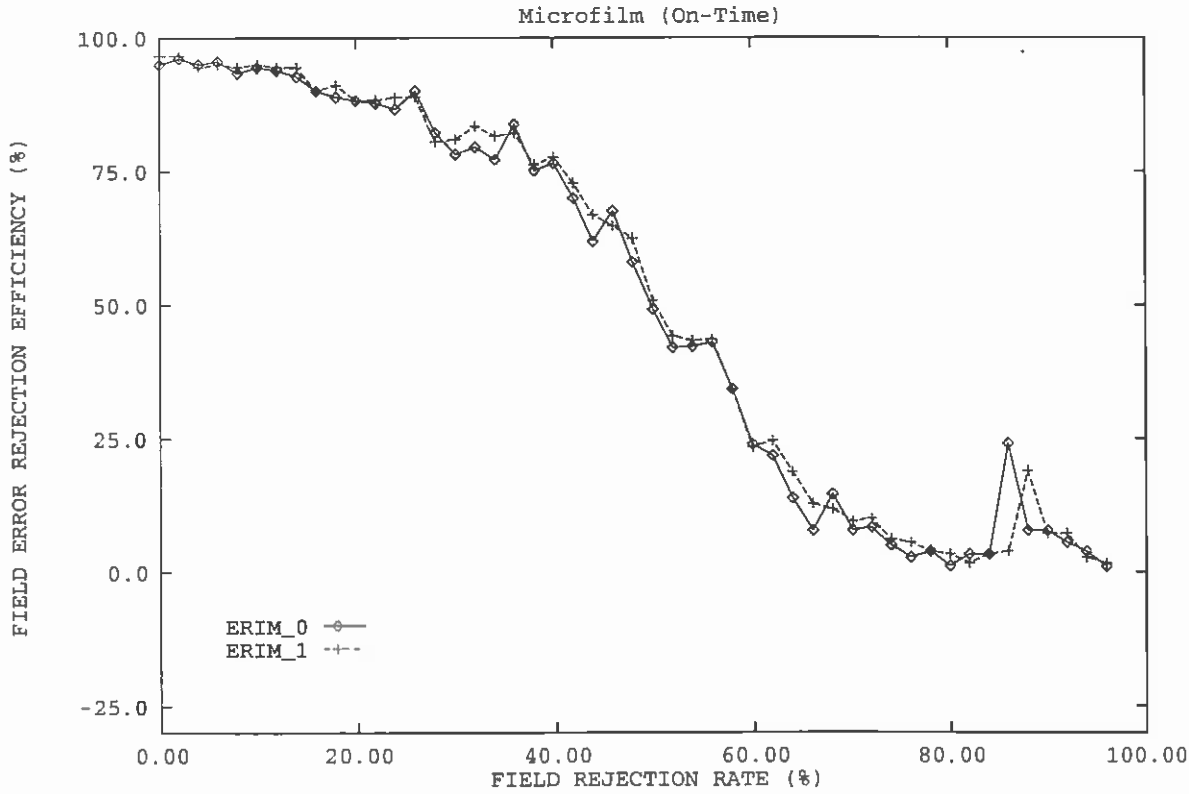

Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
ERIM_0
ERIM_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

NIST'S SECOND CENSUS OCR SYSTEMS CONFERENCE

15-16 FEBRUARY 1994

HUGHES

HUGHES PLAYERS

HUGHES

CONTEXTUAL PROCESSING AND SYSTEM INTEGRATION

TONY BARAGHIMIAN - HUGHES INFORMATION TECHNOLOGY CO.

(703) 759-1392

gab@mitchell.hitc.com

PREPROCESSING AND RECOGNITION

SUSAN CHURCH - HUGHES RECOGNITION SYSTEMS

(818) 702-1455

schurch@lpld01.hac.com

AGENDA

HUGHES

- DESCRIPTION OF HUGHES' SUBMISSIONS
- RESULTS
- COMMENTS

HUGHES SUBMITTED FOUR SETS OF RESULTS

HUGHES

- HUGHES FOCUSSED ON PAPER-BASED IMAGES
- HUGHES TUNED FOR TOTAL FIELD CORRECTNESS (I.E., FIELD ERROR RATE)

HUGHES_0, HUGHES_1 > DIFFERENT RECOGNITION SYSTEMS

HUGHES_2 — FUSION OF HUGHES_0 AND HUGHES_1

HUGHES_9 — RAW OCR READS FROM HUGHES_1

CHARACTERISTICS OF HUGHES_0

HUGHES

- FIELD REGISTRATION BASED ON DOTTED BOX FINDING
- TEXT LINE(S) DETECTION AND DESKEW
- CHARACTER (OVER-) SEGMENTATION USING LOCAL CONTOUR MINIMA AND MAXIMA
- CHARACTER NORMALIZATION
- STROKE FEATURE EXTRACTION
- THREE MLP/BACKPROP CLASSIFIERS: UPPERCASE, LOWERCASE, NUMERIC
  - TRAINED ON 7800/7800/10,000 HANDPRINT SAMPLES FROM NIST SPECIAL DATABASE 3
- DICTIONARY CORRECTION USED LONG WORDS AND SHORT PHRASES SUPPLIED BY NIST CDROM

CHARACTERISTICS OF HUGHES_1

HUGHES

- SAME AS HUGHES_0 EXCEPT:
  - CAVITY FEATURE EXTRACTION

CHARACTERISTICS OF HUGHES_2

HUGHES

- COMBINATION OF HUGHES_0 AND HUGHES_1 USING A SET OF FUSION RULES, TAKING ADVANTAGE OF INFORMATION SUCH AS:
  - MULTIPLE CANDIDATE TEXT STRINGS WITH ASSOCIATED CONFIDENCES
  - INDEPENDENT UPPERCASE/LOWERCASE BELIEFS
  - INTERRELATIONAL STRENGTHS/WEAKNESSES BETWEEN RECOGNITION PARADIGMS

RESULTS

HUGHES

- LESSONS LEARNED FROM THIS CONFERENCE:
  - BETTER RESULTS ARE OBTAINED BY TRAINING THE CLASSIFIERS USING VERY LARGE DATA SETS, PREFERABLY THE NIST-12 DATA
  - MULTIPLE SEGMENTATION HYPOTHESES SEEM TO GIVE BETTER RESULTS, POSSIBLY AT A THE EXPENSE OF ADDITIONAL PROCESSING TIME
  - DICTIONARY AUGMENTATION, SUCH AS SPELL CHECKING AND PROVIDING 100% COVERAGE, BOOST PERFORMANCE

COMMENT 1: PARTICIPATION

HUGHES

- CONTINUED PARTICIPATION IN THIS EVENT IS A STRONG INDICATION OF AN ORGANIZATION'S COMMITMENT TO SUCCEEDING IN THE DOCUMENT PROCESSING BUSINESS MARKET
- COMPLETE PARTICIPATION DEMONSTRATES AN ORGANIZATION'S ABILITY TO:
  - QUICKLY ADAPT TO CUSTOMER REQUIREMENTS
  - PROPERLY FUND AND STAFF UNPLANNED EFFORTS

COMMENT 2: FUTURE NIST OCR TESTS

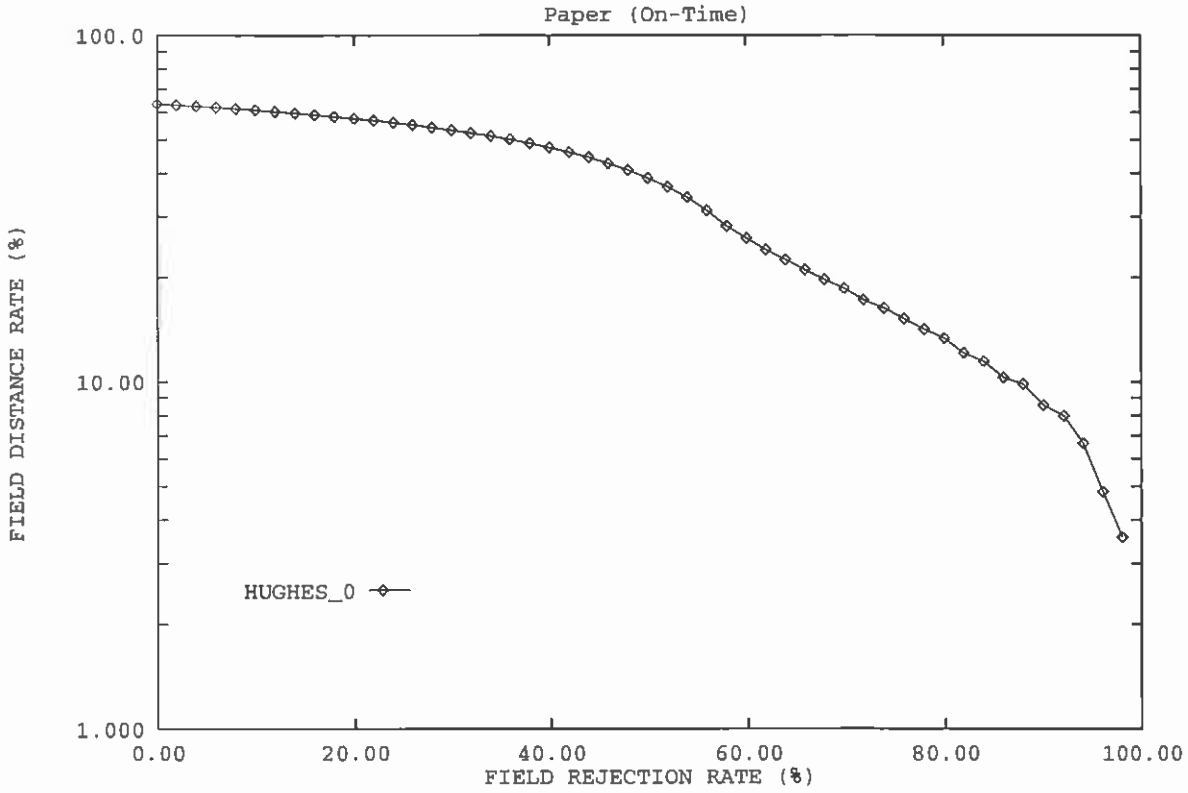
HUGHES

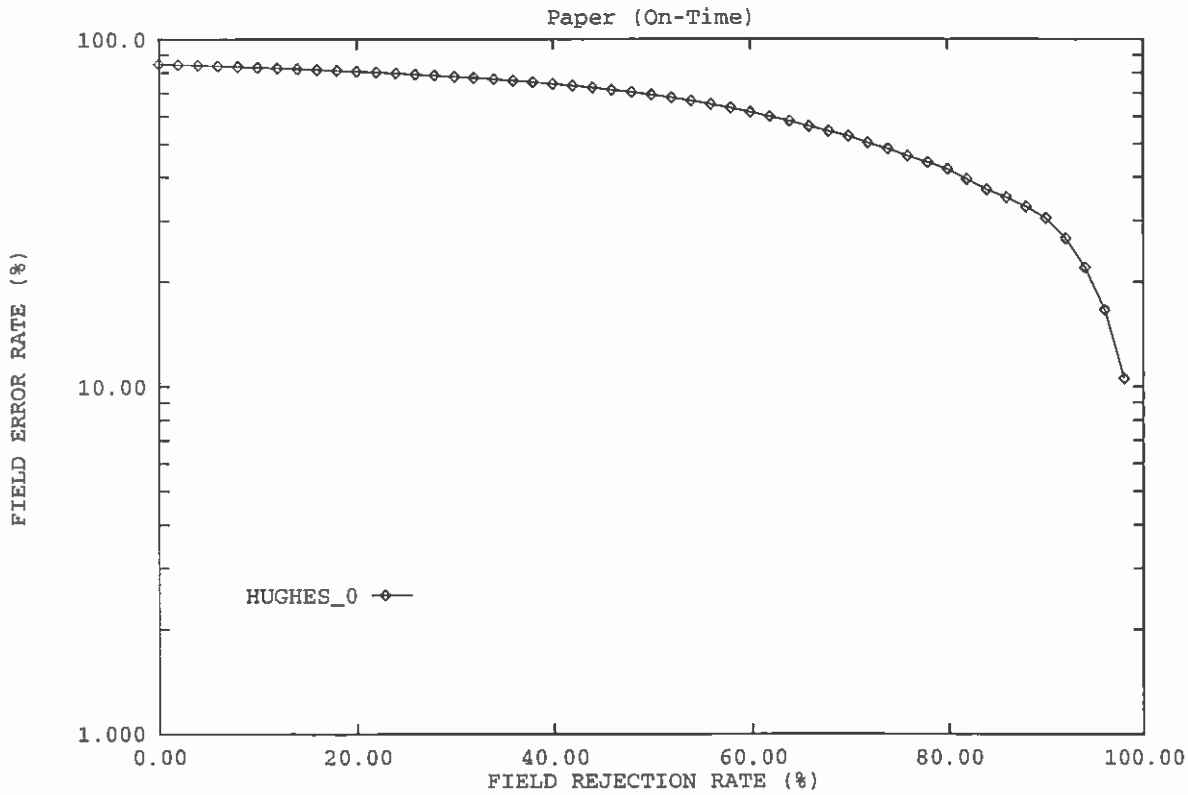
- HUGHES LOOKS FORWARD TO FURTHER PARICIPATION IN THIS SERIES OF OCR TESTS
- SUGGESTIONS FOR FUTURE TESTS INCLUDE:
  - SPEED/THRUPUT TESTING
  - CONTINUED ACCURACY TESTING
  - LARGER SET OF ZONES; ESPECIALLY CONTAINING INTER-ZONE CONTEXT

COMMENT 3: FORM REDESIGN

HUGHES

- HUGHES DOESN'T TELL CUSTOMERS TO REDESIGN THEIR FORMS
- PATRON MENTALITY: "I'VE NEVER MET A FORM I COULDN'T FIX"


Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
HUGHES_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
HUGHES_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

IBM Entries

IBM 1 - Unconstrained Handprint OCR (HOCR)

IBM 2 - Experimental Cursive OCR (COCR)

IBM 0 - Combination of IBM 1 and IBM 2

IBM 3 - IBM 1 with improved dictionary correction (late entry)

IBM 9 - IBM 1 without dictionary lookup (raw OCR)

Comments on the IBM Entries

The IBM entries were submitted from the IBM Almaden Research Center in San Jose, CA. Although we had 4 on-time entries and 1 late entry, there were essentially only 2 basic systems: a handprint OCR (HOCR) system, and an experimental Cursive OCR (COCR) system. IBM 1 was based on HOCR, IBM 2 on COCR, IBM 0 a combination of IBM 1 and IBM 2, and IBM 9 the raw OCR results of IBM 1 (prior to dictionary based correction). IBM 9 was submitted for dictionary correction using software developed by NIST. IBM 3, which was a late entry, is the same as IBM 1 with improved contextual processing.

HOCR is designed to handle unconstrained handprinted characters. It separates touching characters by oversegmenting such patterns and selecting the best segmentation points by dynamic programming. This is a recognition driven segmentation as recognition confidences are made use of in choosing the best path. Due to prior commitments and time pressures, we could not train HOCR using the training data provided for the 2nd Census OCR testing. In fact, it had been trained only using discrete samples collected from the training data (SD 3) and test data (SD 7) of the 1st Census OCR testing, and additional samples collected from a pen based computer. After the 2nd OCR conference, we have trained HOCR using training data provided for this test, and significantly improved our recognition results. These results are being submitted to NIST for evaluation.

COCR is a highly experimental system for recognizing cursive writing. It is in an early stage of development, as demonstrated by the experimental results. It is based on Hidden Markov Models (HMM), and does not use any explicit segmentation. A sliding window is passed over the text field, and features of the image segment in the window are extracted. These features are then used to cluster and classify the patterns. The greatest drawback of the system was that it was trained only on discrete characters collected from SD 3 and SD 7. It is now being trained using data from the training set provided for the 2nd OCR testing.

Some additional details on the 2 systems as well as contextual processing are provided in the accompanying viewgraphs.

Field Image Extraction

Sequence of Operations

- Smoothing and Filtering
- Registration
  - Reference Marks
  - Key Lines
  - Box Boundaries
- Form Type
- Deskewing
- Data Tracking and Extraction
  - Dotted Lines
  - Clipping at Form Text
- Output to Image

Handprint OCR (HOCR)

- Character Segmentation
  - Recognition based segmentation
  - Oversegmentation
  - Select best segmentation by dynamic programming
- Feature Set
  - Contour Direction Features (88-dimensional)
    - Histograms of directions of contour pixels
  - Bending-point Features (96-dimensional)
- Classifier
  - Neural-net classifier to narrow choice of hypotheses to 3
    - Feed-forward BP network
  - Template matching classifier to reorder the 3 choices based on distance
- Training Set: Discrete Characters only
  - SD 3: 42,000 each for Upper Case and Lower Case
  - SD 7: 12, 000 each for UC and LC
  - Pen Input Database: 29,000 each for UC and LC

Cursive OCR

- Basis of IBM 2
- Early prototype - highly experimental stage
- Based on Hidden Markov Model (HMM)
- No explicit segmentation
  - Sliding overlapped window of fixed size
- Features
  - Contour direction features (88-dimensional)
  - 27-dimensional after data reduction to 90% of energy
- Training set
  - Trained on discrete characters only
  - Approx. 42,000 samples from SD 3, and 12,000 samples from SD 7 each for Upper Case and Lower Case
- Language Model
  - Unigrams modified from NIST-supplied ascii files, separate dictionaries for each field

Language modeling

Two databases, with 32,000 matched tiff+ascii examples for each field (approx. 250k words). Only ascii available for another 880k words now, up to 20m words later?

	field0	field1	field2
# of field tokens	46901	47005	61399
# of word types	15179	15573	18050
# wty in rep. fields	8209	8502	7822
words per field	2.618	2.527	2.876
w/f in rep. fields	2.037	1.993	2.077

	field0	field1	field2
# of field tokens	53005	52484	68939
# of word types	14545	14503	17239
# wty in rep. fields	9310	9575	8757
words per field	2.590	2.495	2.841
w/f in rep. fields	2.018	1.967	2.054

Actual dictionary sizes

	field0	field1	field2
# of rep. words	6290	6263	7935
# of words	7449	7082	9588
w/f in rep. fields	2.01805	1.9671	2.05379

(False hapaxes were added)

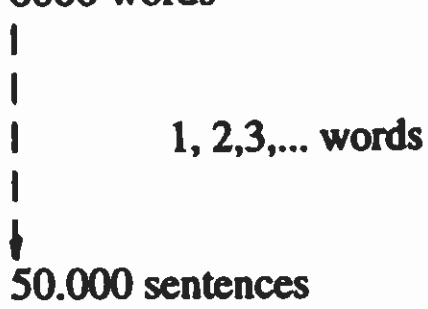
Using dictionaries to improve recognition

Problem:

Assuming 60% character recognition rate,
the probability of recognizing a field correctly
is extremely low:

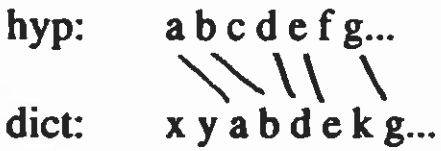
60 60 60 60 60 60 60.
36 22 13 7 4 2.... %

Use of dictionaries (context) is a MUST.



The method

1. Try to match word



2. Since letter separation is a problem, try all alternatives returned by OCR.

3. How to score "similarity"
   - match character
   - match on 2d choice
   - match with skip, 2 skips
   - increase score for "sure" chars
   - increase score for successive matching chars
   - increase score for frequent words

4. Keep reasonable alternatives

5. Consider all combinations for first three words, and look up dictionary
   - use 1st word as index if high confidence,
   - else, scan full dictionary

```
phrase - template: sEcRETARy

enter  SECRETARY   EGEFZRHEV
enter  SECRETIARY  EGEFZZOHEV
enter  SECREFRARY  EGEFZSTHEV

best set 17    RECREATION 43
best set 14    SCREENER 6
best set 14    CREATIVE 7
best set 25    SECONDARY 24
best set 32    SECRETARIAL 66
best set 49    SECRETARY 256
best set 17    SECURITY 155
best set 15    DECORATOR 21
best set 17    CARETAKER 17
best set 30    SECRATARY 6

SECRETARY (49)

best set 71     SECRETART
best set 71     SECRATARY
best set 70     SECRTARY
best set 70     SECRETRY
best set 71     SECRETARTY
best set 74     SECRETARY
best set 71     SECREETARY
best set 71     SECRETARYU
best set 70     SECRETAY
best set 71     SECRETERY

solution: SECRETARY
```

SECRETARY

score (word)

(phrase)

phrase - template: NImVfmiuanG

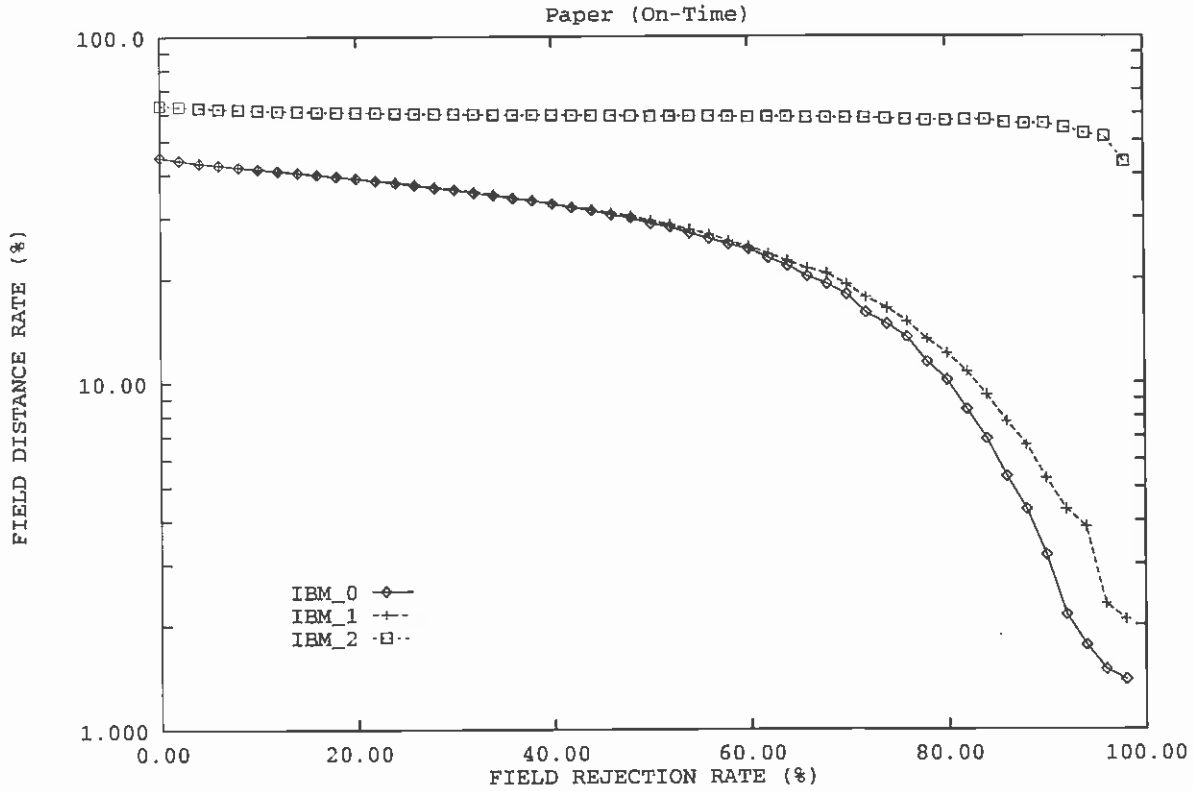
Manufacturing

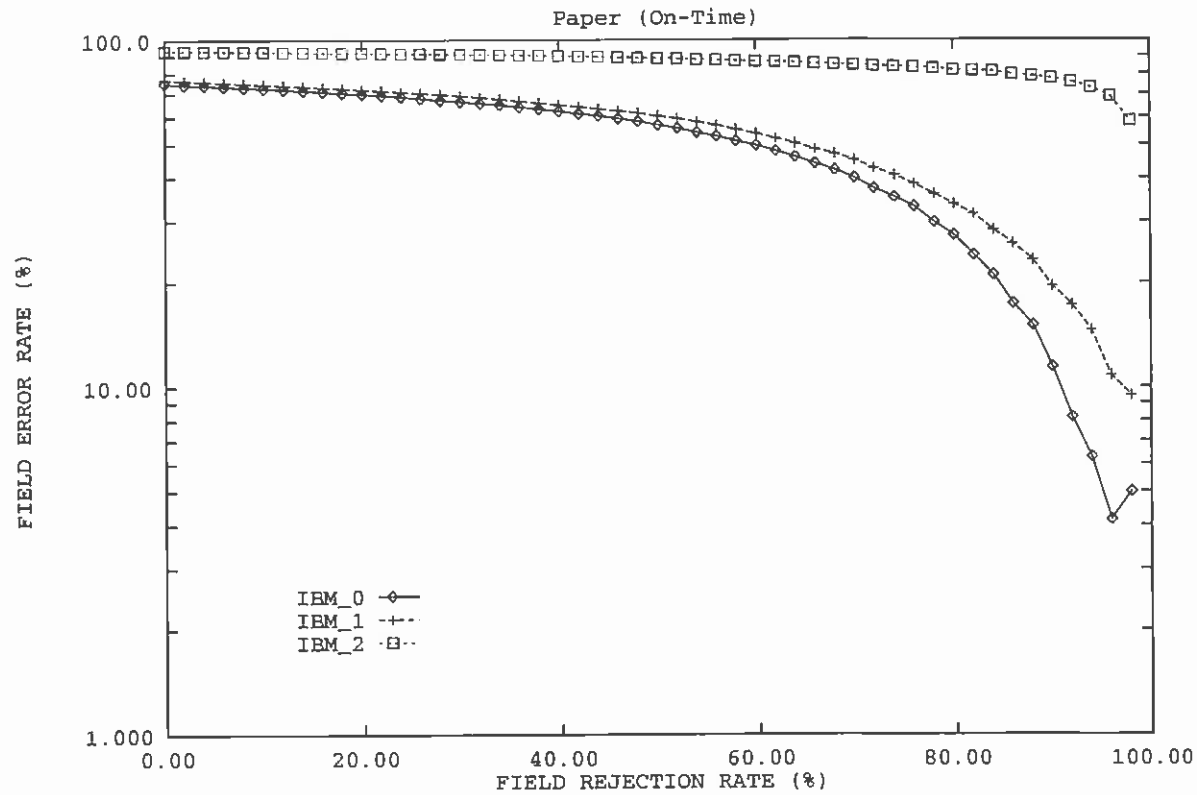
enter NIMVFMIUANG XLWUINJVMXS
enter NIMVMIUANG XLWUAJVMXS
enter NIMVIMIUANG XLWUJTJVMXS
enter KMVFMIUANG MWUINJVMXS
enter KMVMIUANG MWUAJVMXS
enter KMVIMIUANG MWUJTJVMXS
enter IKMVFMIUANG LUWUINJVMXS
enter IKMVMIUANG LUWUAJVMXS
enter IKMVIMIUANG LUWUJTJVMXS

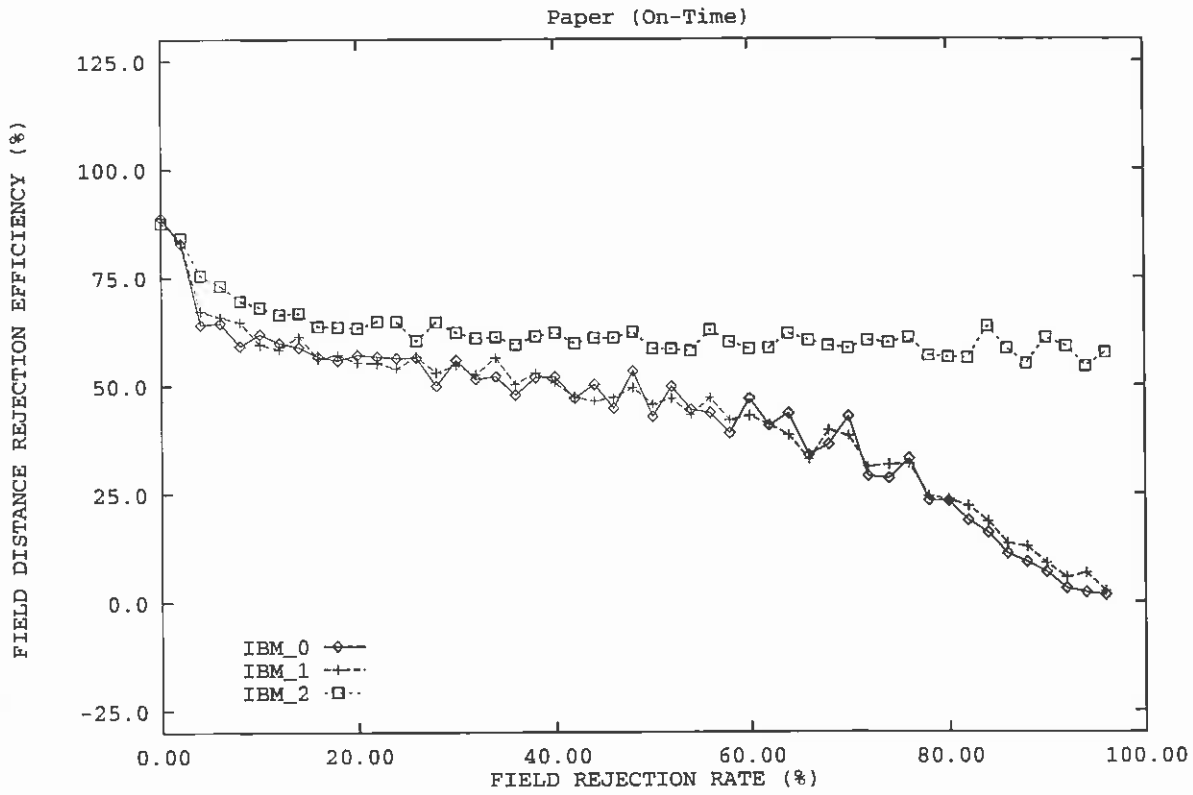
best set 20 MAINTAINING 11
best set 17 MANUFATURING 13
best set 13 MONITORING 5
best set 17 SMELTING 8
best set 14 MANUFACTRING 7
best set 15 MACHINING 16
best set 12 CONTINUING 5
best set 16 MANUFACTUING 6
best set 16 MANUFACURING 5
best set 21 MANUFACTING 31

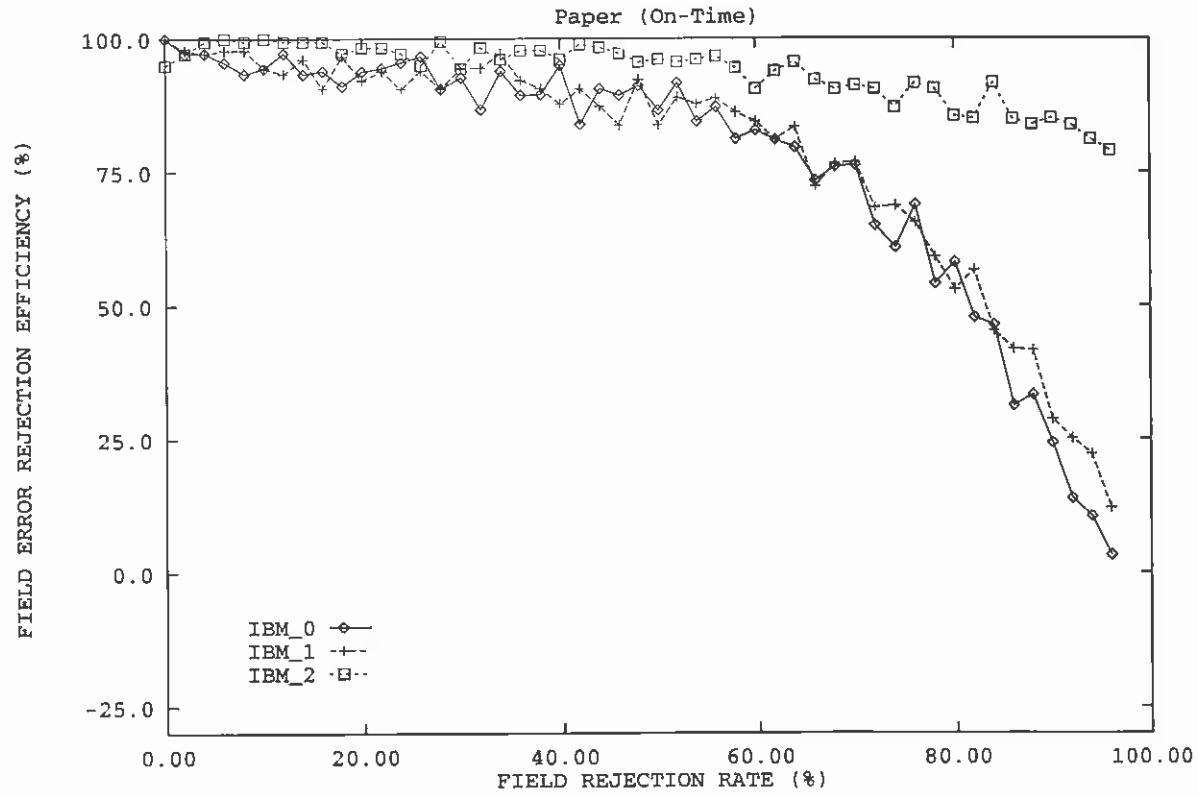
best set 42 MANUFACTURING
best set 43 MANUFACTRING
best set 42 MANUFACTERING
best set 46 MANUFACTING
best set 42 SMELTING
best set 42 MANUFACTORING
best set 42 MANUFACTRUING
best set 41 MANUFACURING
best set 43 MANUFARTING
best set 42 MANUFATURING

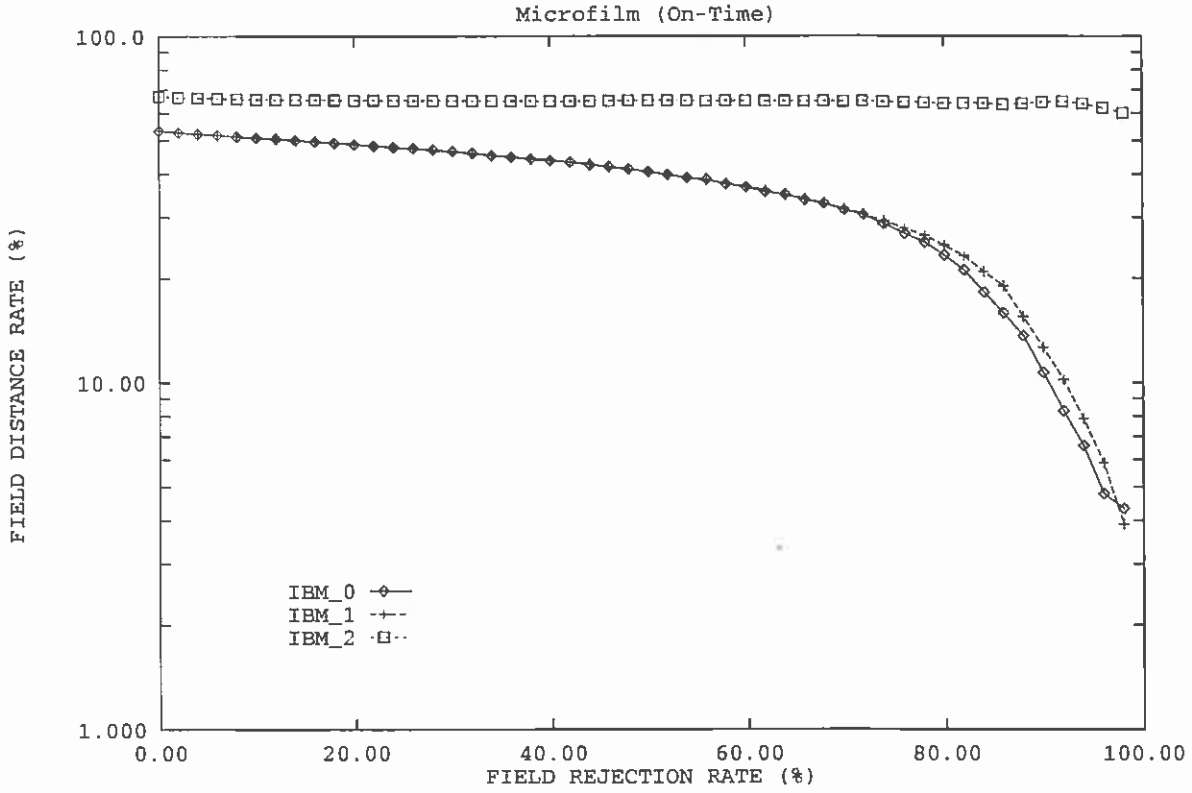
Solution: MANUFACTING

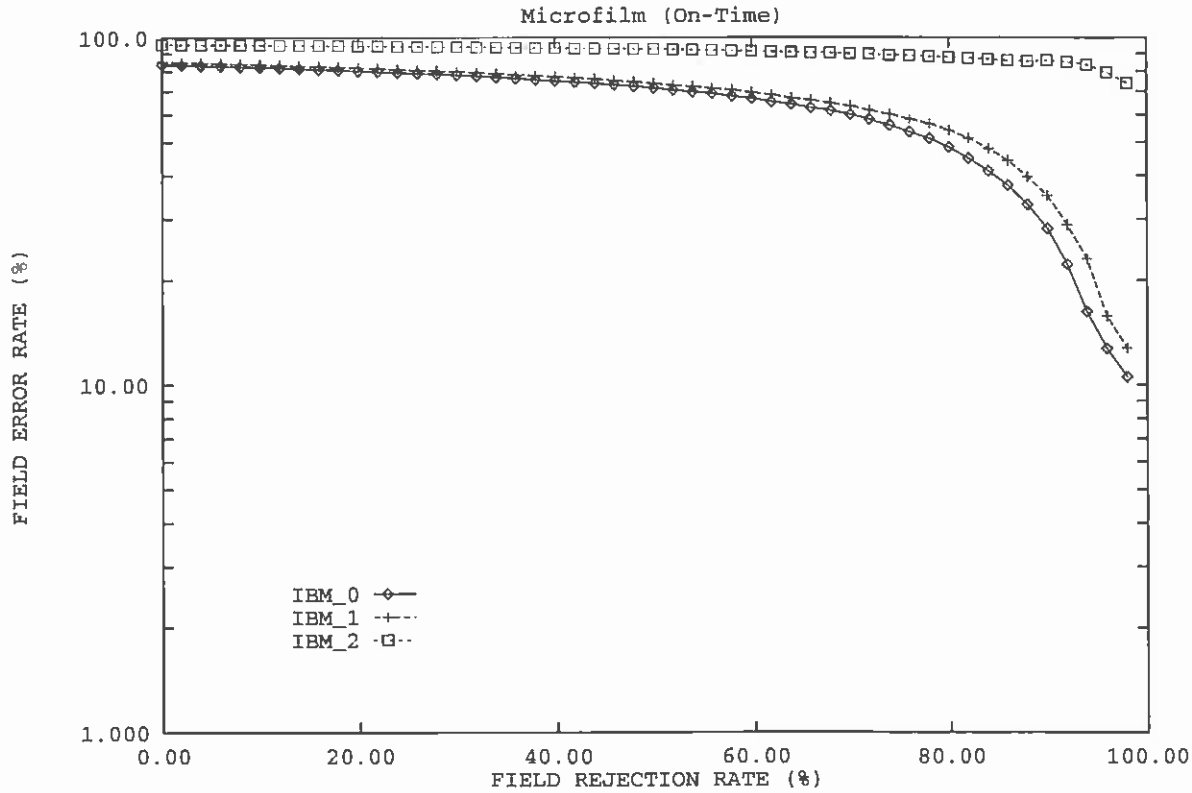

Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

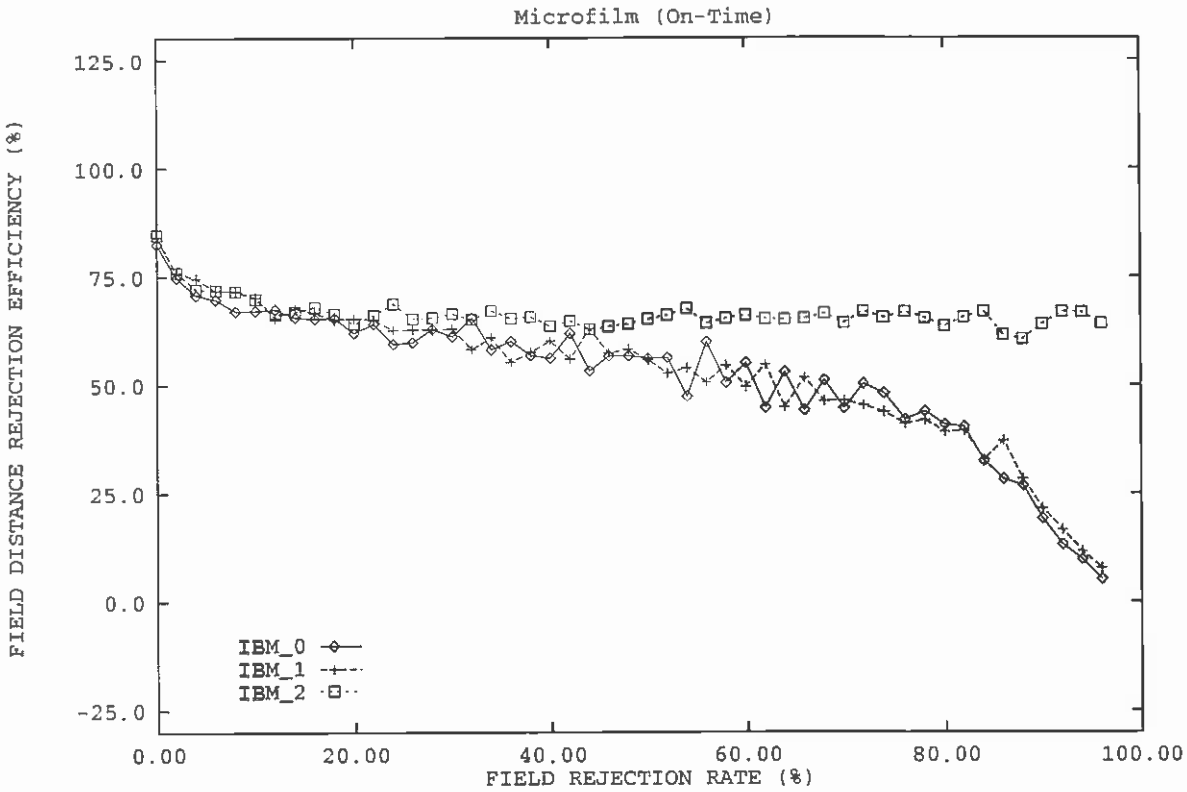

Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

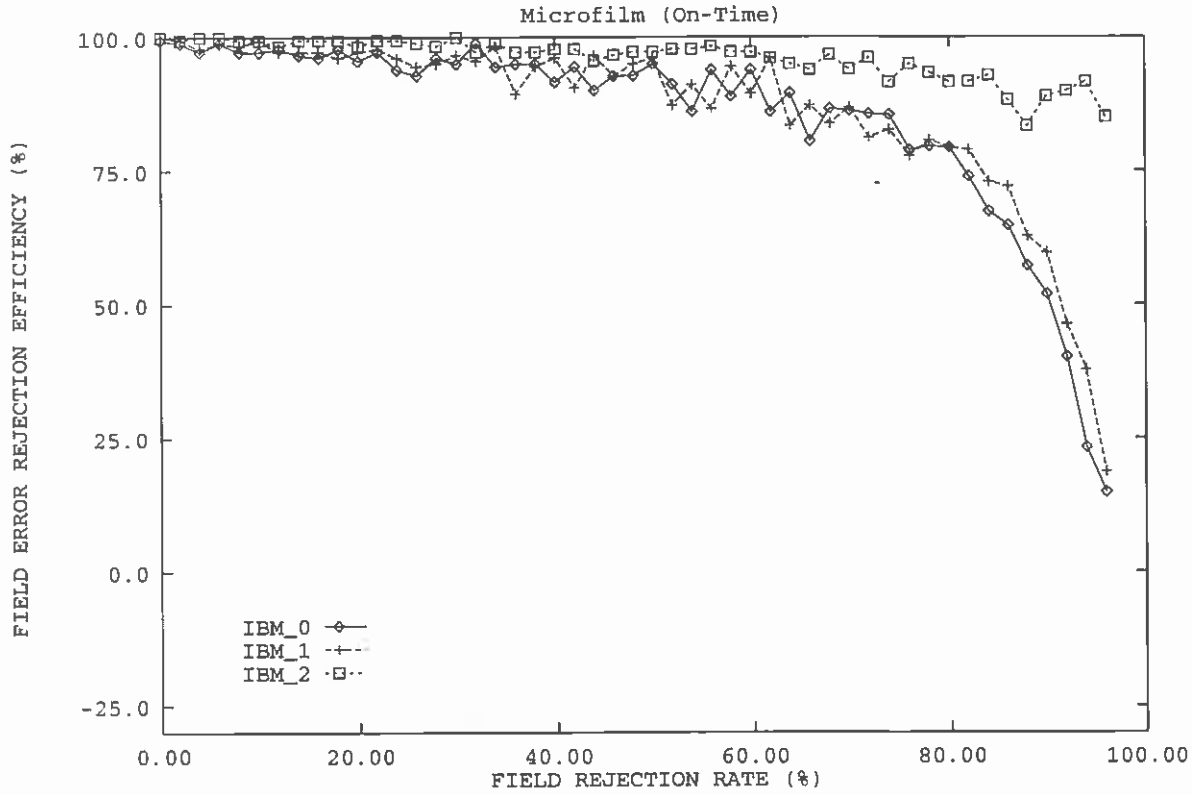

Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IBM_0
IBM_1
IBM_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)
IBM_0
IBM_1
IBM_2


Microfilm (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)
IBM_0
IBM_1
IBM_2

MEMO #94-02
January 1994

IDIAP

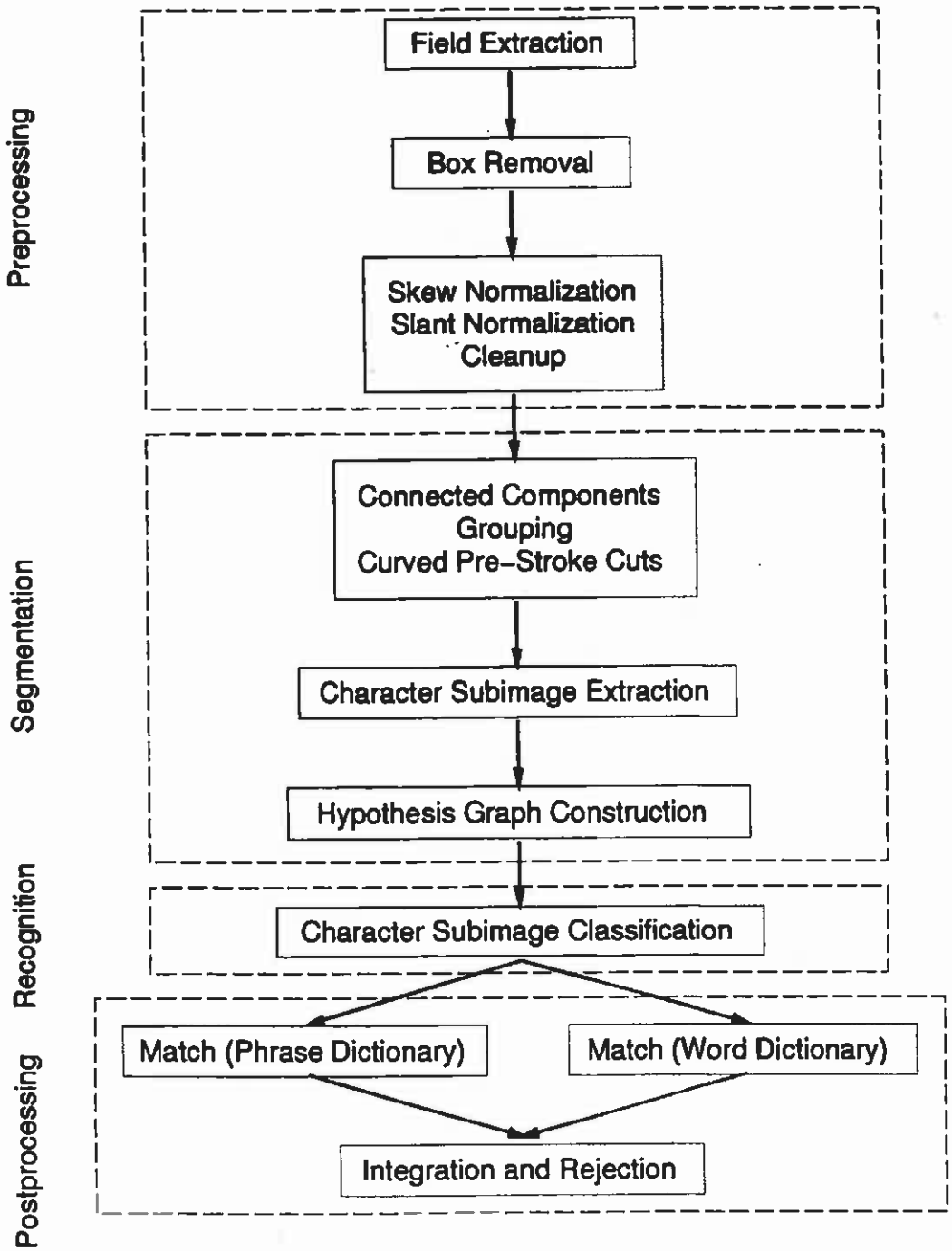
A System for the Off-Line Recognition of Handwritten Text

Thomas M. Breuel

IDIAP, C.P. 609, 1920 Martigny Switzerland
Tel: +41 (26) 22 76 64, FAX: +41 (26) 22 78 18
tmb@maya.idiap.ch

ABSTRACT

A new system for the recognition of handwritten text is described. The system goes from raw, binary scanned images of census forms to ASCII transcriptions of the fields contained within the forms. The first step is to locate and extract the handwritten input from the forms. Then, a large number of character subimages are extracted and individually classified using a MLP (Multi-Layer Perceptron). A Viterbi-like algorithm is used to assemble the individual classified character subimages into optimal interpretations of an input string, taking into account both the quality of the overall segmentation and the degree to which each character subimage of the segmentation matches a character model. The system uses two different statistical language models, one based on a phrase dictionary and the other based on a simple word grammar. Hypotheses from recognition based on each language model are integrated using a decision tree classifier. Results from the application of the system to the recognition of handwritten responses on U.S. census forms are reported.


Preprocessing
Field Extraction
Box Removal
Skew Normalization
Slant Normalization
Cleanup
Segmentation
Connected Components
Grouping
Curved Pre-Stroke Cuts
Character Subimage Extraction
Hypothesis Graph Construction
Recognition
Character Subimage Classification
Postprocessing
Match (Phrase Dictionary)
Match (Word Dictionary)
Integration and Rejection

Noteworthy Features

- complete forms-to-ASCII system
- strictly bottom-up processing
- box/underline removal
- sophisticated character segmentation
- Bayesian, segmental recognizer
- MLP-based character classification
- dictionary backoff using decision trees
- optimized field error (trained)

Writing Styles

	NIST	CEDAR
printed, segm., upper	22%	8%
printed, linked, upper	18%	9%
printed, segm., mixed	21%	3%
printed, linked, mixed	28%	9%
cursive	6%	67%

Notes
- very different composition
- cursive segmenter works poorly for printed styles
- need different segmentation algorithm

Results

- intrinsic errors:
  - not-in-language-model: 34%
  - poor quality: 12%
- evaluation:
  - letter-accurate field transcriptions
  - spaces are not counted (about 1.5%)

Results

All inputs (n=1500)

	0%	25%	50%	75%
Both	37%	21%	6.1%	1.9%
Phrases	42%	27%	8.9%	1.9%
Words	50%	36%	20%	5.1%

Inputs in the language model

	0%	25%	50%	75%
Phrases	12%	2.6%	1.4%	0%
Words	37%	21%	11%	3.7%

Sources of Errors

At 50% rejection (6.1% total):

2.0%	poor estimate of $P(W_i\|x)$
1.6%	minor emendation
1.3%	truncation yielding good phrase
0.3%	transcription error
0.3%	non-character accepted
0.5%	other

Submitted Systems

IDIAP_0	phrase-based, no kerning
IDIAP_1	phrase-based, kerning
IDIAP_2	phrase-based and word-based
IDIAP_3	corrects clerical error in IDIAP_2

Throughput

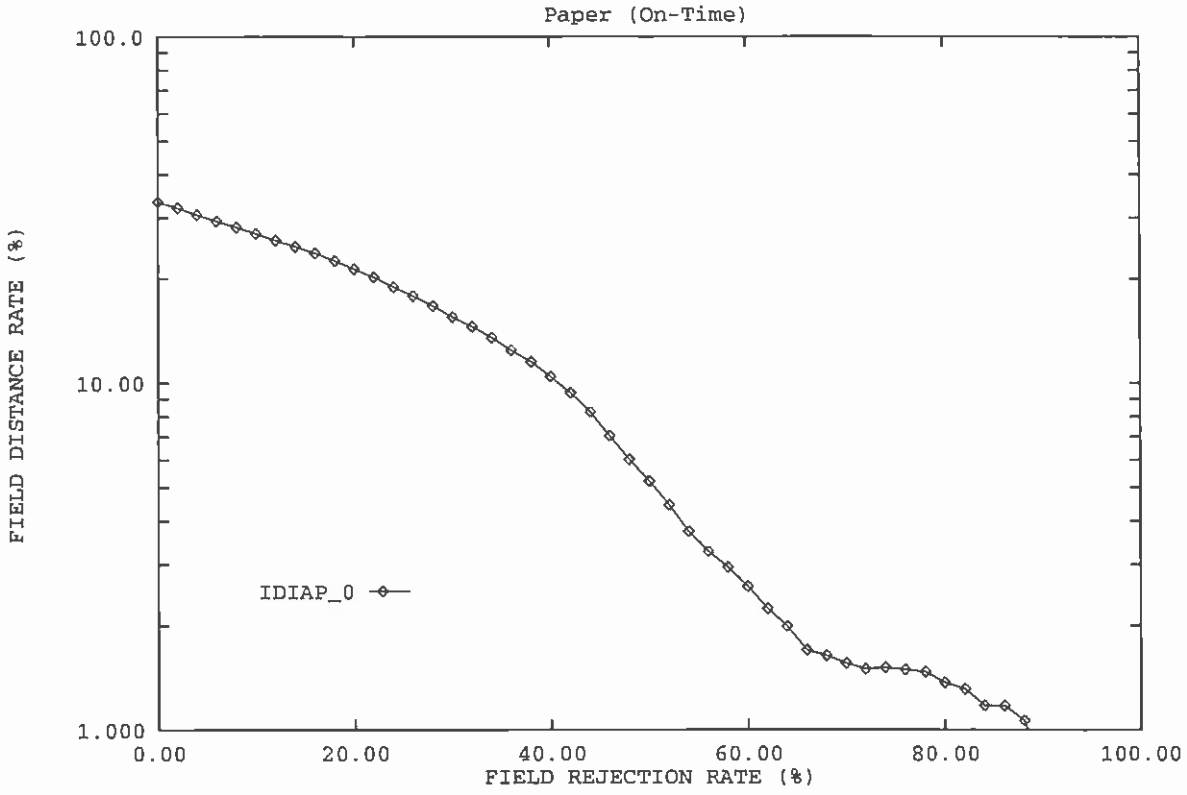
- field extraction: 6 sec
- cleanup: 20 sec
- segmentation/recognition 60 sec

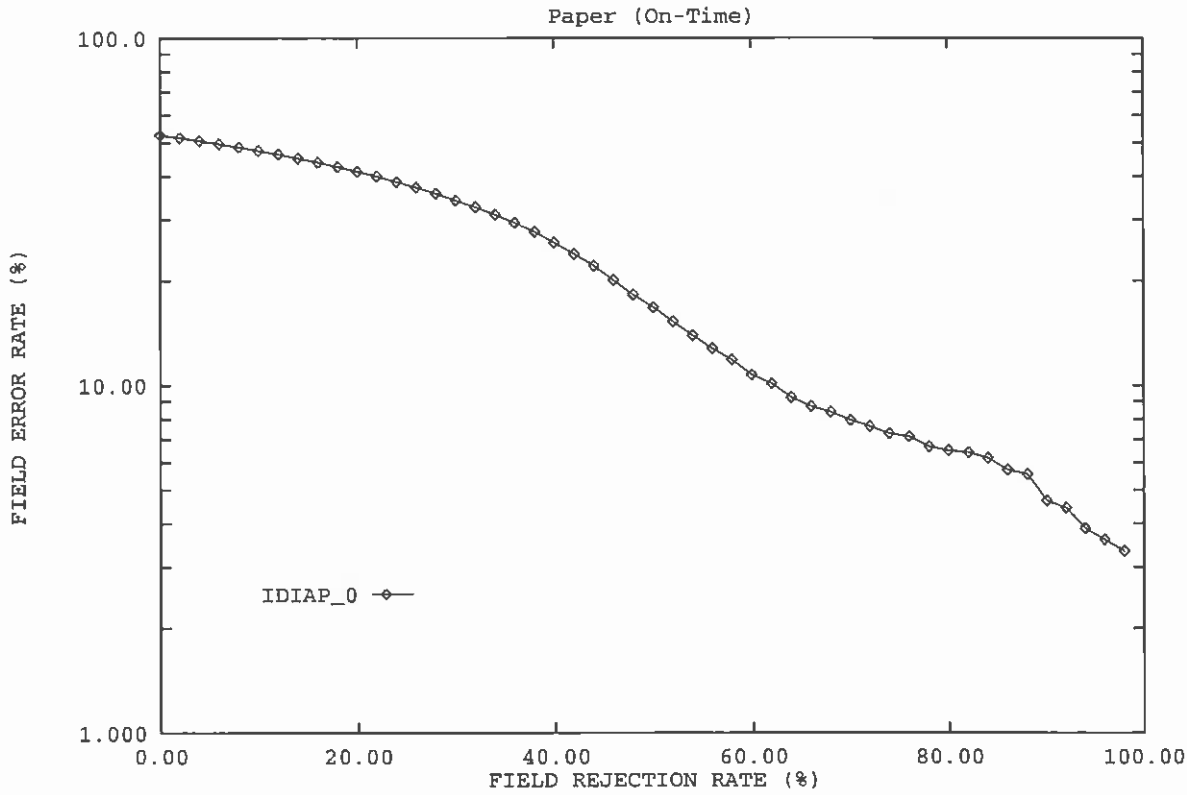
Notes

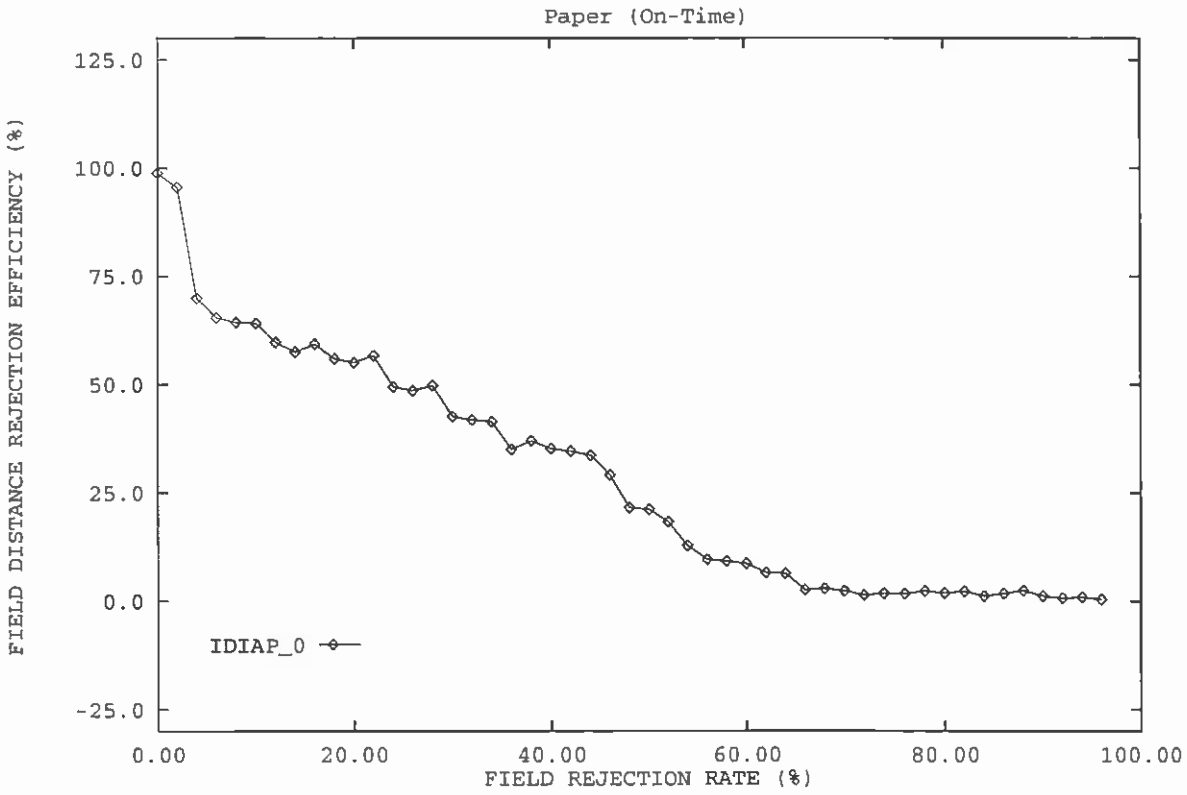
- running on diskless SPARCstation ELC
- expensive intermediate results are recomputed multiple times (e.g., Gaussian convolutions, dilations)
- implemented as UNIX processes communicating via temporary files; significant file I/O overhead
- dictionary reloaded from disk for every field

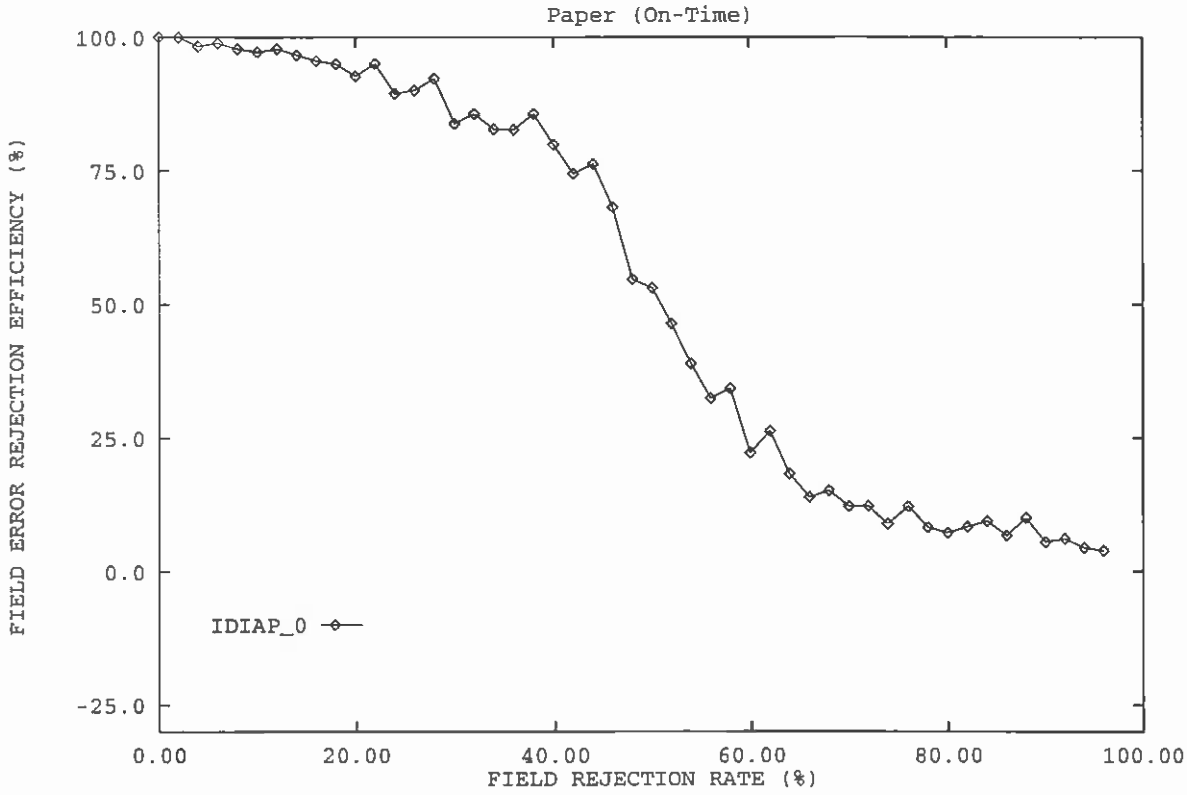
Further Improvements

- better character subimage classification [2.0%]
- task-specific pre-processing (writing outside box, two-line input) [1.3%]
- better language models
- integrate cursive segmentation and cursive models
- better forms removal
- apply to other tasks


Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IDIAP_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IDIAP_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
IDIAP_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
IDIAP_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

The NIST System

Connected Component Segmentation

For each 4-way connected region, "blob" use two NNs and rules to determine if more than one character is present

If lone character recognize it, else attempt segmentation using "snake" to find white space minima between strokes.

Recognition: Size, Translation, Orientation normalization, KL transform, PNN classification to 26 classes.

Post processing: reject or alter improbable hypotheses using digraph probabilities. Spell Correction: Digraph and Levenstein retrieval from whole phrase dictionaries.

The NIST System

Binary Minipages

Blob positioning. Fix orientation by skew

Cut 64 x 512 Fields from fixed coordinates

Morphological Prcossessing: Dilate and Erode

Kill unconnected components on criteria of
Size, Aspect ratio, Location, Proximity
to neighbors, Shape

Identify word spaces from smoothed
minima of the column summations

Writer Dependency

3000 writers.

Number of writers with all three fields correct.

CGK_3	ERIM_0	IDIAP_2	PLUR_0
950	841	766	1283

Number of writers with any two out of three fields correct.

CGK_3	ERIM_0	IDIAP_2	PLUR_0
1018	1073	1084	1040

Number of writers with all three fields incorrect.

CGK_3	ERIM_0	IDIAP_2	PLUR_0
349	330	403	179

No rejection.

PLUR_0 correctly classifies all fields of 43% of the writers.

Separation of Recog nition Performance and Spell Correction Performance?

Performance as submitted to NIST.

CGK_3.P

- field error = 0.381 @ 40% reject 0.141
- field distance = 0.207 @ 40% reject 0.037

ERIM_0.P

- field error = 0.397 @ 40% reject 0.212
- field distance = 0.187 @ 40% reject 0.109

IDIAP_2.P

- field error = 0.421 @ 40% reject 0.149
- field distance = 0.220 @ 40% reject 0.050

Apply NIST Spell correction Algorithm to real submission.

CGK_3.P

- field error = 0.210 @ 40% reject 0.080
- field distance = 0.154 @ 40% reject 0.056

ERIM_0.P

- field error = 0.190 @ 40% reject 0.091
- field distance = 0.141 @ 40% reject 0.048

IDIAP_2.P

- field error = 0.247 @ 40% reject 0.086
- field distance = 0.190 @ 40% reject 0.052

Possible solutions to the Coverage Problem

Augment the SD13 "phrase_123.sht" dictionaries.
Number of unique phrases used.

CGK_3.P	ERIM_0.P	IDIAP_2.P
5929	5326	5226

Number of phrases that are NOT in the SD13 phrase_123.sht dictionaries, including multiple occurences.

CGK_3.P	ERIM_0.P	IDIAP_2.P
3869	2708	1998

Do not force spell correction. Use raw hypotheses or partial words. Number of phrases that fail UNIX spell.

CGK_3.P	ERIM_0.P	IDIAP_2.P
2261	660	171

Number of phrases that are NOT in the reference files, including multiple occurences.

CGK_3.P	ERIM_0.P	IDIAP_2.P
2886	2963	3033

Cheat! Use the concatenated reference files as dictionaries!

How well do they SD13 dictionaries cover the actual references?

Number of reference answers unavailable to systems which used only the "phrase_?.sht" dictionaries of SD13 is 3105 out of 5378.

and counting multiple occurences in the reference files number rises to 3206 out of 9000.

The number of reference answers unavailable to systems which used only the "phrase_?.sht" dictionaries of SD13 for their intended field is 3375 out of 9000.

Field error rate cannot be below 37.5%.

That assumes phrases are not split into words. Coverage is extended greatly if they are.

Number of reference words unavailable to systems which used only the "word_123" dictionaries of SD13 is 1789 out of 3186.

and counting multiple occurences in the reference files the number rises to 10309 out of 17770.

Are Fields Different?

Number of unique entries in SD13 phrase_123.sht and word_123.sht dictionaries.

	Phrase	Word
• Field 1	8340	14447
• Field 2	8633	14432
• Field 3	7956	17110

Number of unique phrases in the reference files. Phrases broken into words and sorted uniquely.

	Phrase	Word
• Field 1	1910	1502
• Field 2	1784	1379
• Field 3	2125	1847

Number of phrases in the reference files occuring only once. Also as a percentage of the total number of phrases = 3000.

	Number	%age
• Field 1	1628	54.3
• Field 2	1494	49.8
• Field 3	1895	63.1

Are Fields Different?

FIELD ERROR

	CGK_3.P	ERIM_0.P	IDIAP_2.P
• Field 1	0.372	0.375	0.413
• Field 2	0.345	0.355	0.370
• Field 3	0.426	0.462	0.479

FIELD DISTANCE

	CGK_3.P	ERIM_0.P	IDIAP_2.P
• Field 1	0.196	0.172	0.217
• Field 2	0.173	0.153	0.181
• Field 3	0.245	0.230	0.256

IDIAP have a 10.9% more success in field 2 than field 3.

Is this Significant?

Complementary Systems
Superior Performance from a Majority System

Plurality PLUR_0 system from CGK_3 ERIM_0 IDIAP_2.

- If all three concur use that hypothesis with highest confidence.
- If only any two concur use that hypothesis with highest confidence of the majority pair.
- If all three differ use the hypothesis with highest confidence.

PLUR_0.P

- field error = 0.286 @ 40% reject 0.079
- field distance = 0.158 @ 40% reject 0.029

CGK_3.P

- field error = 0.381 @ 40% reject 0.141
- field distance = 0.207 @ 40% reject 0.037

ERIM_0.P

- field error = 0.397 @ 40% reject 0.212
- field distance = 0.187 @ 40% reject 0.109

IDIAP_2.P

- field error = 0.421 @ 40% reject 0.149
- field distance = 0.220 @ 40% reject 0.050

Improvements on this Majority.

- Use more voting systems, if available, and a plurality.
- Coherent generation or use of confidences.
- Normalization Scheme.

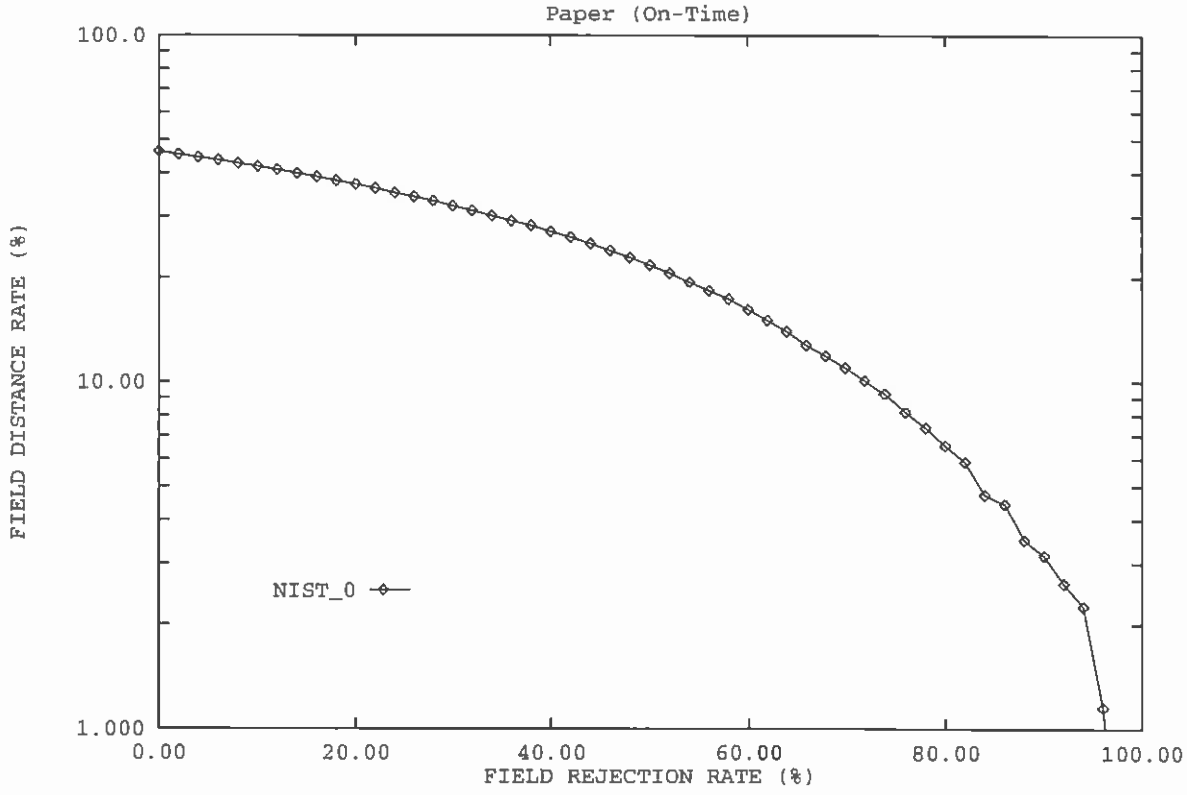
Are the best systems failing on the same fields?

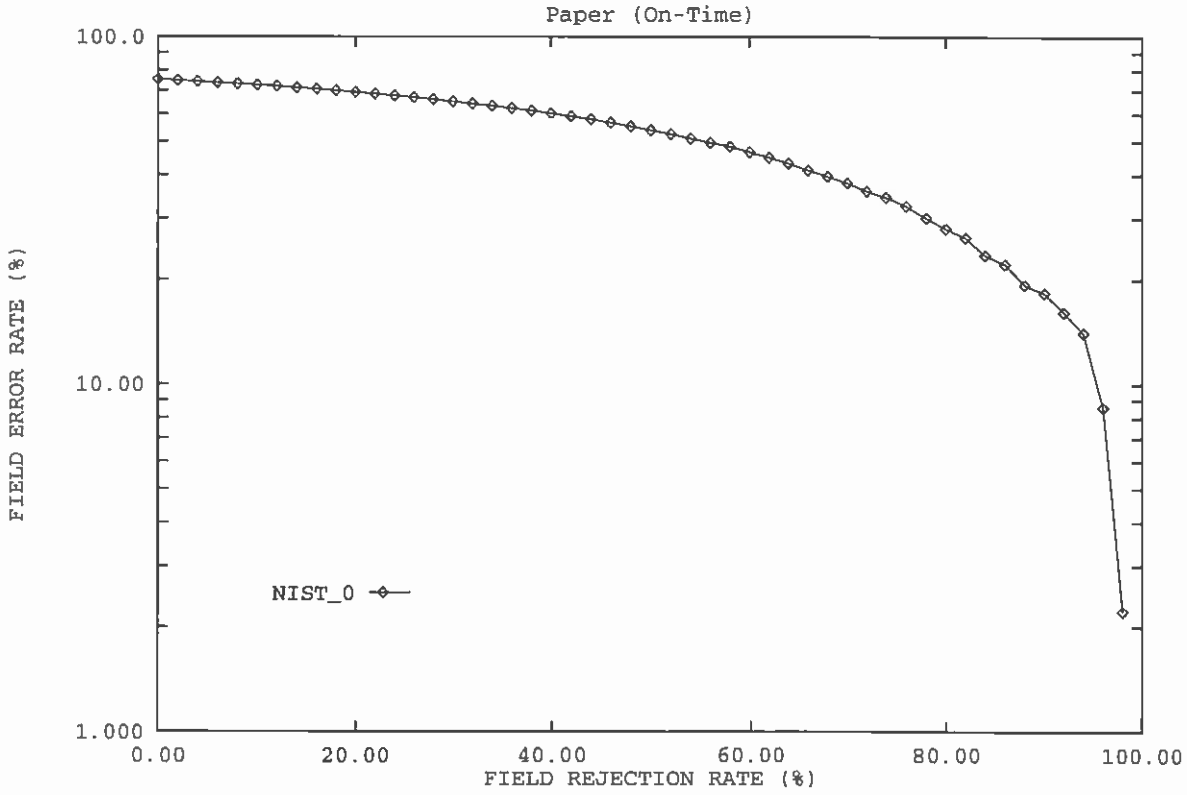
Number of fields where both systems wrong = W

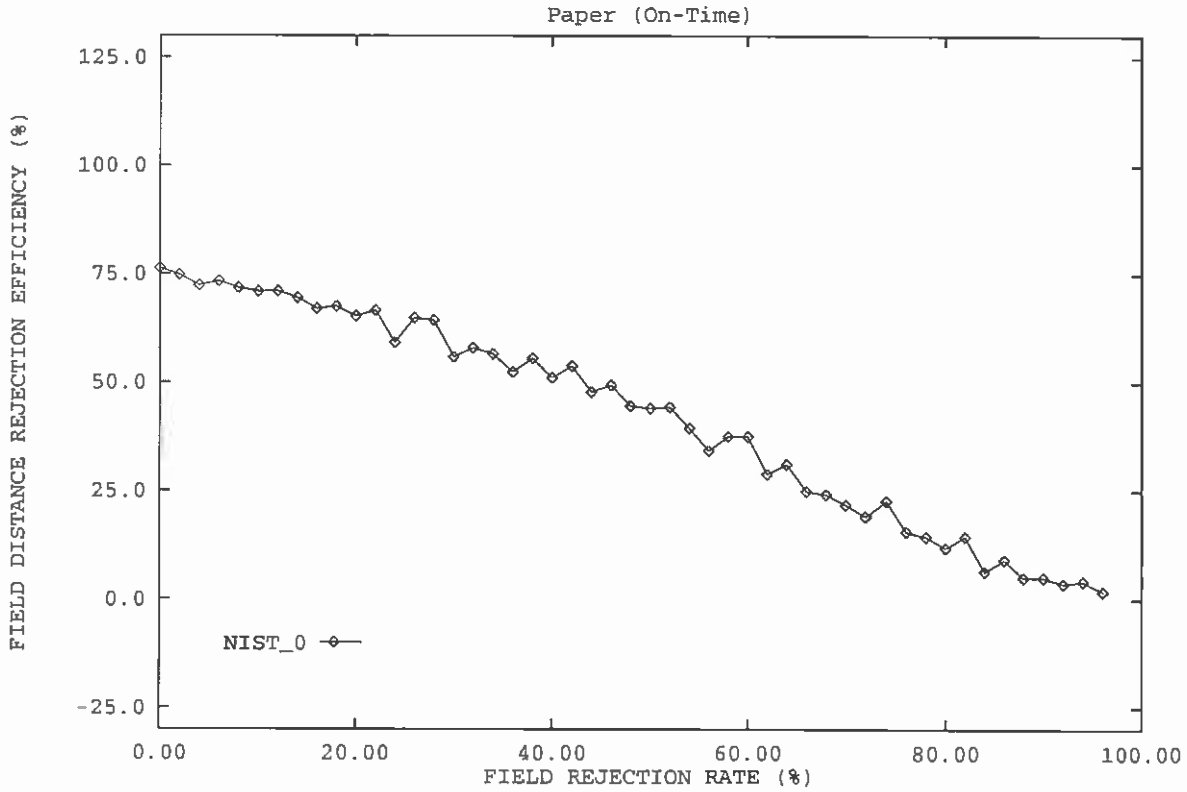
	CGK_3	ERIM_0	IDIAP_3
CGK_3	W = 3431		
ERIM_0	W = 2161	W = 3575	
IDIAP_3	W = 2332	W = 2709	W = 3787

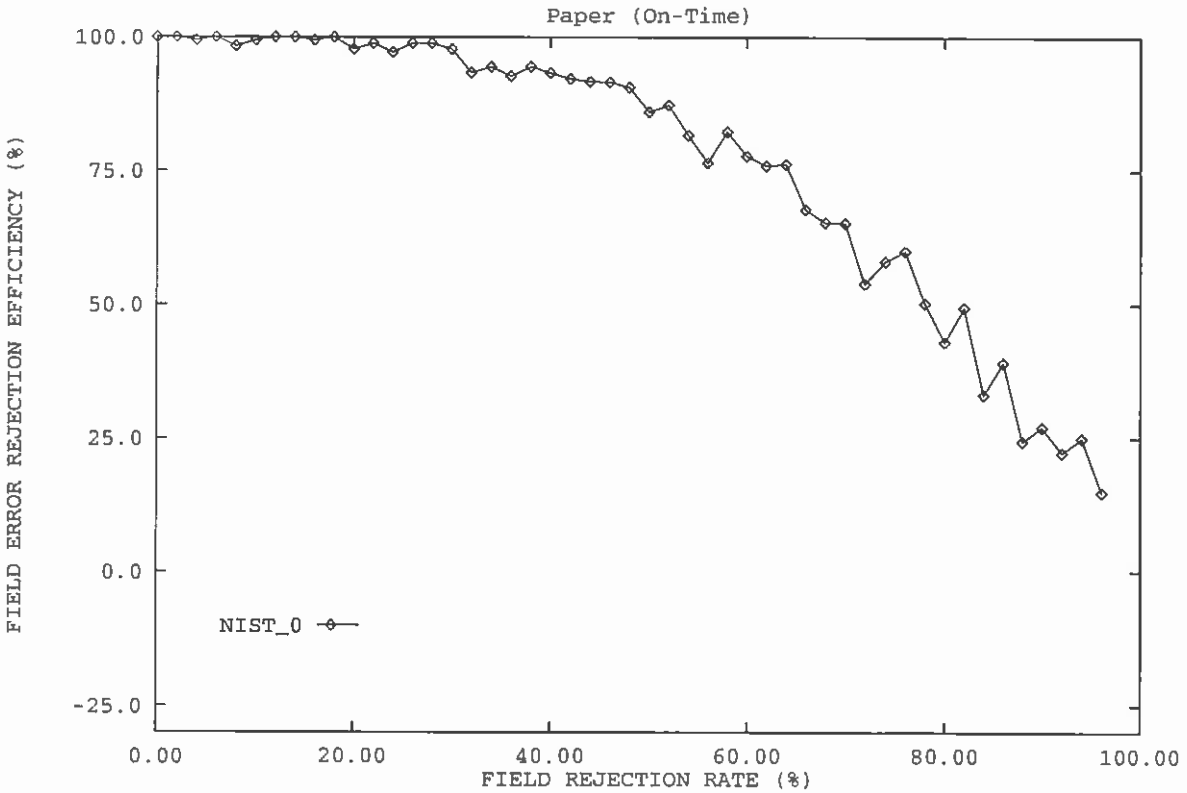
CGK fail on 1270 fields that ERIM succeed on
ERIM fail on 1414 fields that CGK succeed on
IDIAP fail on 1078 fields that ERIM succeed on

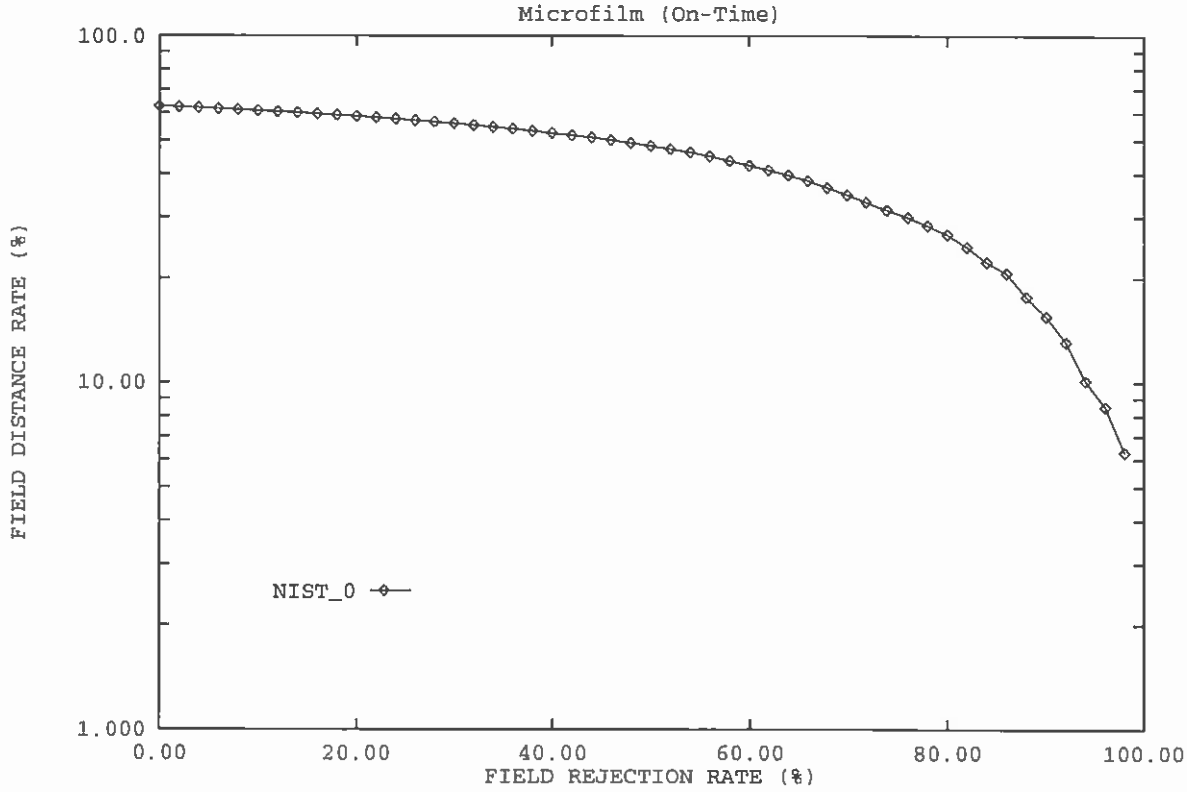
The systems fail on different fields.
Can this be exploited?

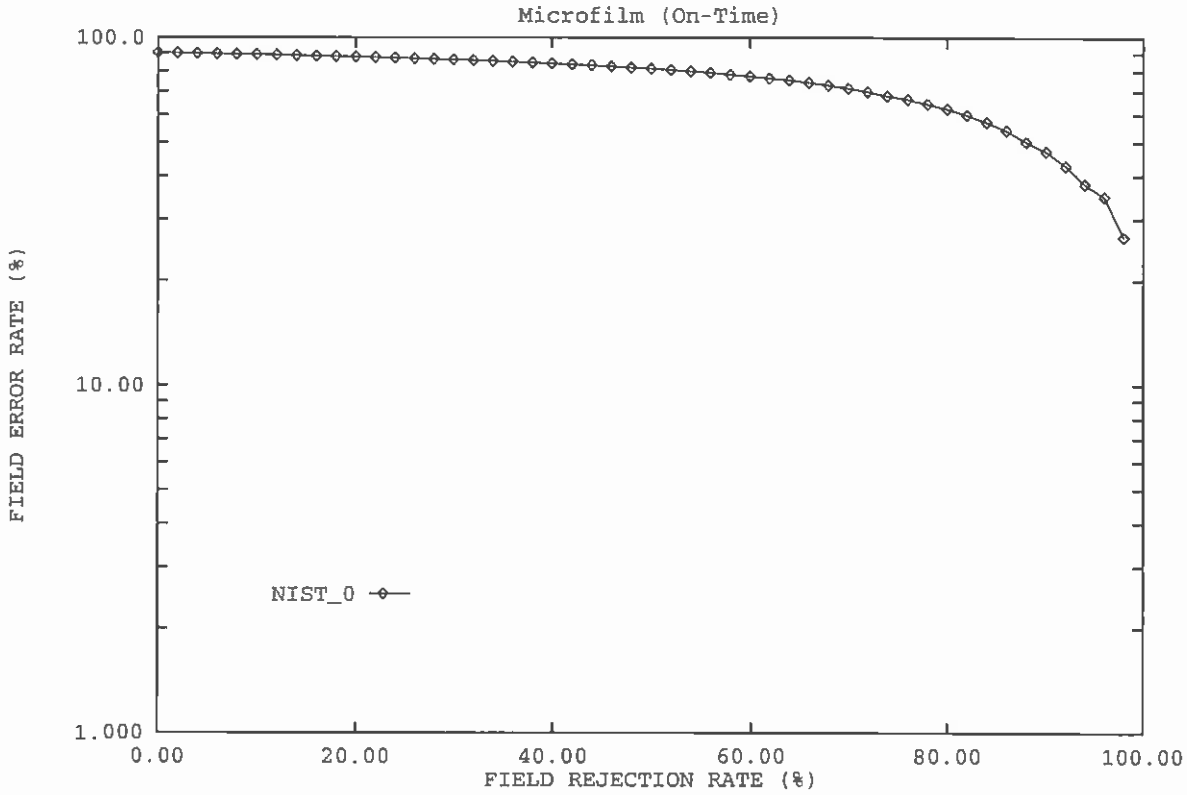

Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

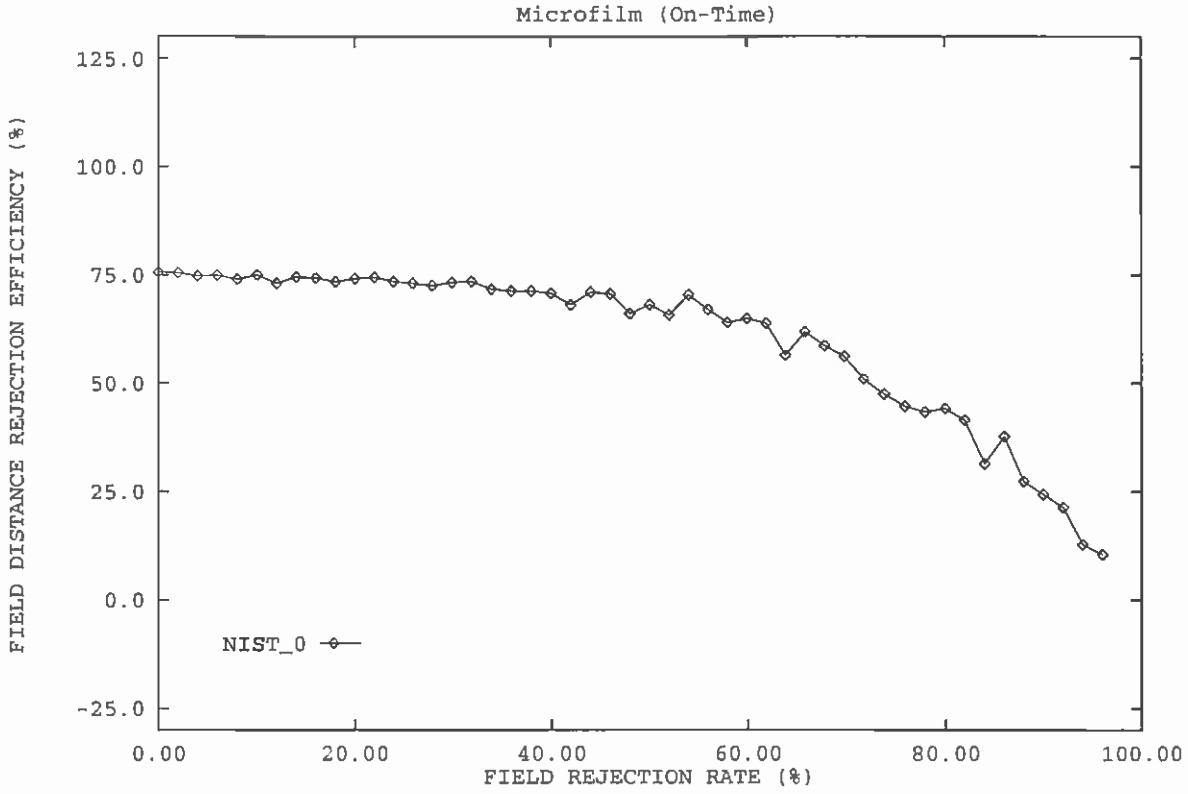

Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

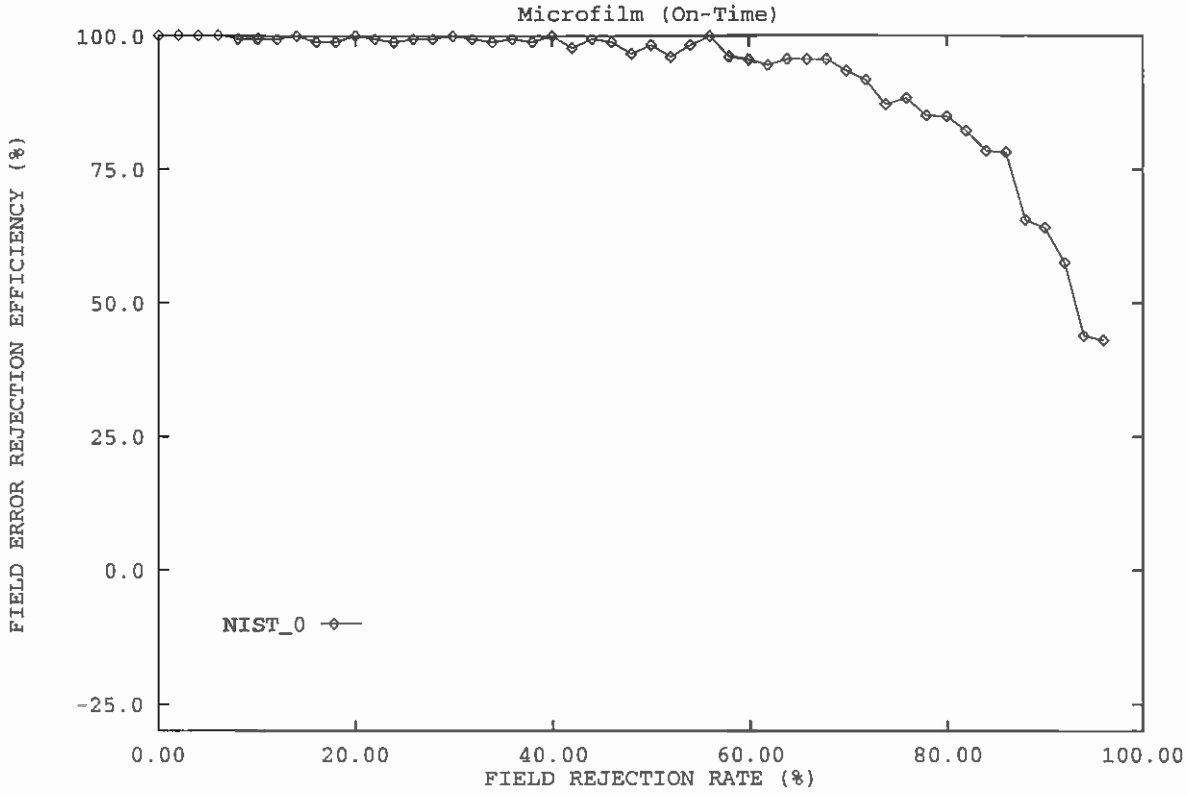

Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2
UBOL
1994

University of Bologna

Department of
Electronics and Computer Sciences
(D.E.I.S.)

ITALY

Zs. M. Kovacs L. Simoncini

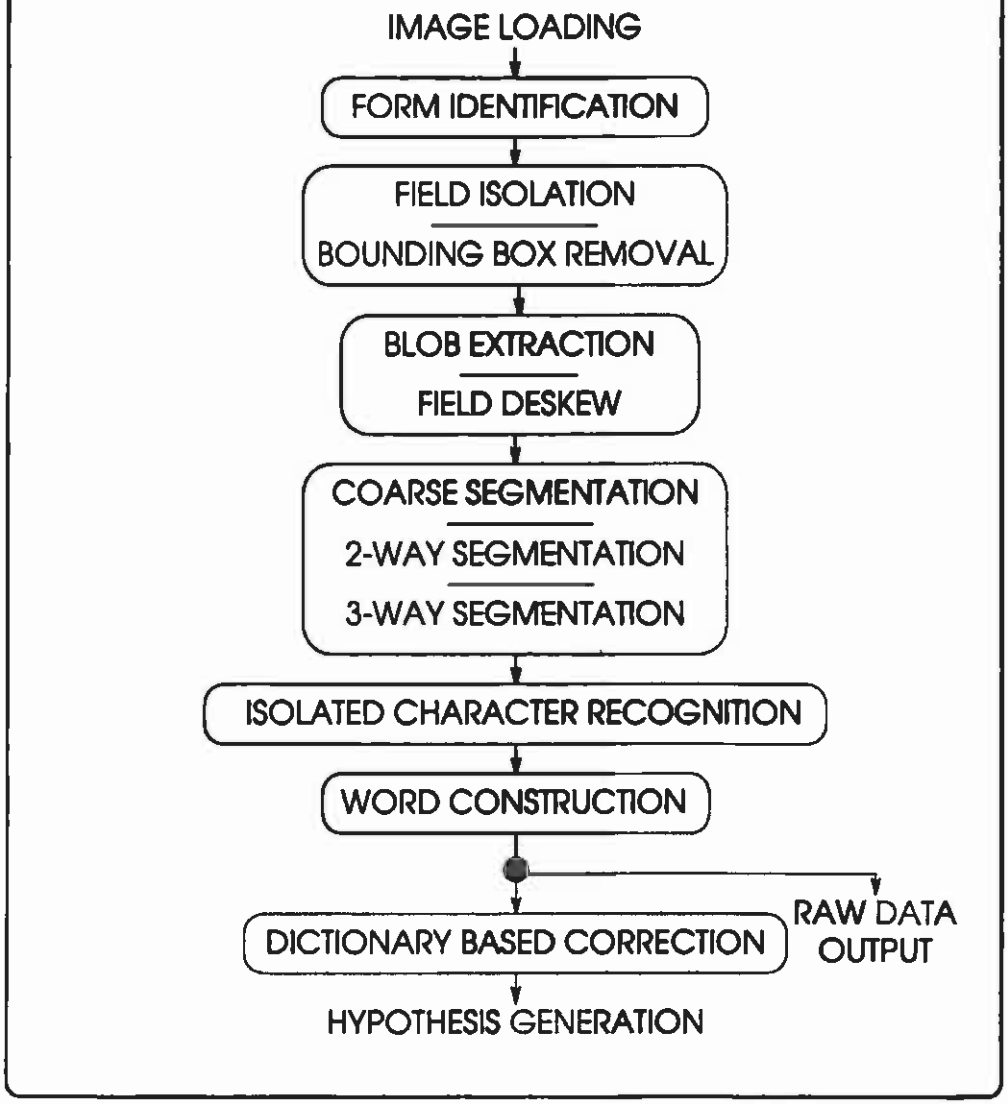
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

System overview



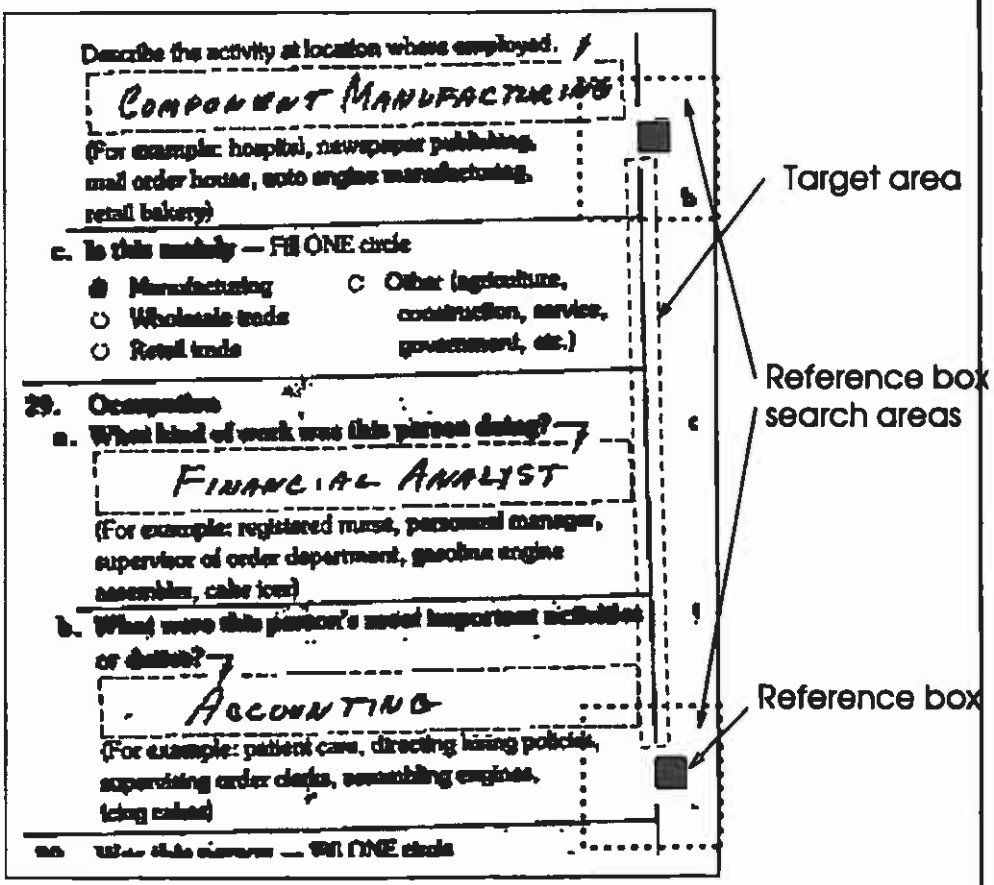
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Form identification



- Reject rate: 1/4000

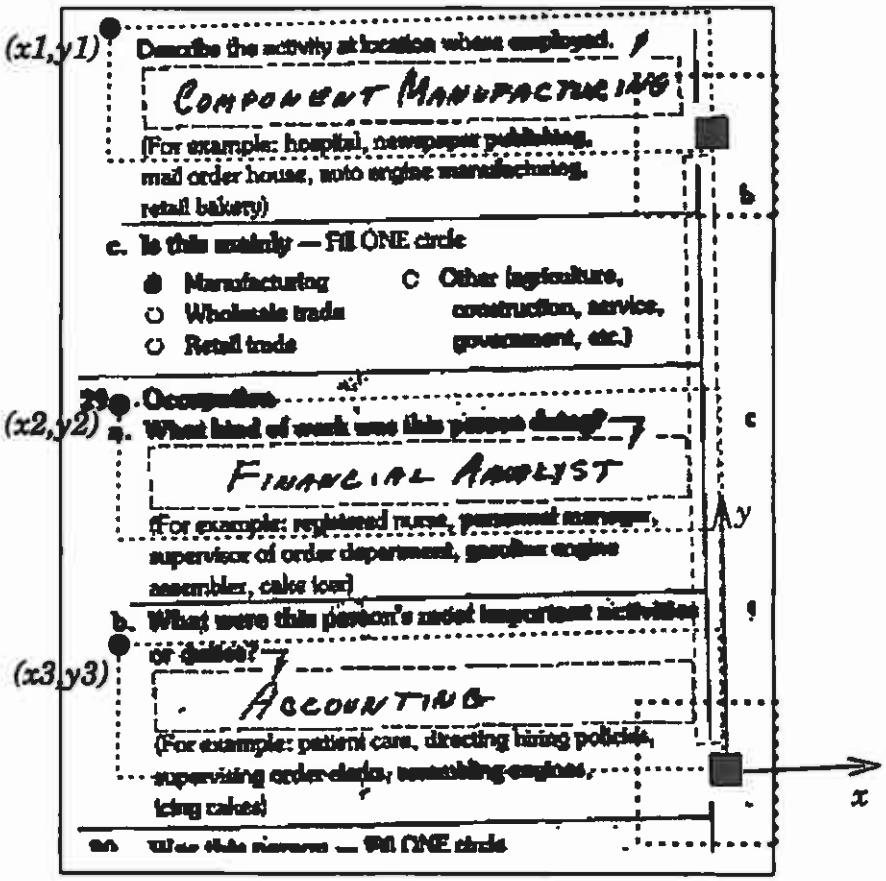
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Form isolation



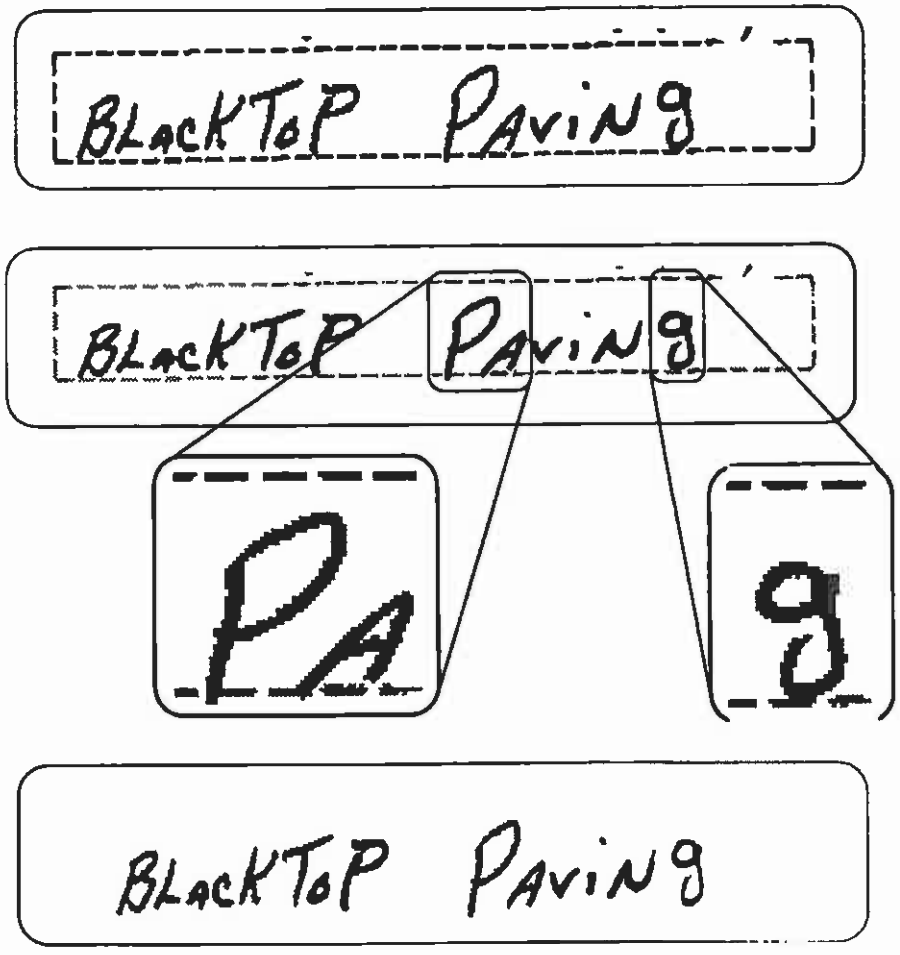
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Bounding box removal



UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Blob extraction

JOB SETTER

JOB SETTER

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Field deskew

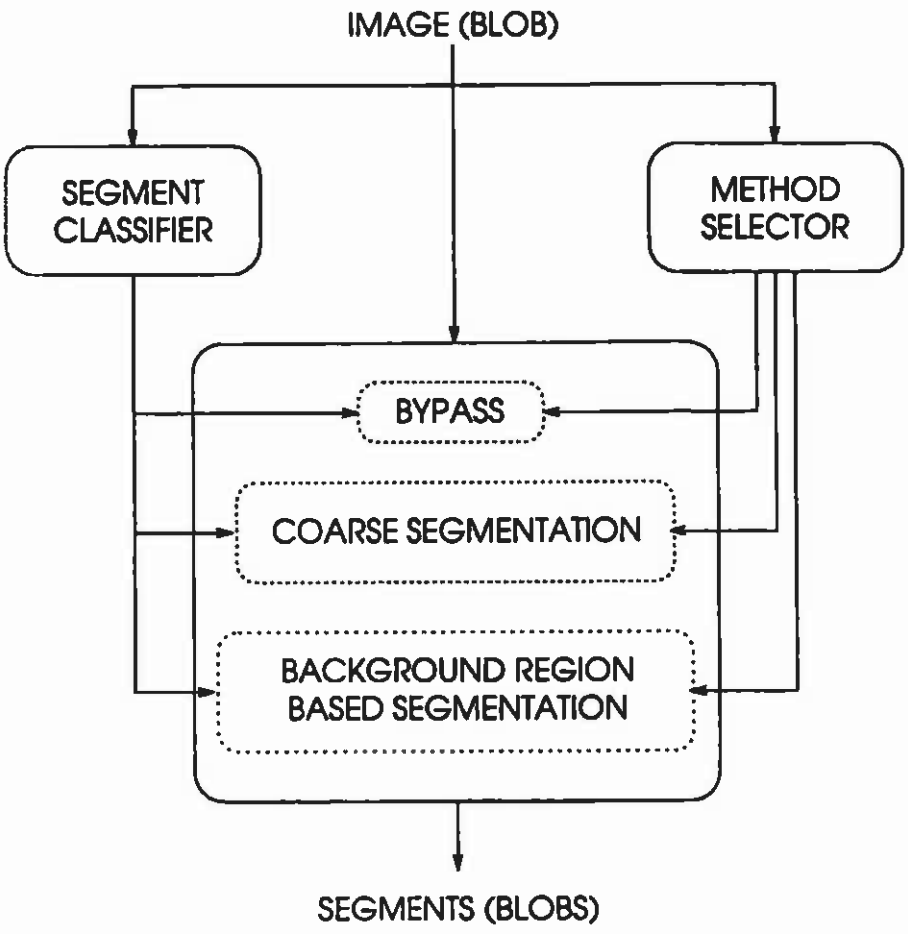
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Segmentation



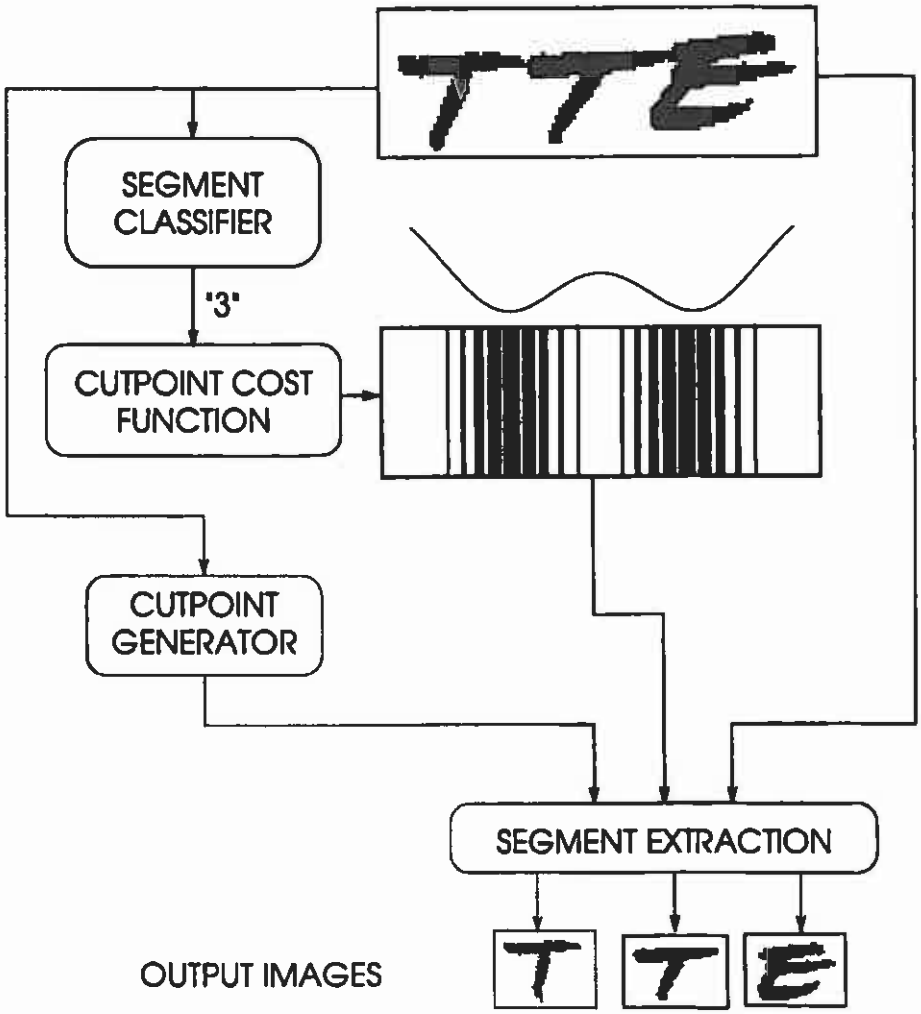
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Coarse segmentation



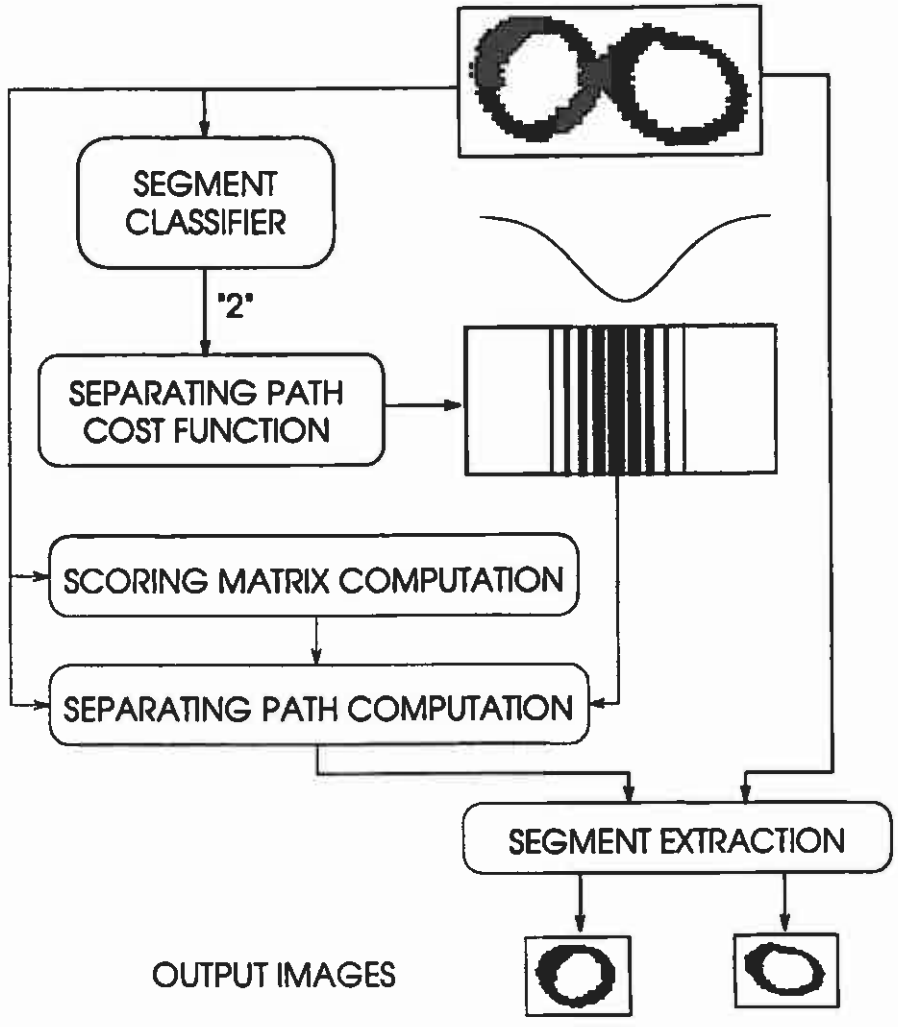
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Background region based segmentation



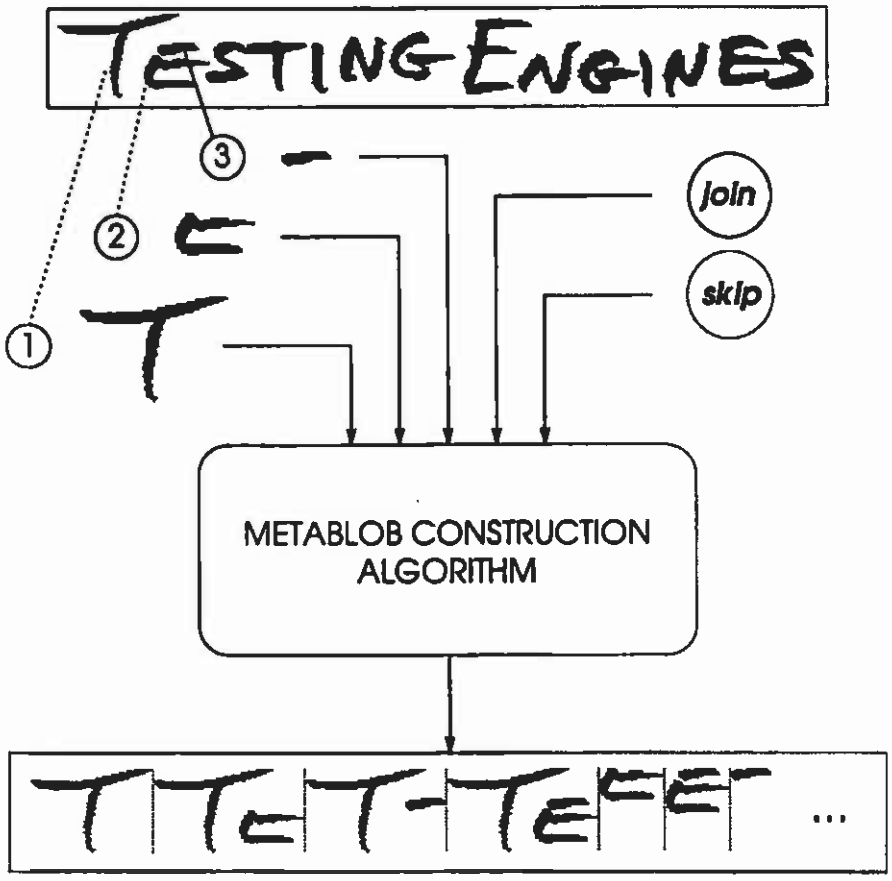
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Metablob construction



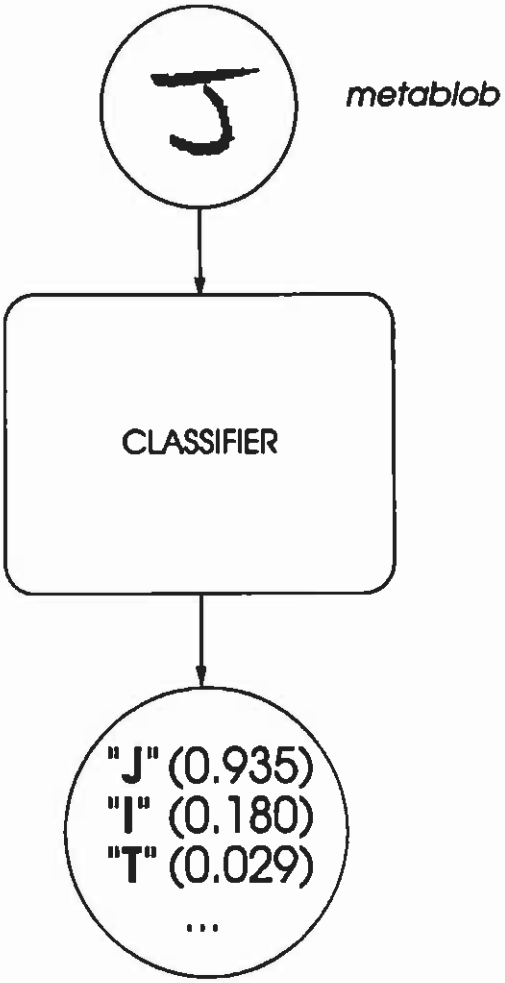
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Isolated character recognition



UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2
UBOL
1994

Features

DISTANCE TRANSFORM

CHAIN CODE HISTOGRAMS

OTHER GEOMETRIC INFORMATIONS

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Classifier

NETWORK

MULTI-LAYER PERCEPTRON

TRAINING

NIST Special Database 3
45,000 uppers
45,000 lowers
223,000 digits

CEDAR CDROM 1
12,000 uppers
8,000 lowers

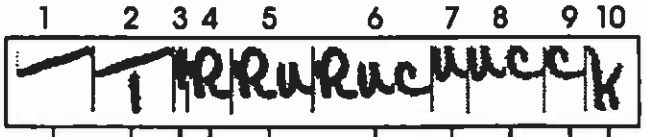
UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Word building



1	2	3	4	5	6	7	8	9	10
"I" 0.511 "L" 0.505 "J" 0.502 "I" 0.501 "R" 0.5 "A" 0.5	"T" 0.86 "F" 0.521	"L" 0.998 "I" 0.997 "J" 0.506	"R" 0.996 "Q" 0.56	"M" 0.806 "Q" 0.519	"M" 0.968 "U" 0.593	"U" 0.807 "L" 0.53	"N" 0.648 "W" 0.503	"C" 0.996 "L" 0.512	"K" 0.548 "R" 0.501 "G" 0.5

WORDS:

2, 5, 9, 10
1, 3, 4, 7, 10
2, 4, 7, 9, 10
3, 5, 9, 10
4, 7, 9, 10

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2
UBOL
1994

Dictionary based Correction

AGREP package

Developed by

Sun Wu **Udi Manber**

at the University Of Arizona
Department Of Computer Science

Three level Dictionary Check:

1 - Top choice match

2 - Multiple choice match

3 - Inexact match

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Correction step 1

TOP CHOICE MATCH

Truck

T R U C K

J A V E R

N L

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Correction step 2

MULTIPLE CHOICE MATCH

Truck

T R N E K

J A U C R

V L

UNIVERSITA' DEGLI STUDI DI BOLOGNA
DIPARTIMENTO DI ELETTRONICA
INFORMATICA E SISTEMISTICA
Viale Risorgimento 2 - 40136 Bologna - ITALY

COCR2

UBOL

1994

Correction step 3

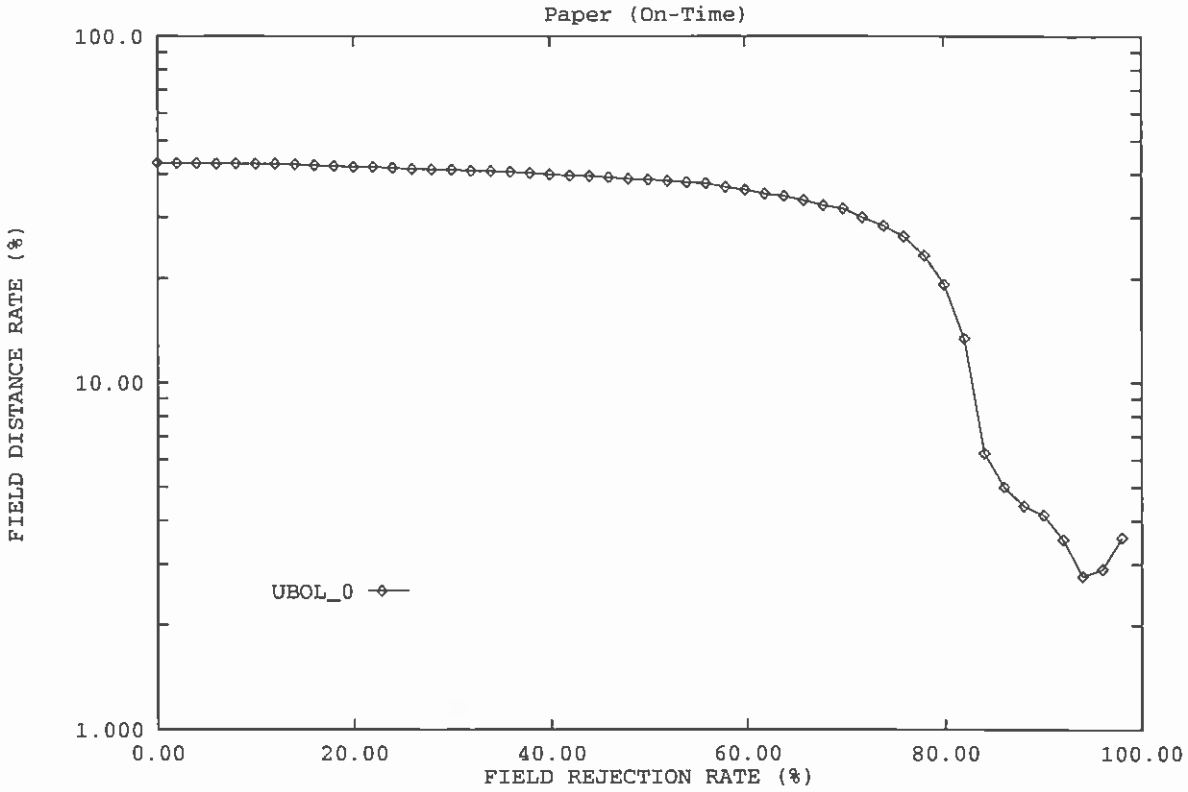
INEXACT MATCH

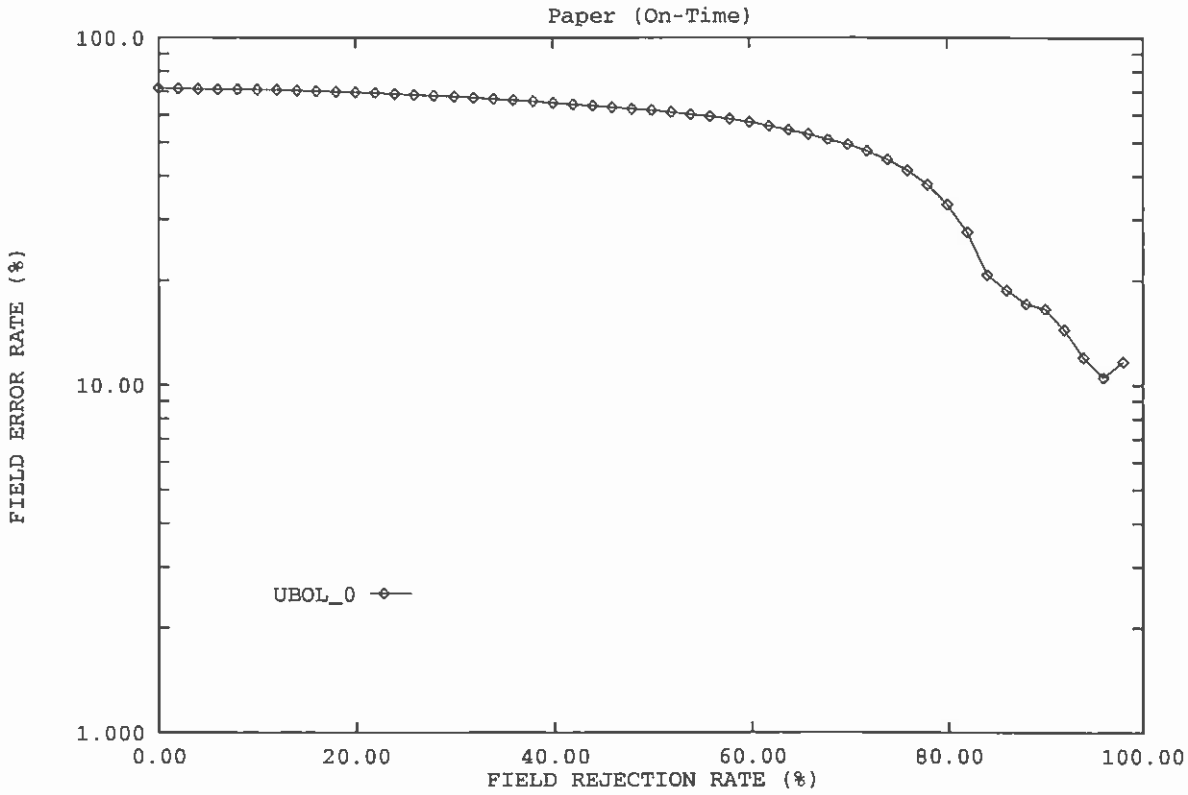
K

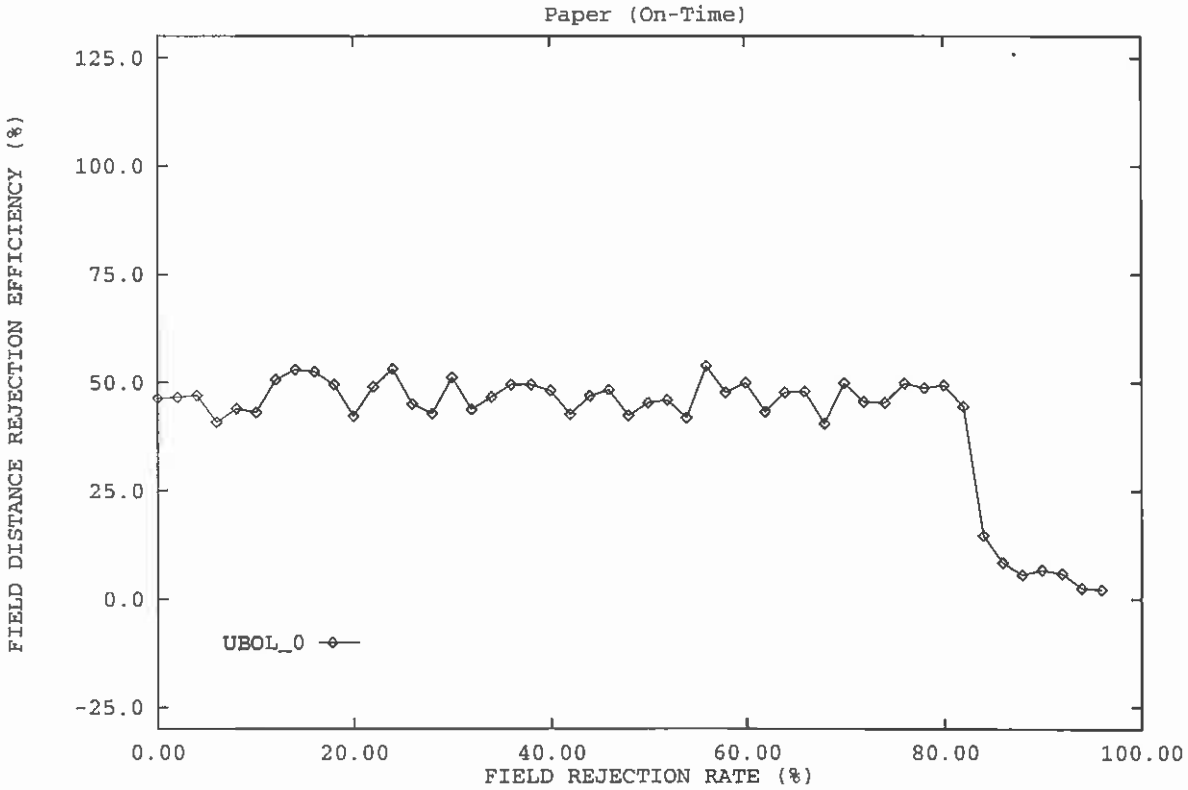
T R N C R

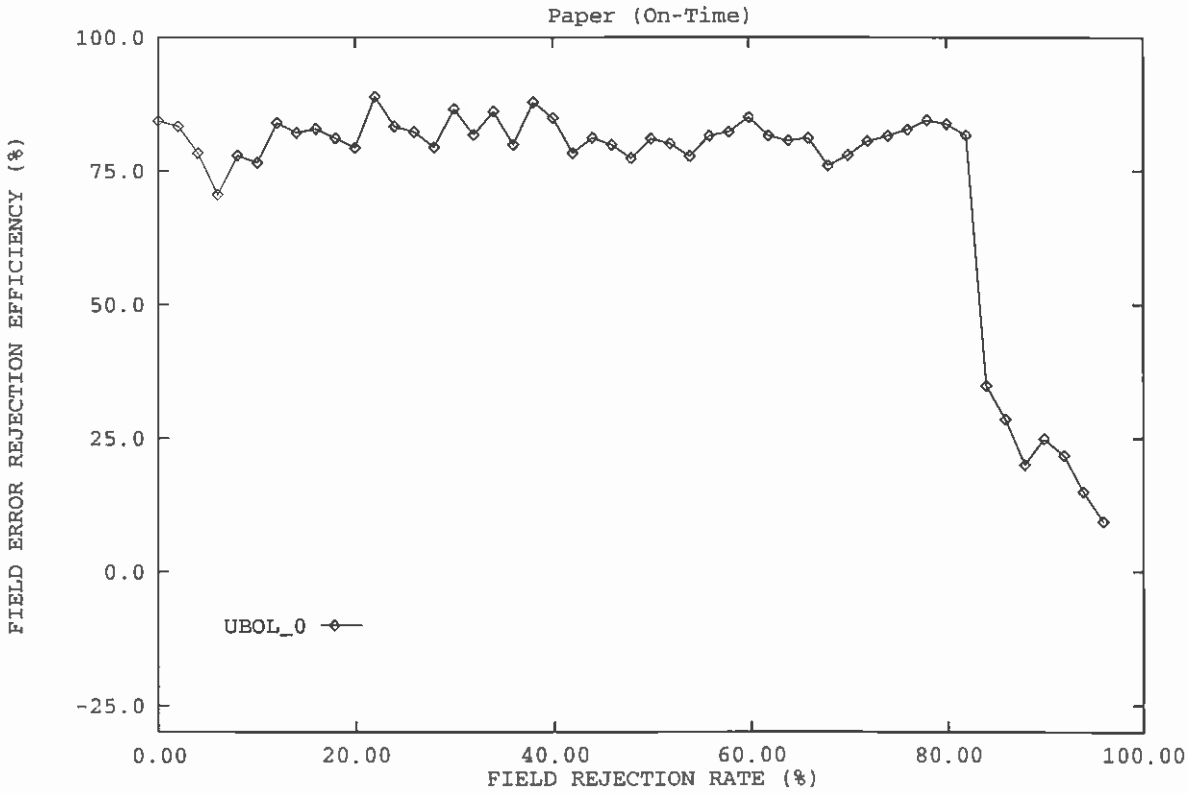
J A U E

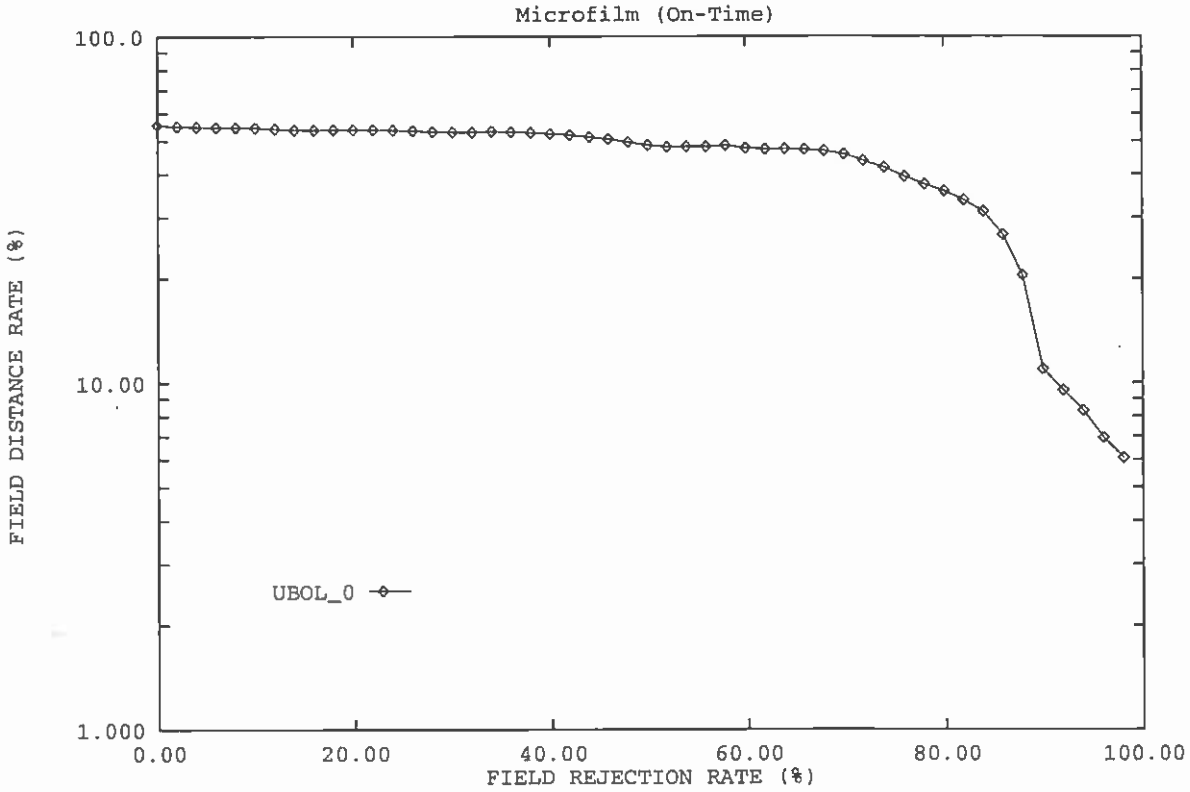
V L

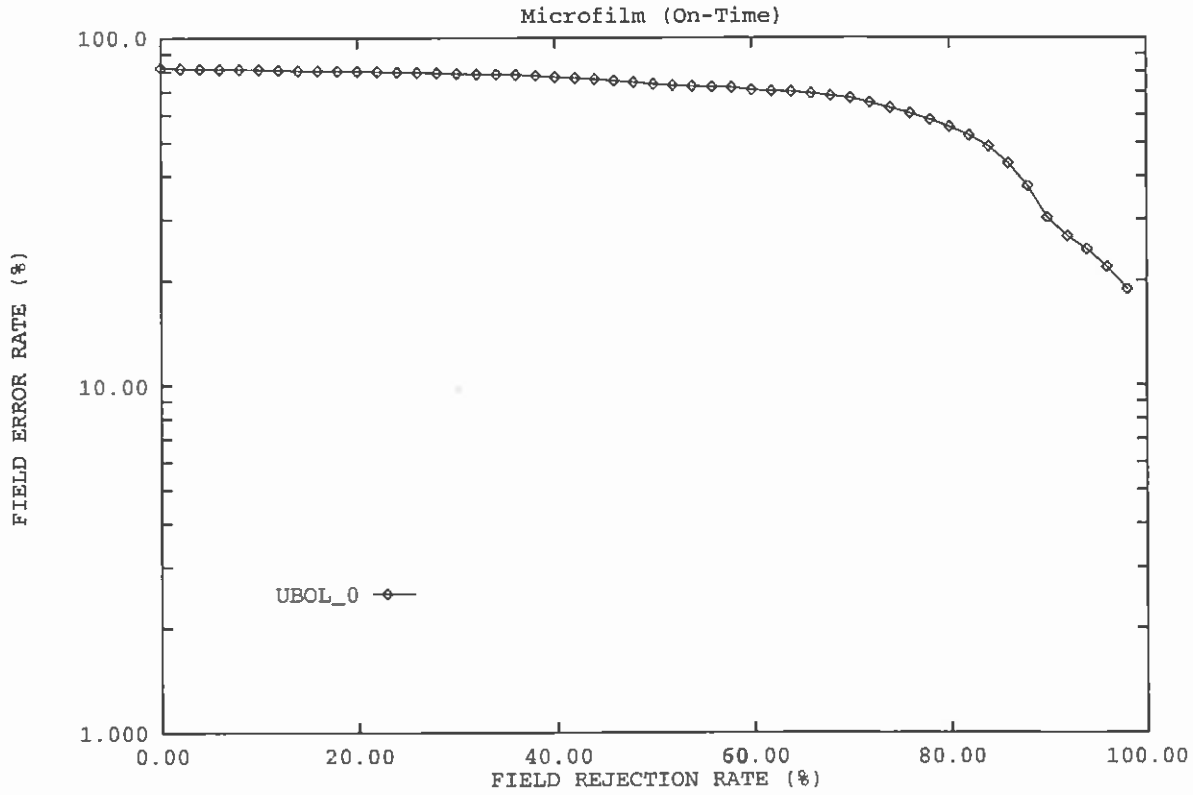

Paper (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

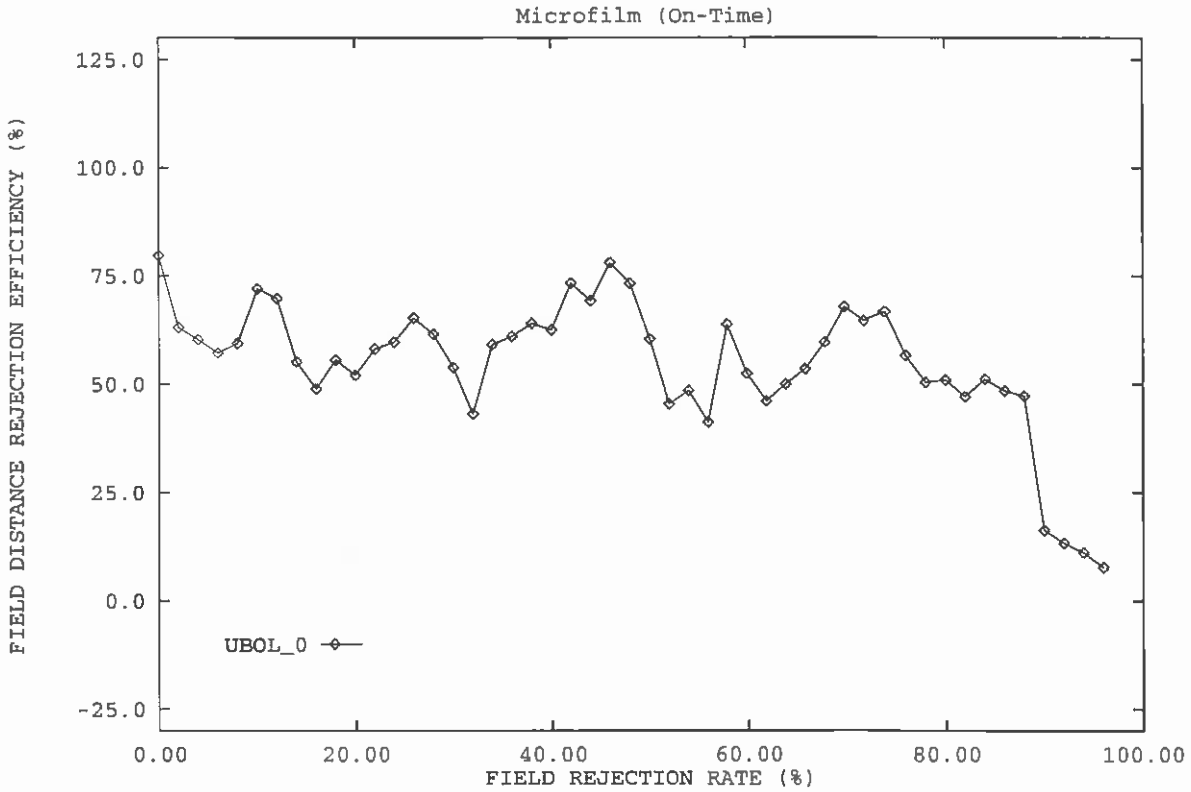

Paper (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

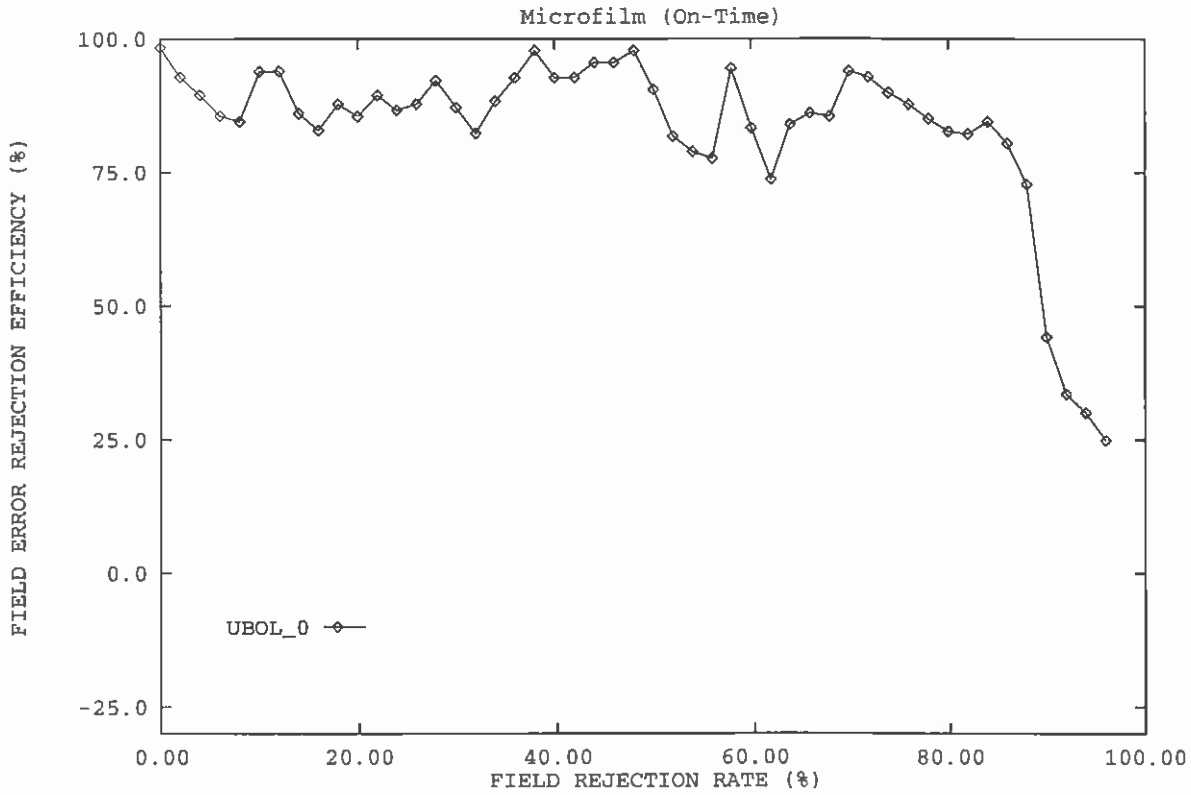

Paper (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
100.0
10.00
1.000
FIELD ERROR RATE (%)
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (On-Time)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_0
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

D Summaries For Late Submissions

Stanley Janet and Jon Geist

This appendix contains summaries for all system results that were received late for the test but early enough to be scored for the Conference. All organizations submitting results on time except ERIM also submitted late results. Comparison of the on-time and late results shows that many late submission were significantly more accurate than the on-time submissions from the same organization. In some cases this reflected a continuation of planned improvements to the OCR system that were not finished in time for the on-time submissions. In other cases, it reflected the discovery of a bug in the OCR system upon examination of graphs of the on-time scores. These were circulated among the on-time participants following scoring of the on-time submissions to help the participants prepare for the meeting portion of the Conference.

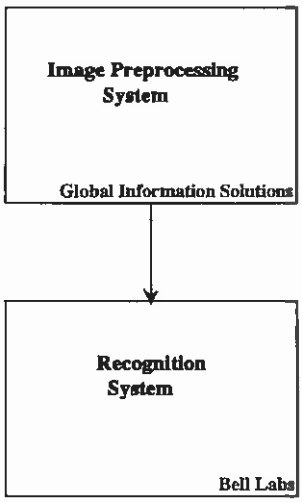
For instance, a number of the curves show regions of decreasing field error and distance rates followed by regions where these quantities increase with increasing rejection rate. This means that the confidences are not properly correlated with the actual OCR errors in some regions of rejection rate. This can happen at high rejection rates with a number of algorithms. However, in at least one case, a large effect of this nature was caused by the programming error described below. This description gives some insight into what causes the field error rate to increase instead of decrease with increasing rejection rate.

In the Second Conference tests, empty fields were to be classified with BLANK to distinguish them from fields that were left empty because the OCR was hopeless. (The latter was a good strategy for minimizing the field distance rate.) In its on-time submissions, CGK also assigned BLANK with a confidence of 1.0 to fields that were not even extracted. Since their confidence was 1.0, these fields were never rejected as the rejection rate was increased. At some rejection rate, the only fields left were the correct fields and incorrectly assigned BLANK fields. As the rejection rate was increased, correct fields with confidences less than 1.0 were rejected and the number of accepted (unrejected) fields decreased accordingly. On the other hand, no fields that had been incorrectly assigned BLANK were rejected, so the ratio of the fields with errors to the accepted fields increased with increasing rejection rate. A search by the CGK participant for the cause of the significant rise in the field error and distance rates in his on-time results uncovered this programming error, and made it clear that this was not the correct assignment for fields that were not extracted, only for fields that were empty. The CGK late submissions corrected this programming error, and as a result, were greatly improved over the on-time submissions at intermediate to high rejection rates.

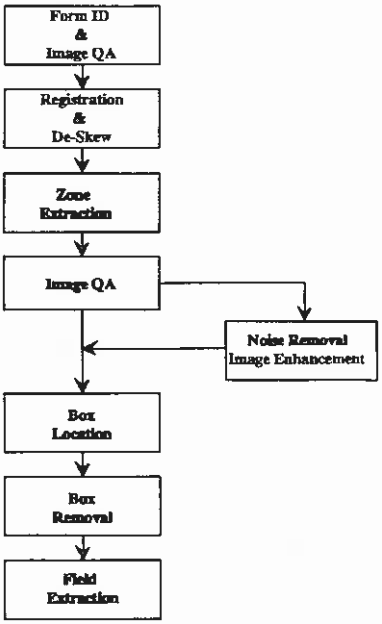
Graphs of field distance rate, field error rate, field distance rejection efficiency, and field error rejection efficiency are presented for the results obtained for the images scanned from paper. The same graphs are also presented for the results obtained from the images scanned from microfilm, if these were submitted. Note that the same system name will apply to results from microfilm and from paper when both were provided. The field distance and error rates are defined in Chapter 6, and the field distance and error rejection efficiencies are defined in Appendix C.

Two organizations, AT&T and MCC, submitted only late results for scoring. Viewgraphs describing their systems are presented here along with the graphs. The viewgraphs for the other systems are presented in Appendix C.

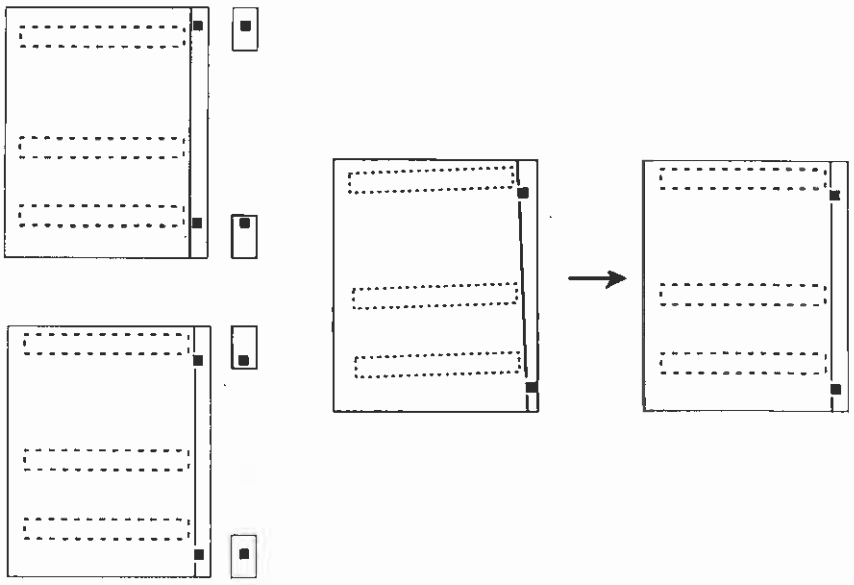
– AT&T Bell Labs & AT&T Global Information Solutions Colaborative Experiment



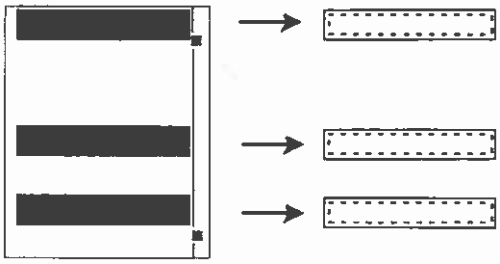
- Image Preprocessing System



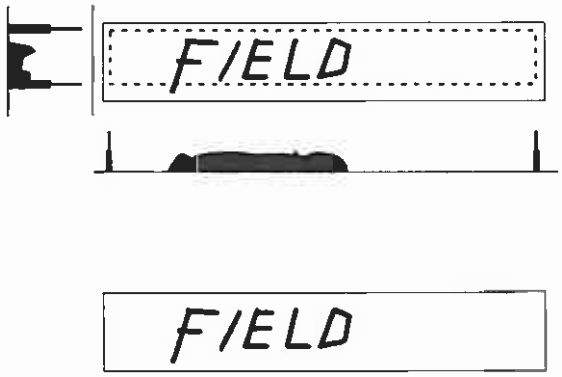
- Form ID, Image QA, Registration & Skew Correction



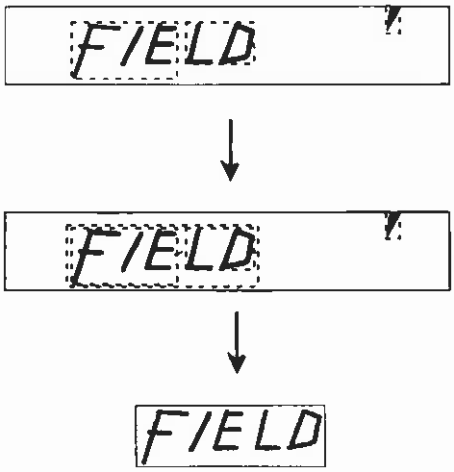
- Zone Extraction

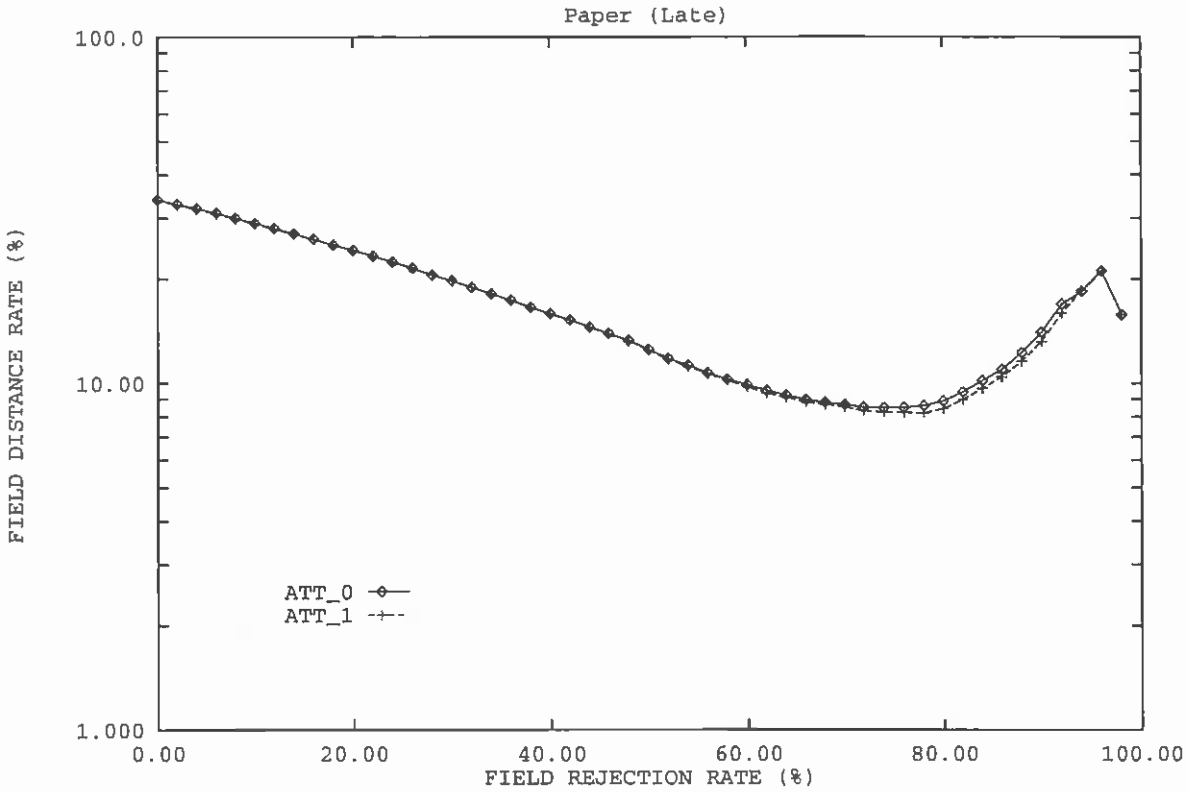


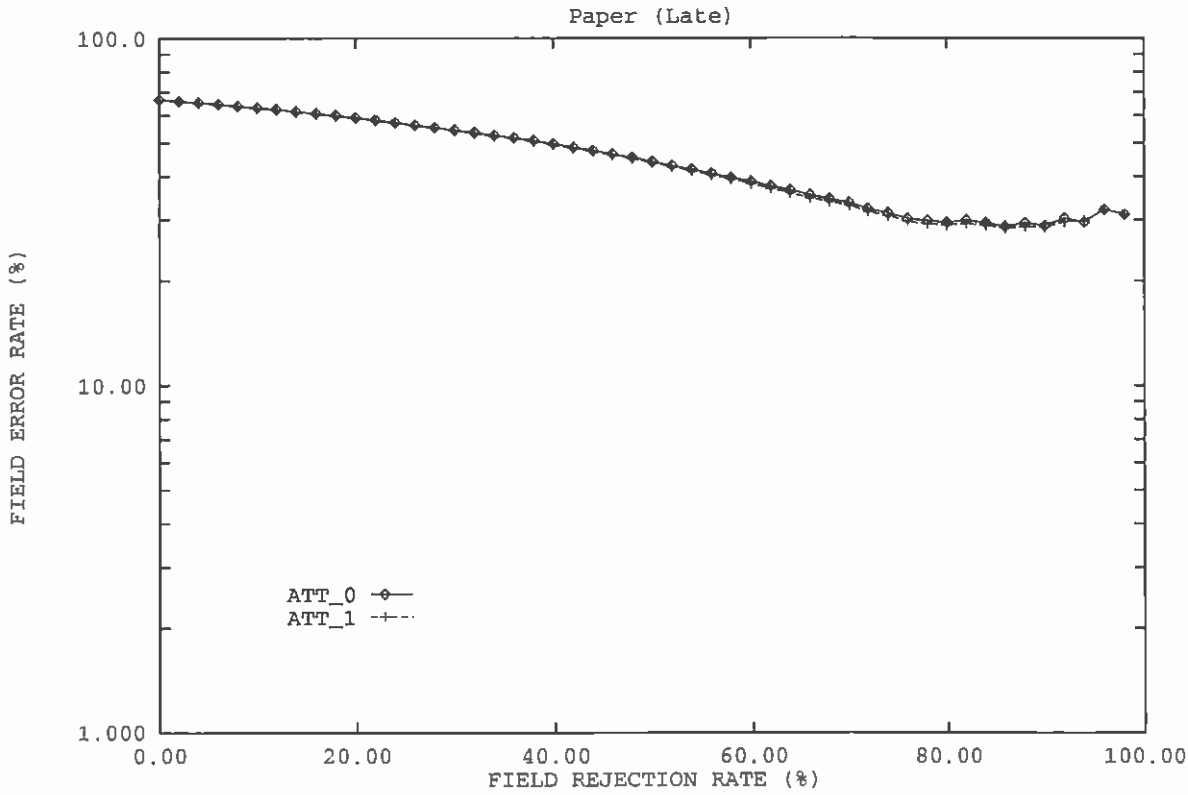
- Box Detection/Removal

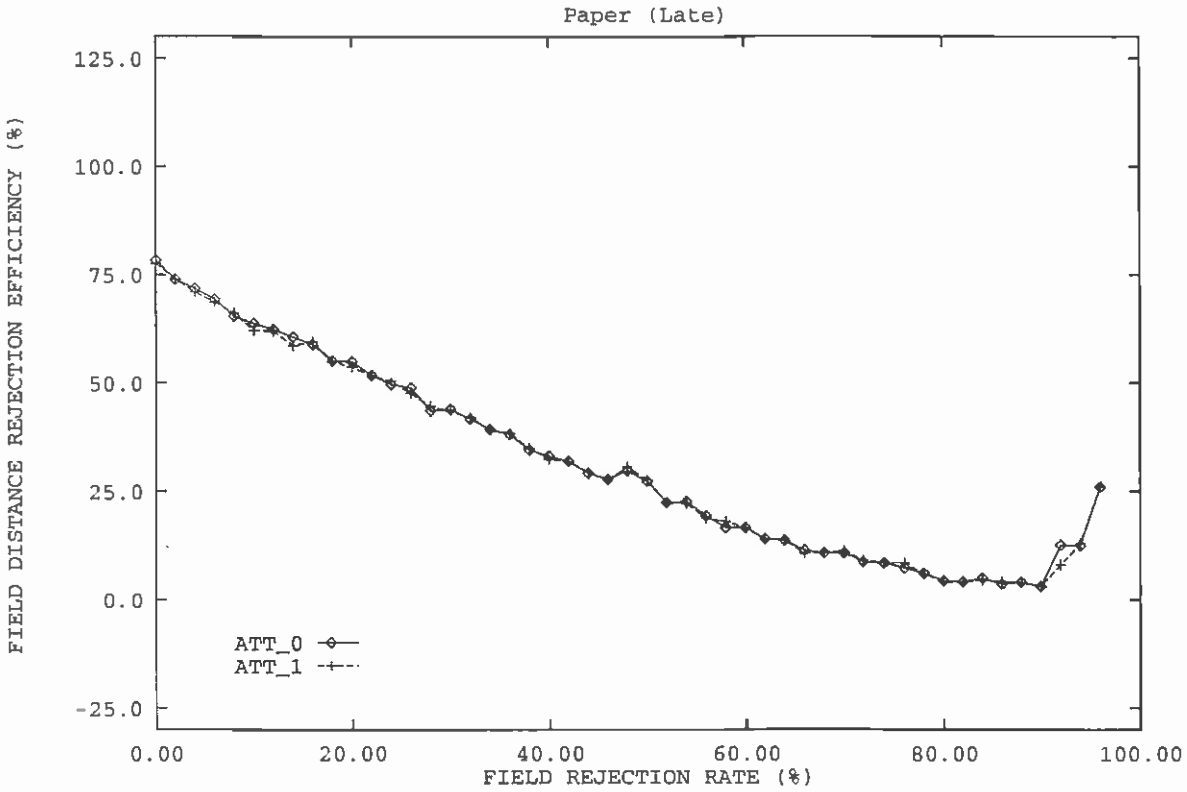


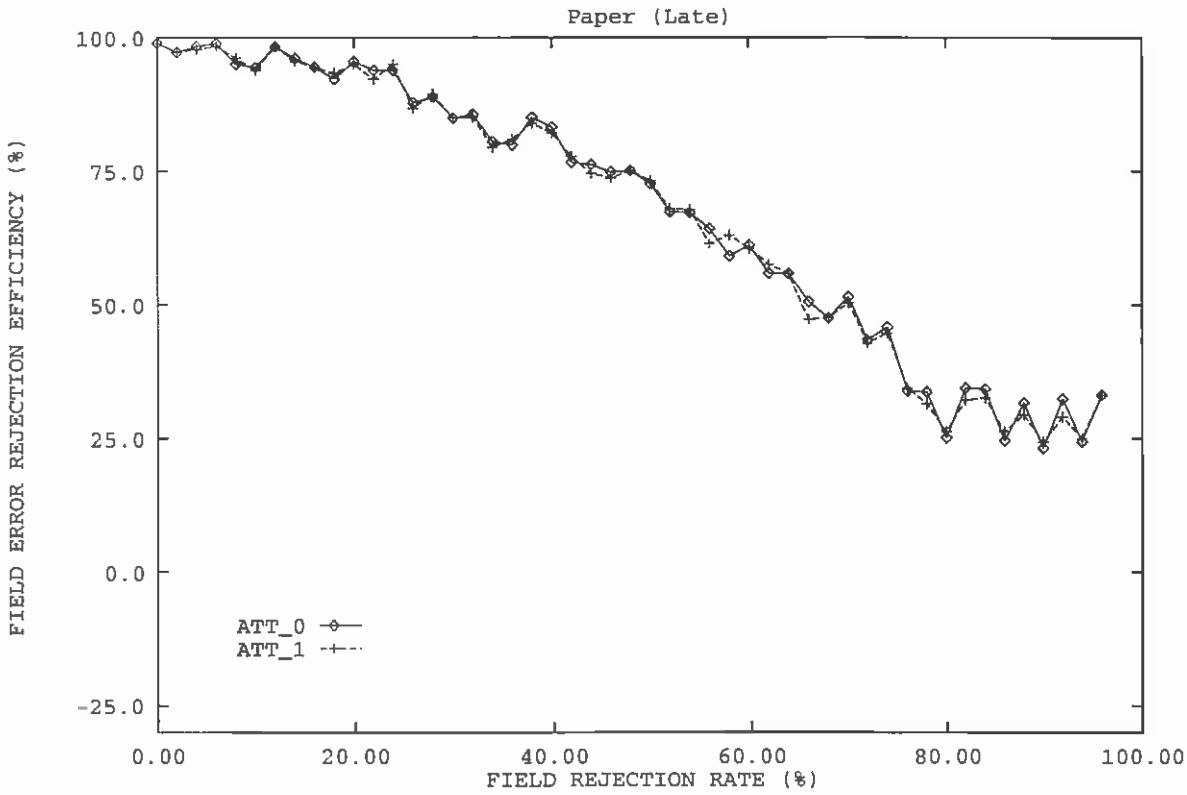
- Field Extraction




Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
ATT_0
ATT_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
ATT_0
ATT_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

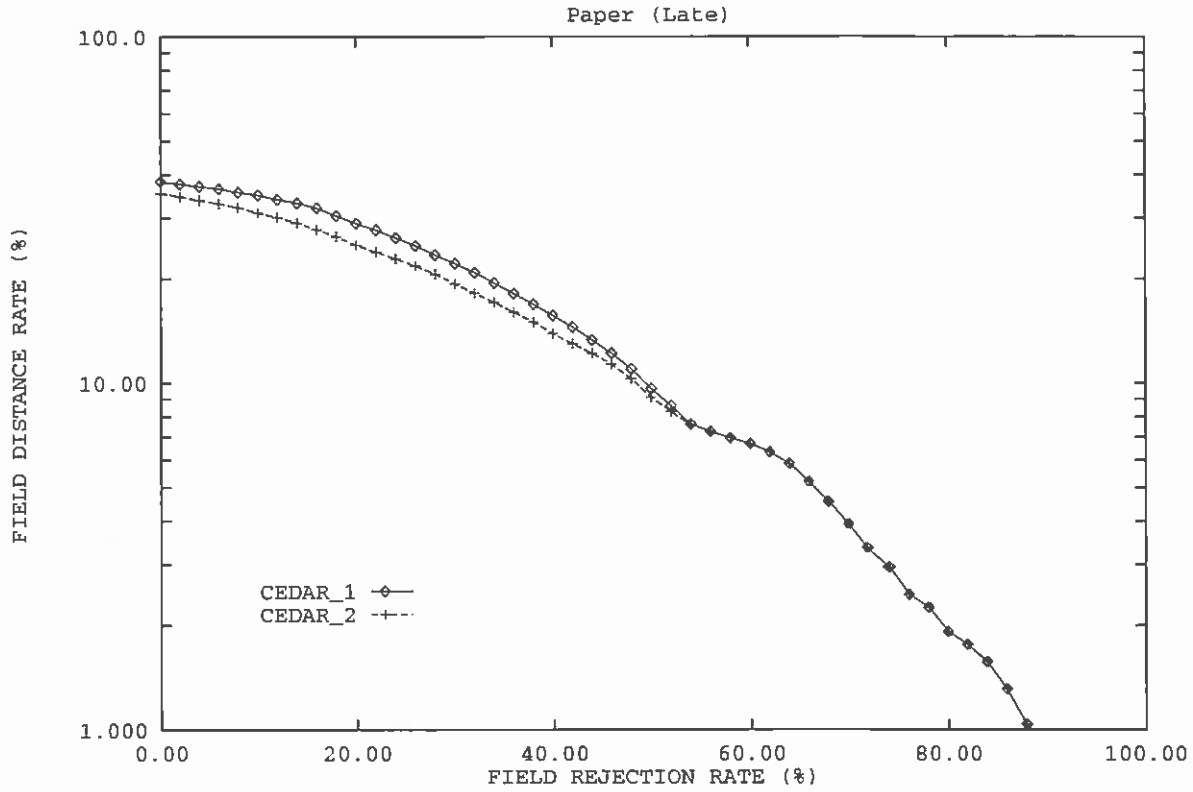

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
ATT_0
ATT_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

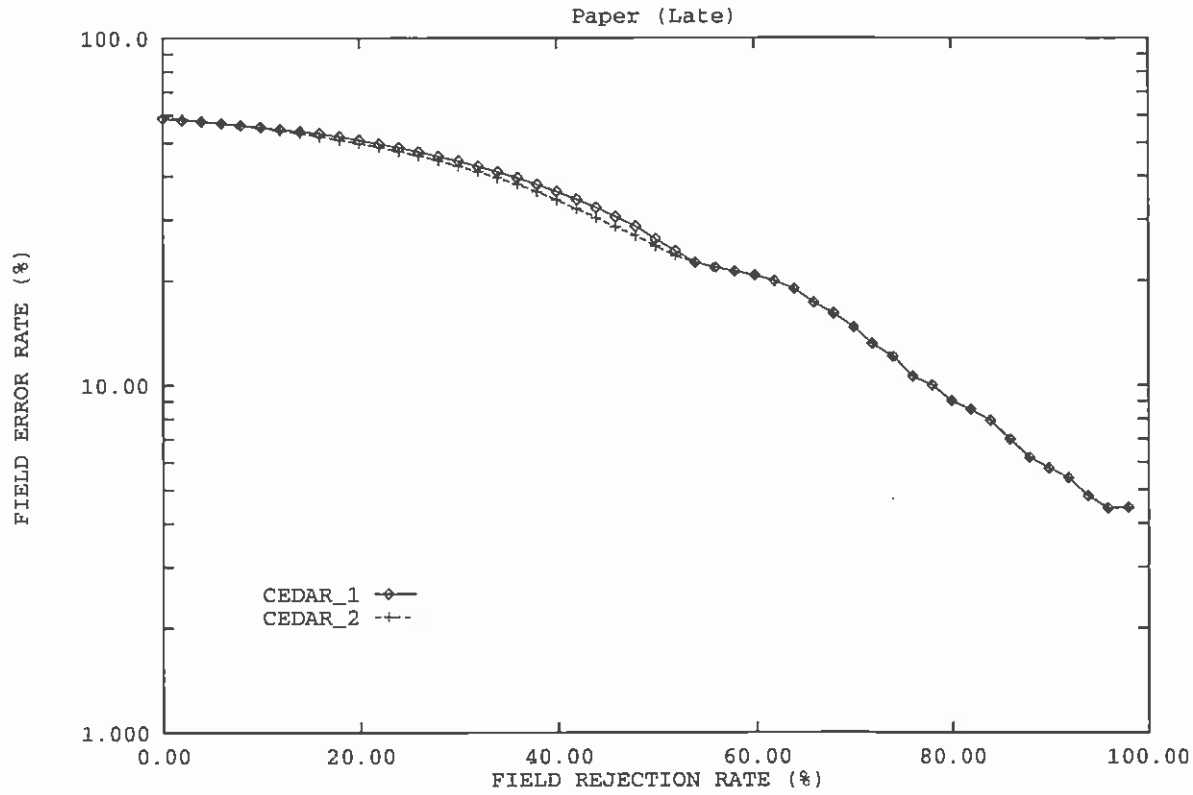

Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
ATT_0
ATT_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

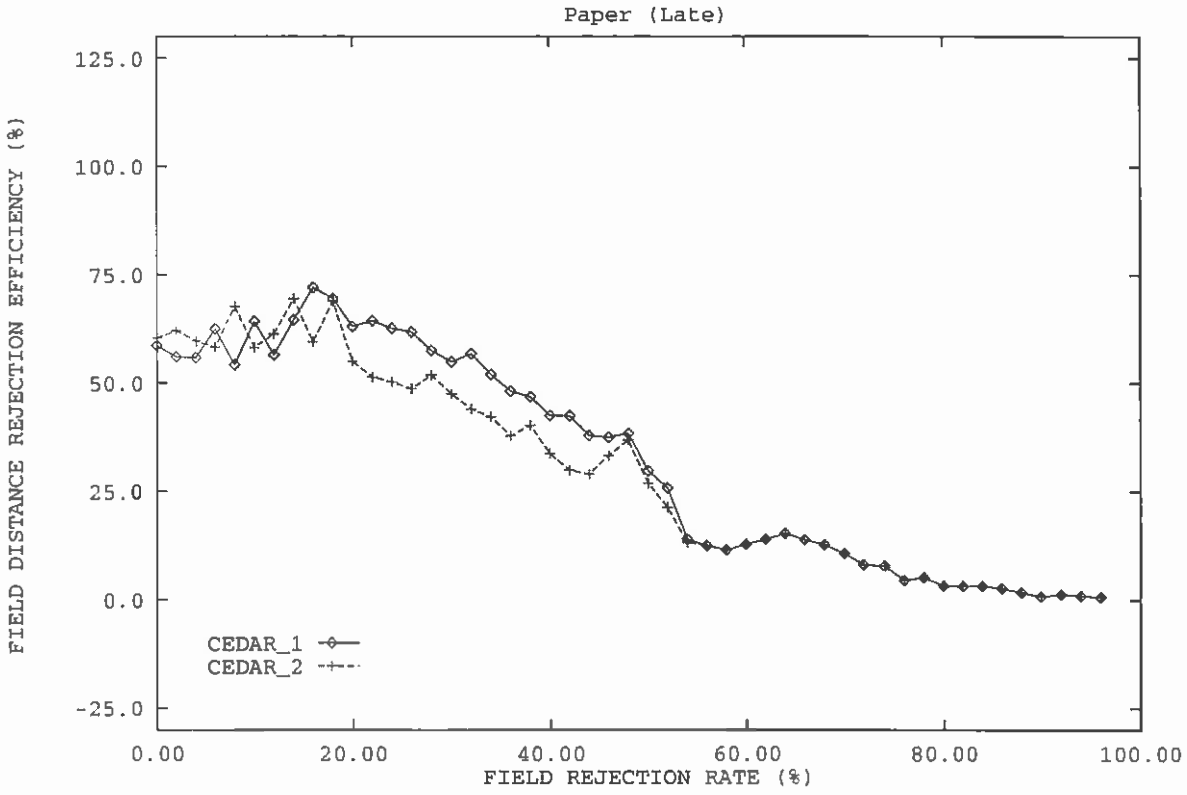
CEDAR

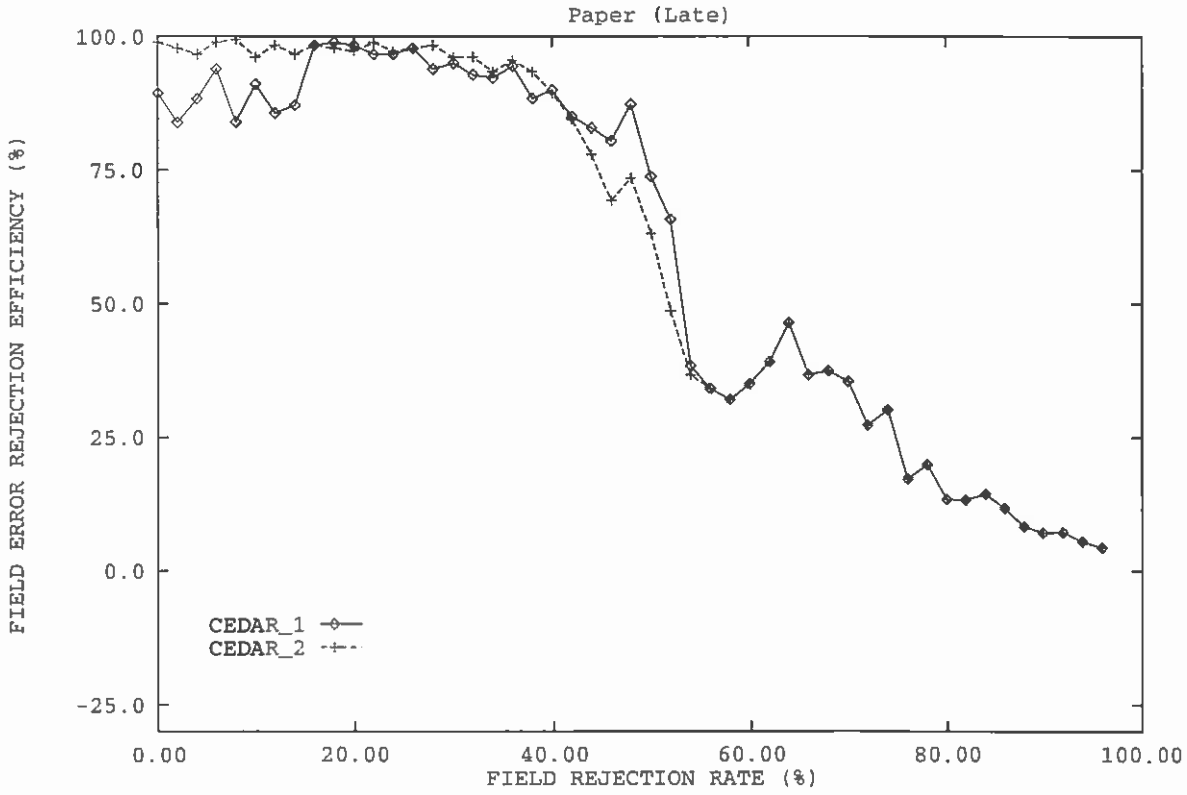
Late Submissions

See also: Summary for On-Time Submissions


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CEDAR_1
CEDAR_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CEDAR_1
CEDAR_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

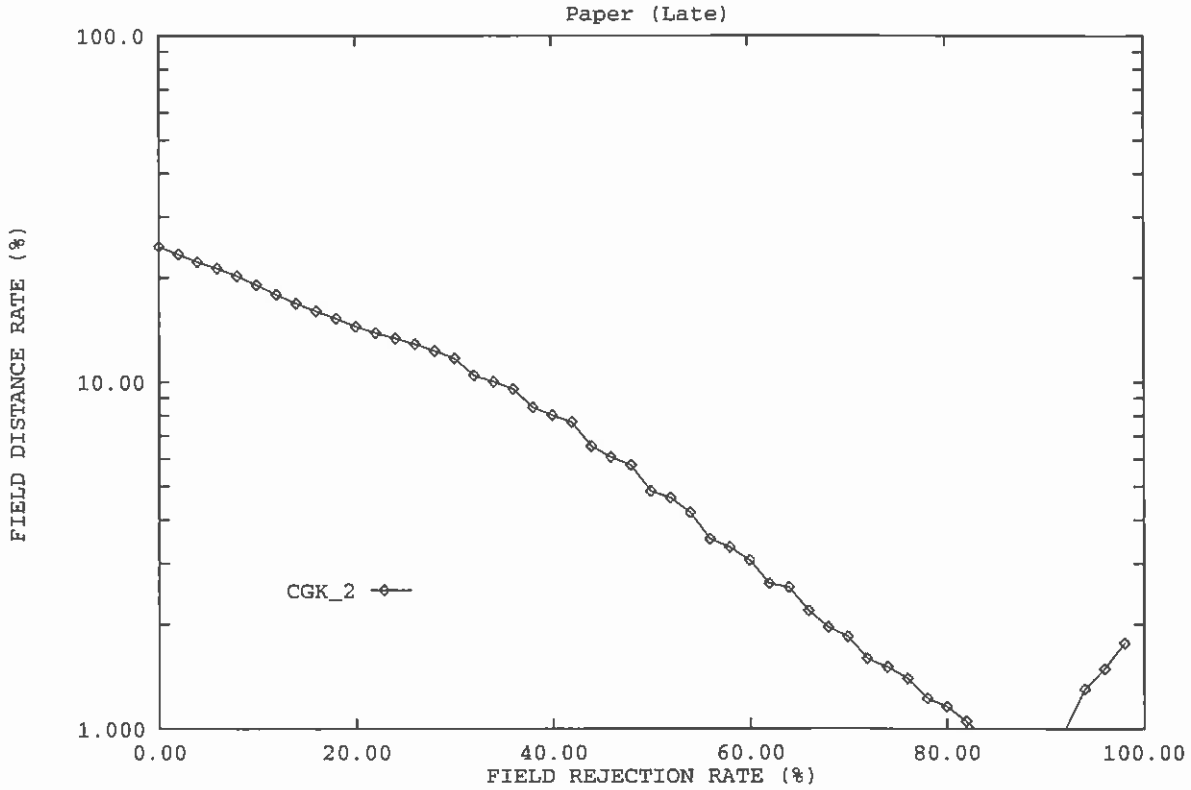

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CEDAR_1
CEDAR_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

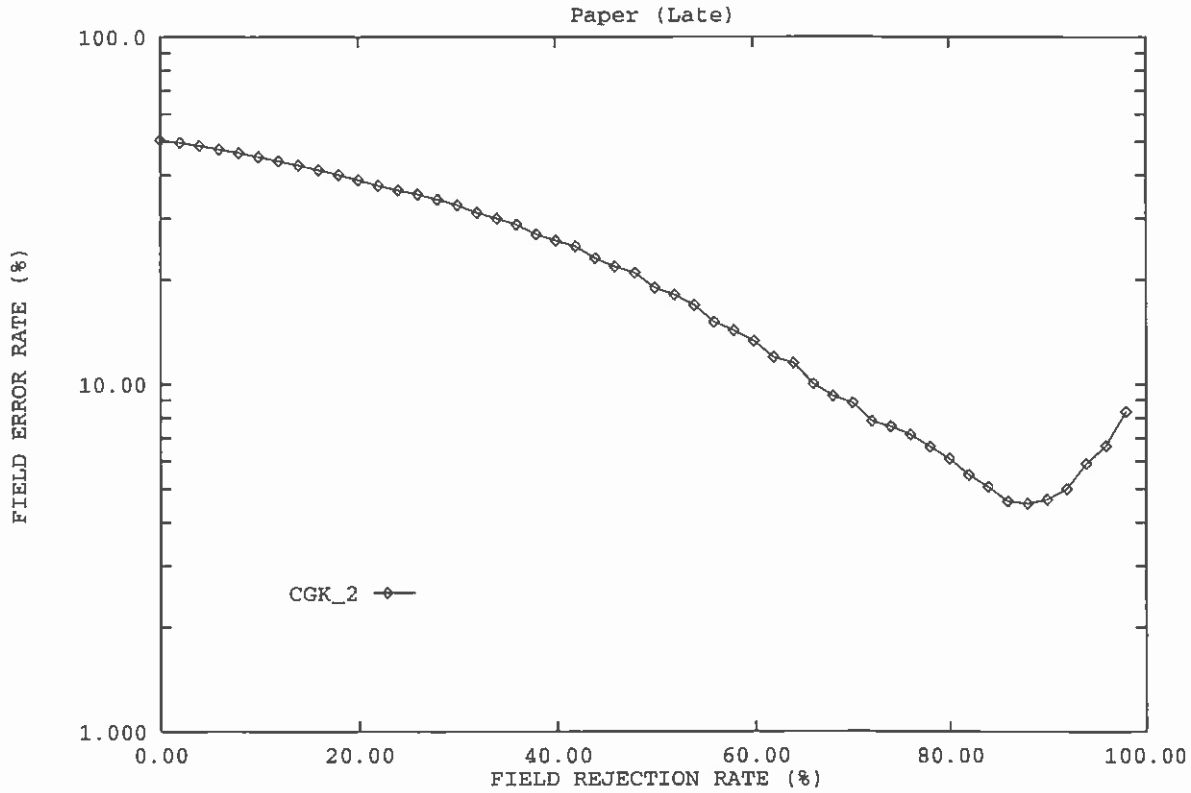

Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CEDAR_1
CEDAR_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

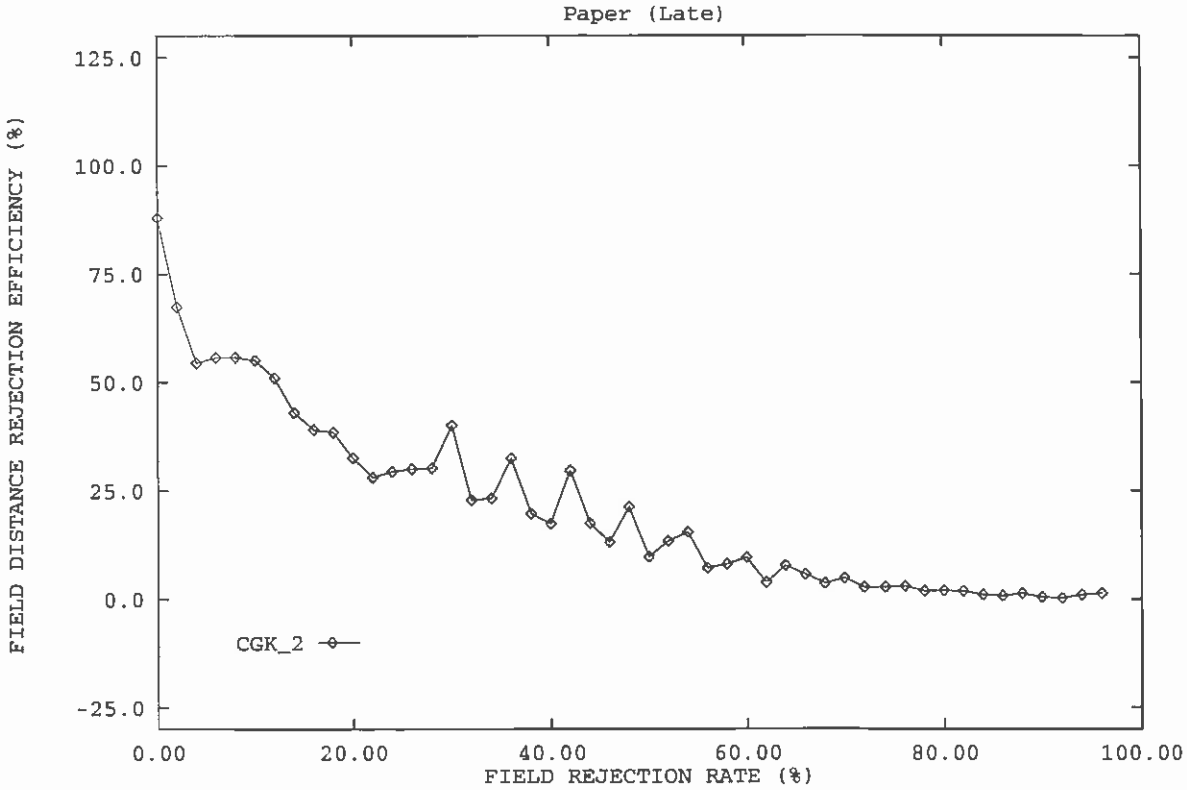
CGK

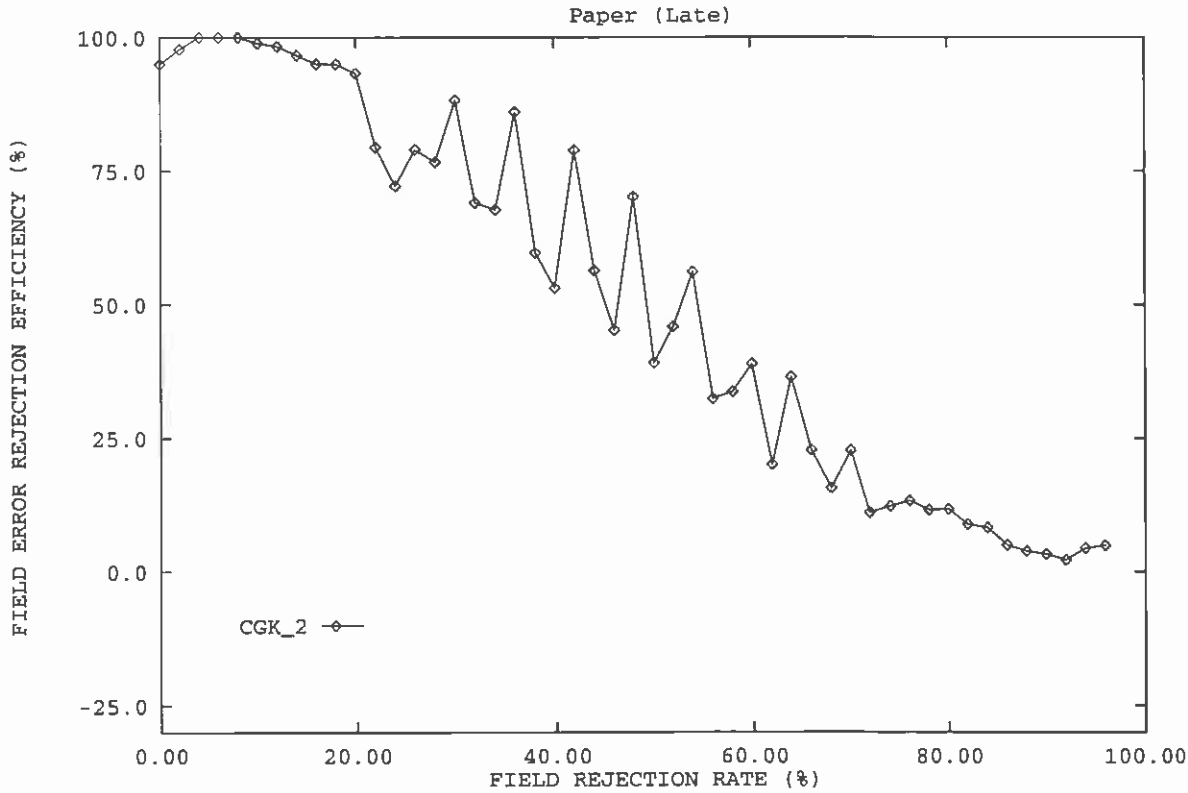
Late Submissions

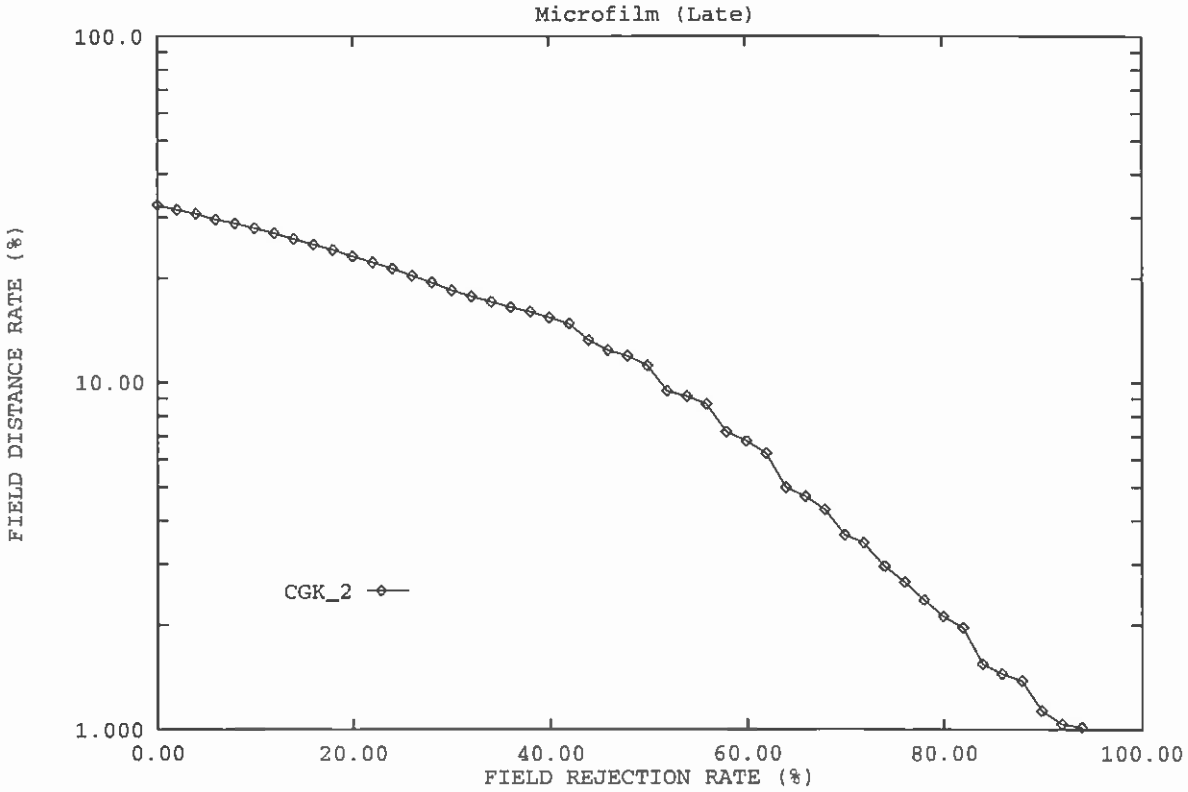
See also: Summary for On-Time Submissions

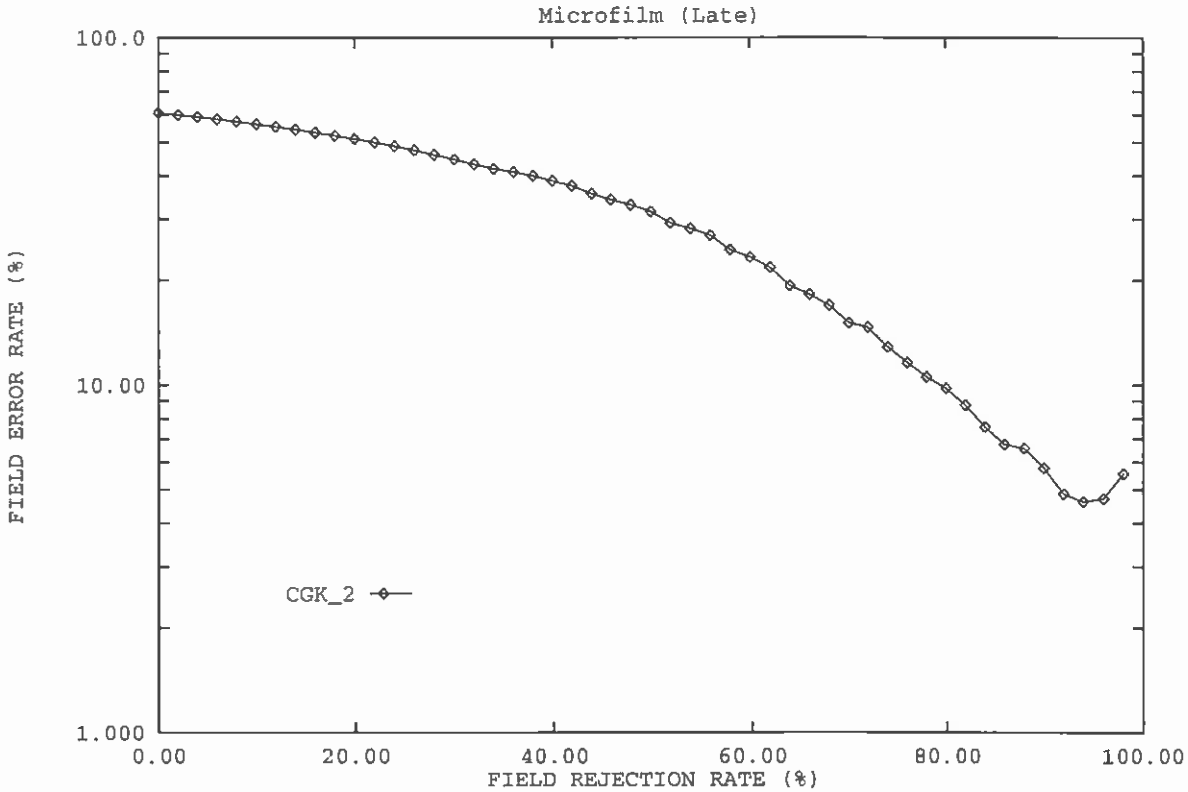

Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

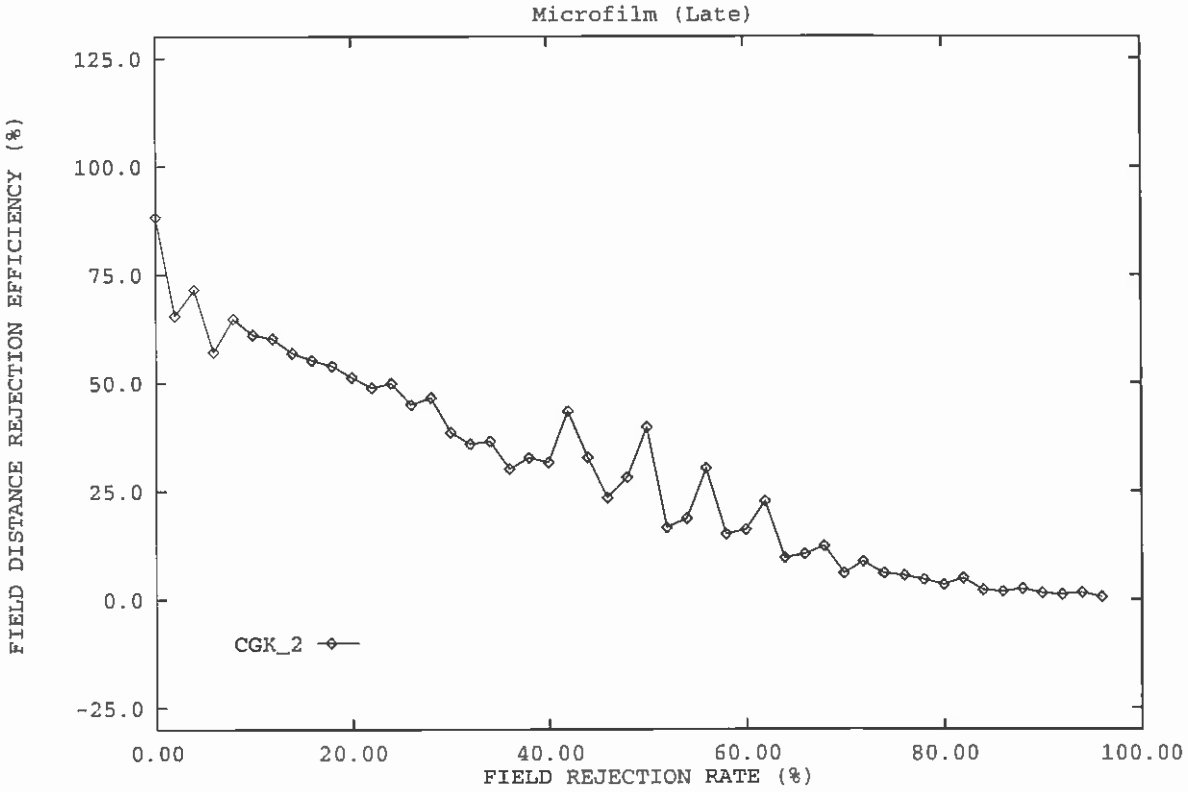

Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

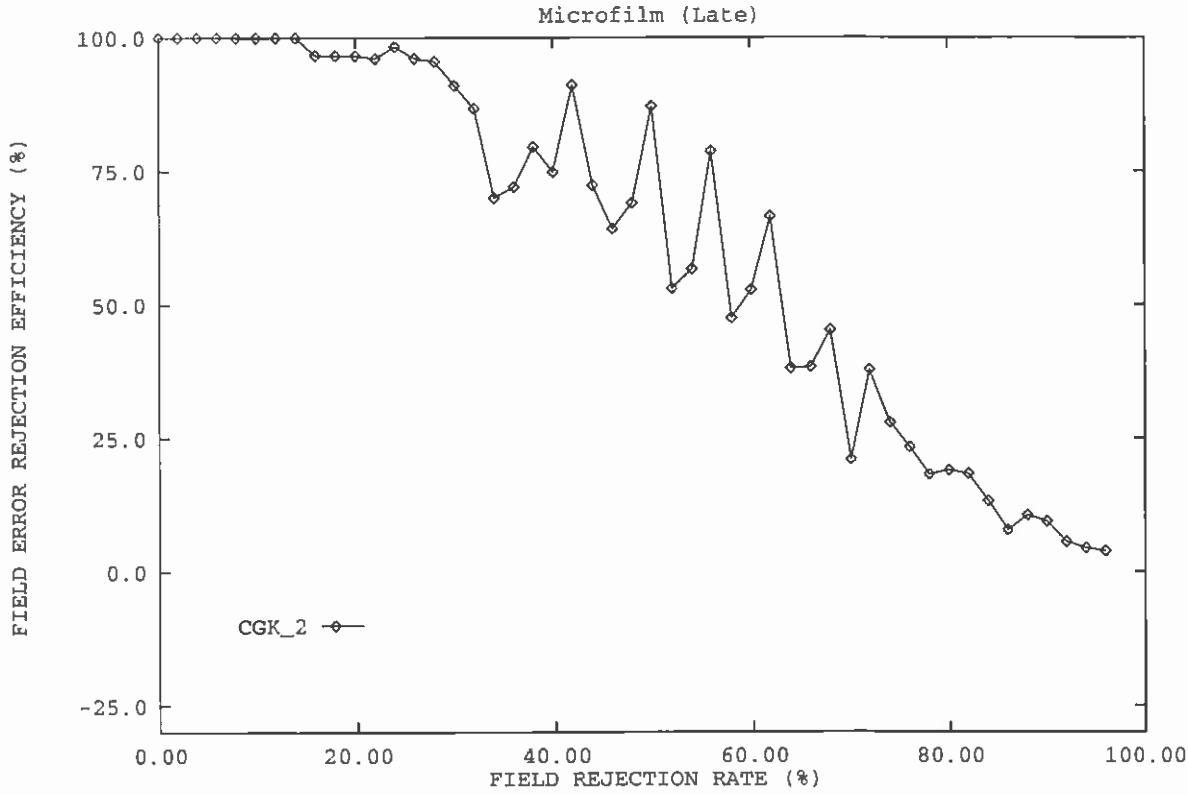

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

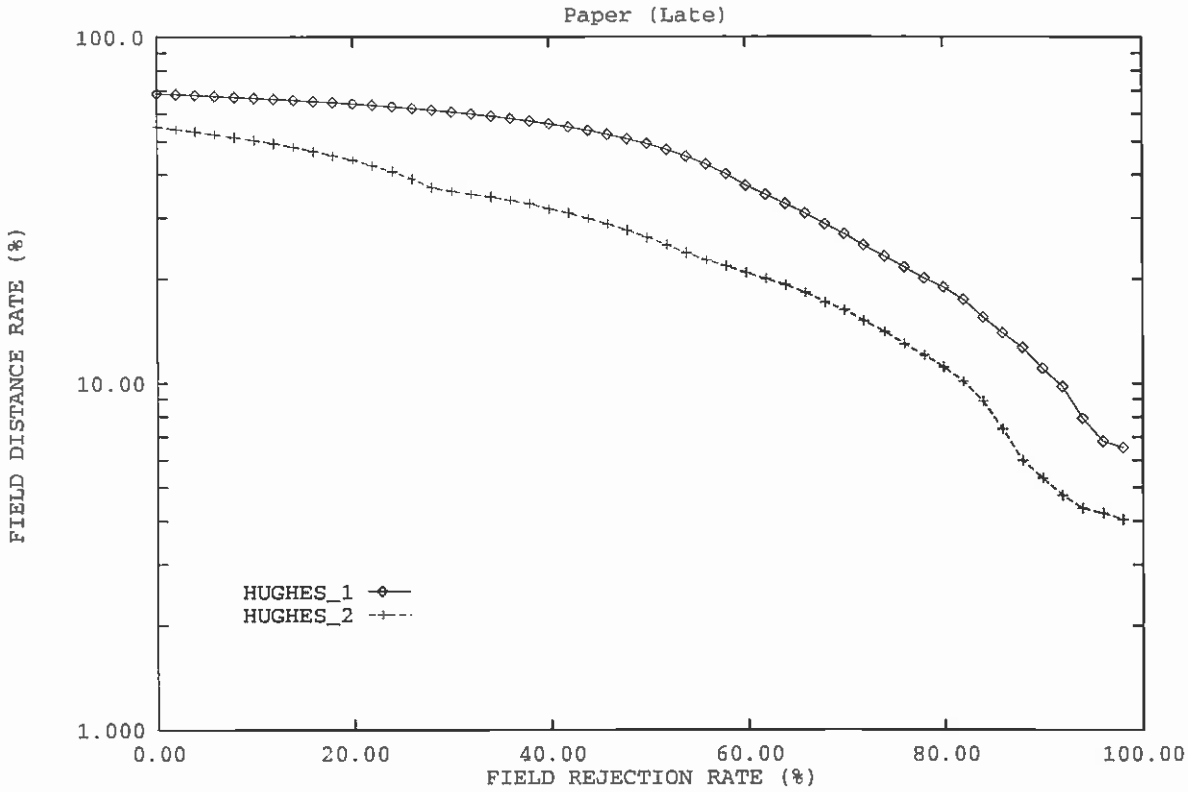

Microfilm (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

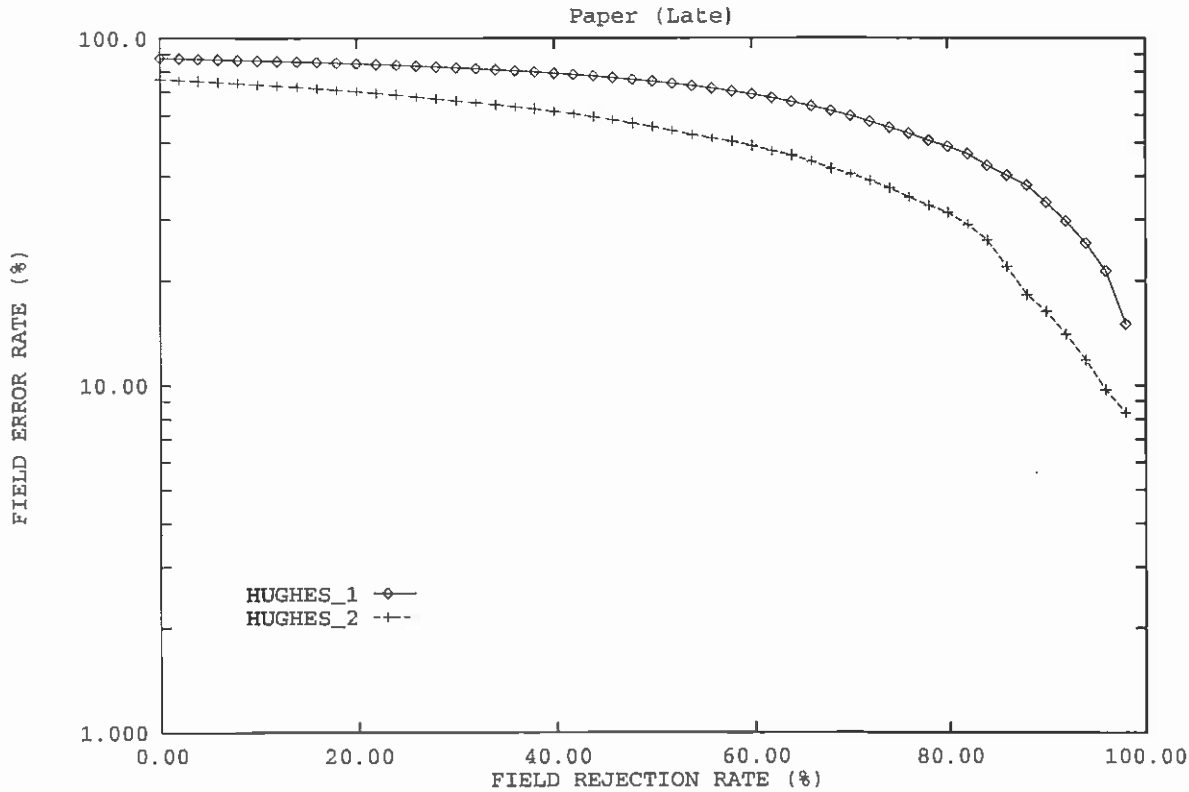

Microfilm (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
CGK_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

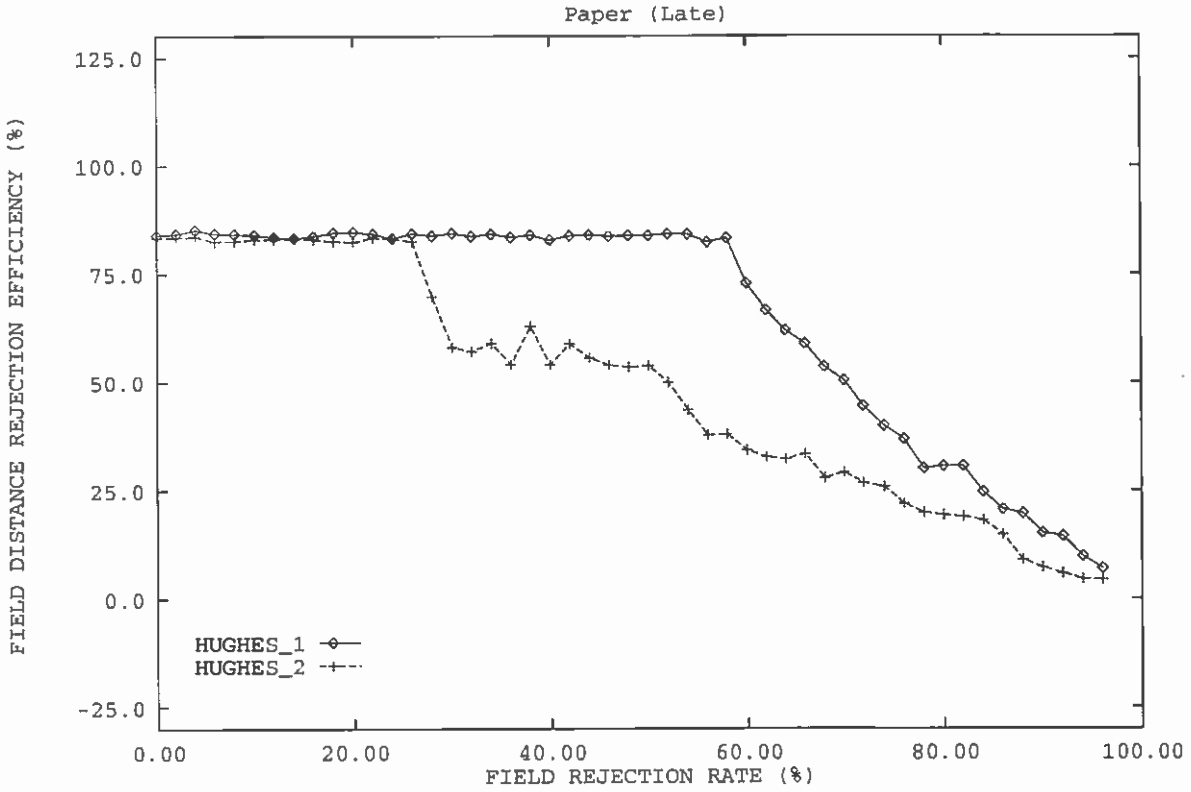
HUGHES

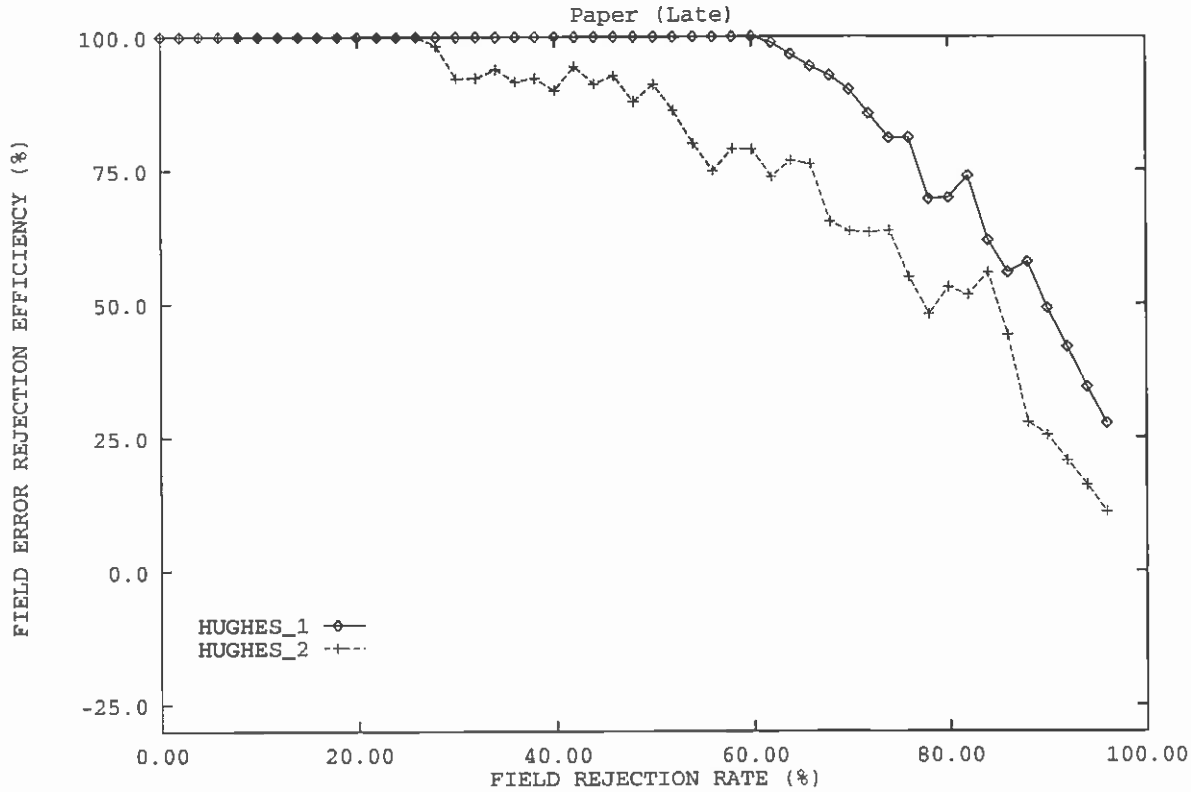
Late Submissions

See also: Summary for On-Time Submissions


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
HUGHES_1
HUGHES_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
HUGHES_1
HUGHES_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

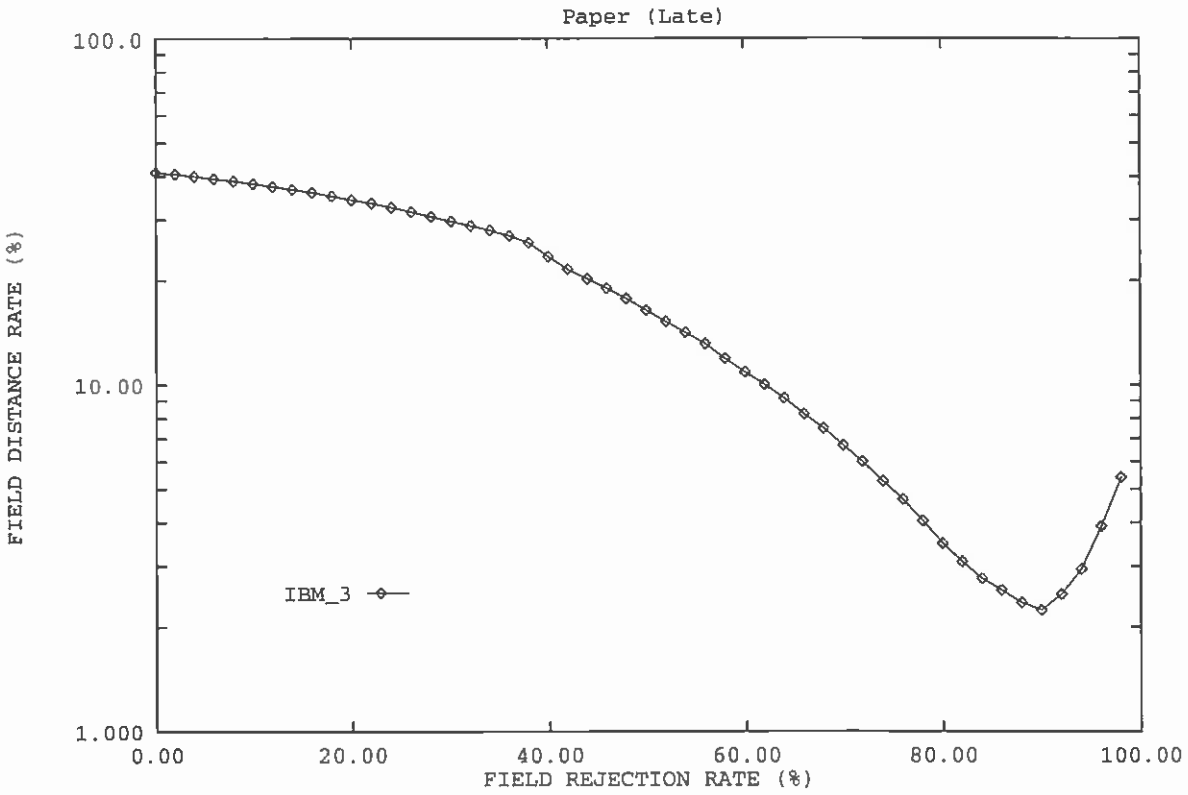

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
HUGHES_1
HUGHES_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

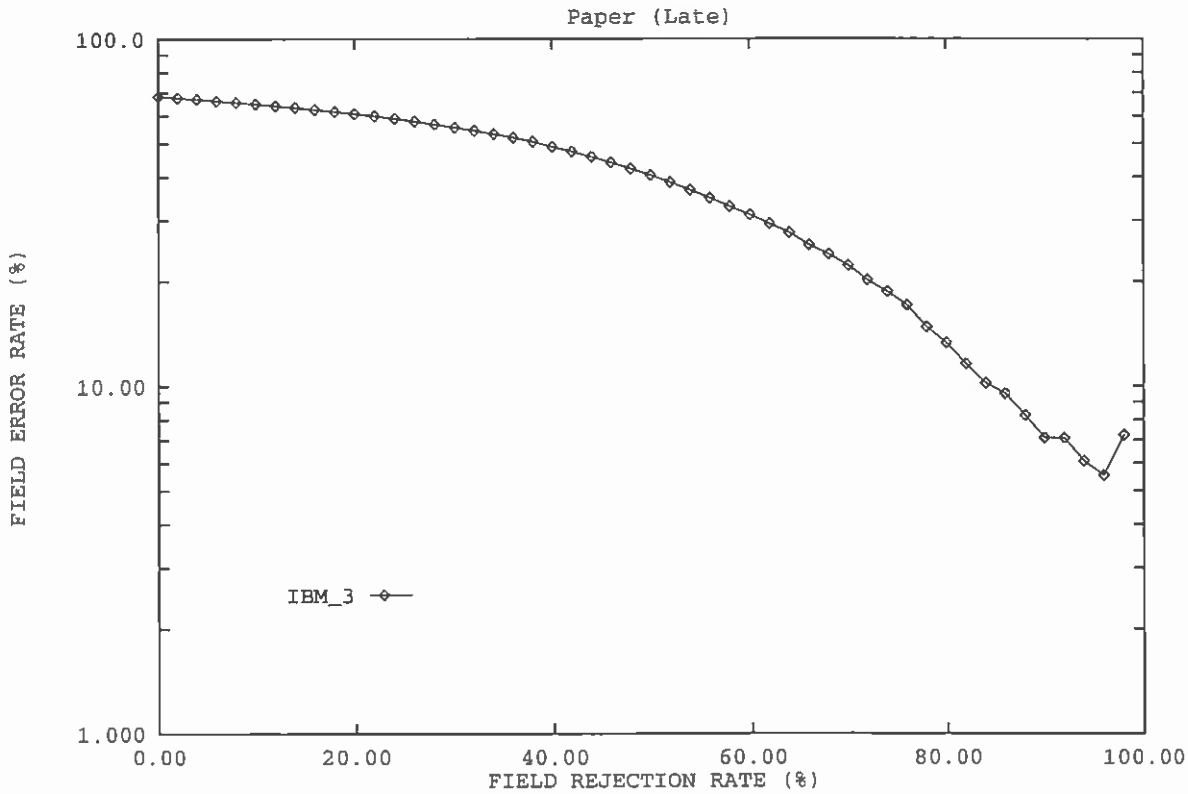

Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
HUGHES_1
HUGHES_2
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

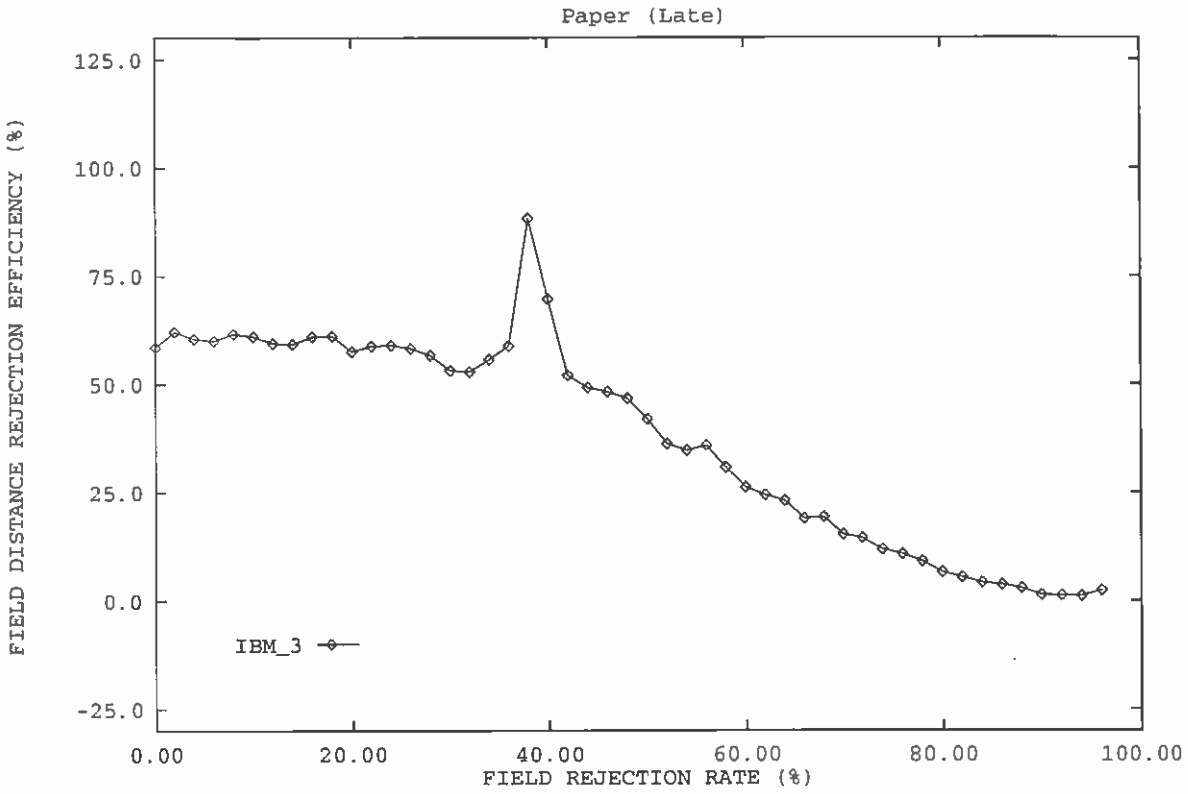
IBM

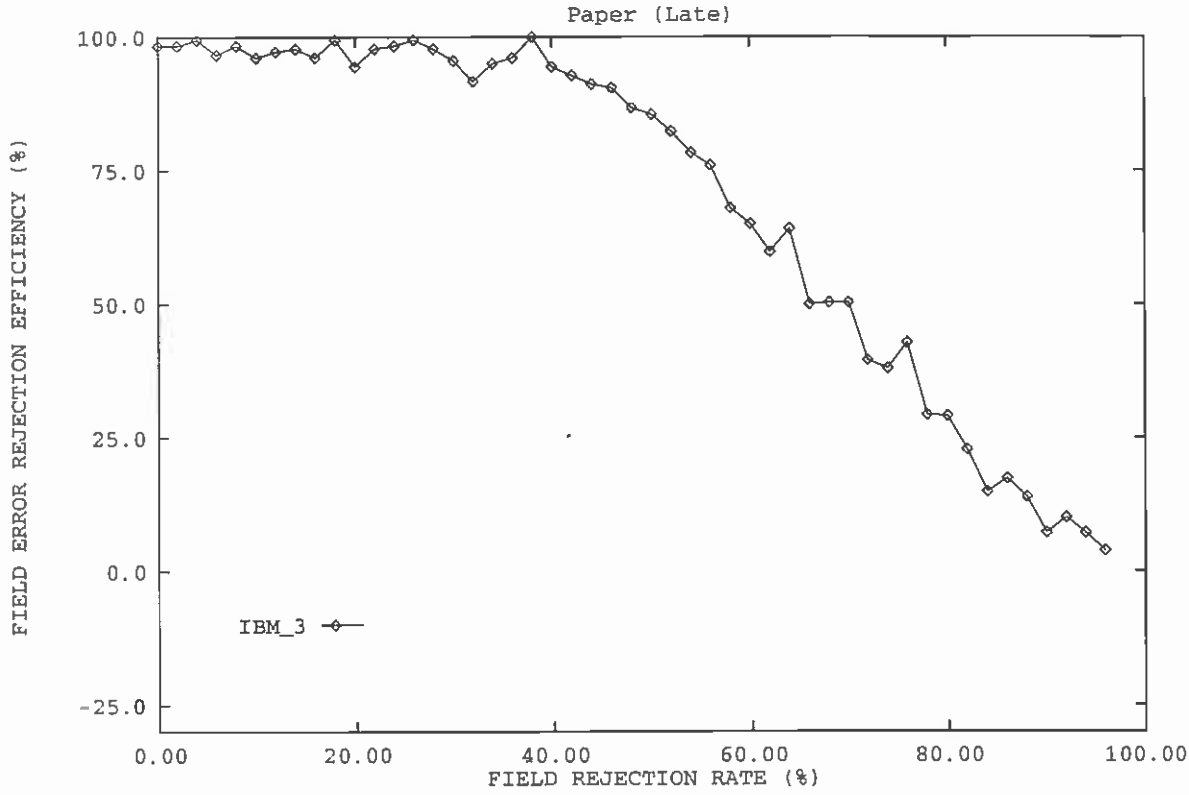
Late Submissions

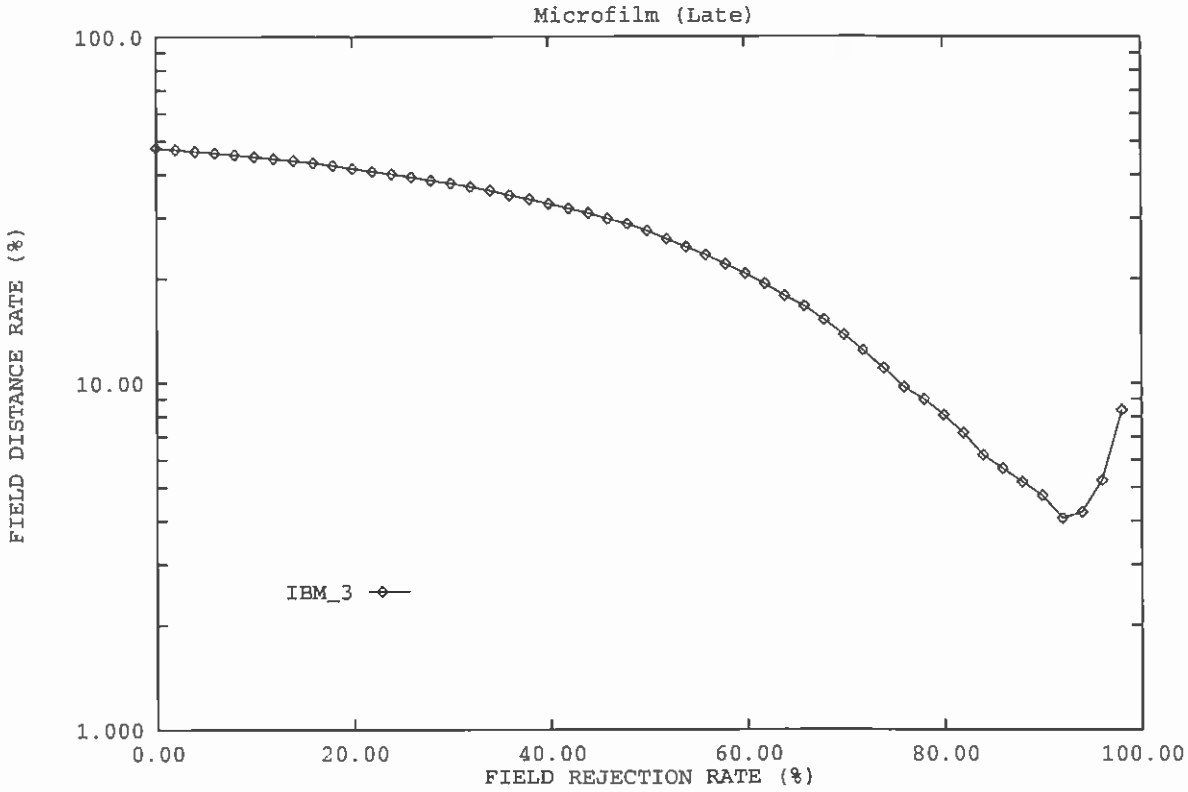
See also: Summary for On-Time Submissions

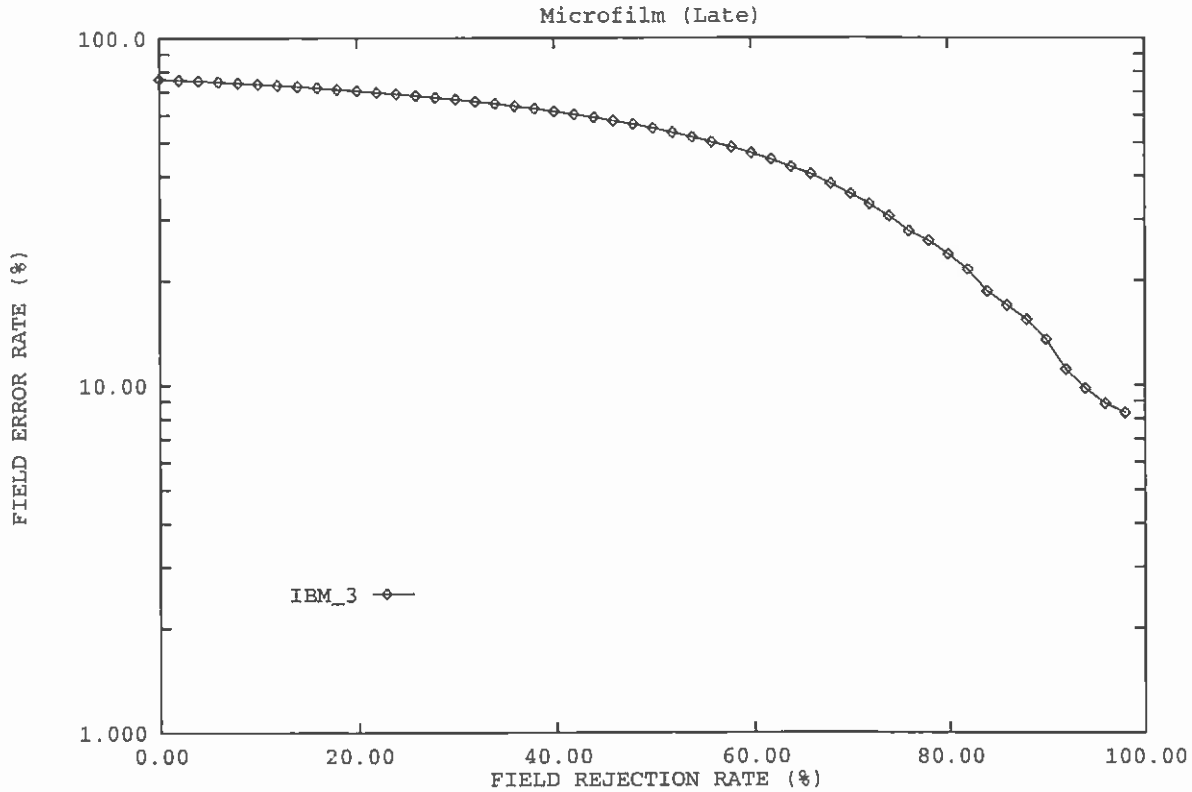

Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

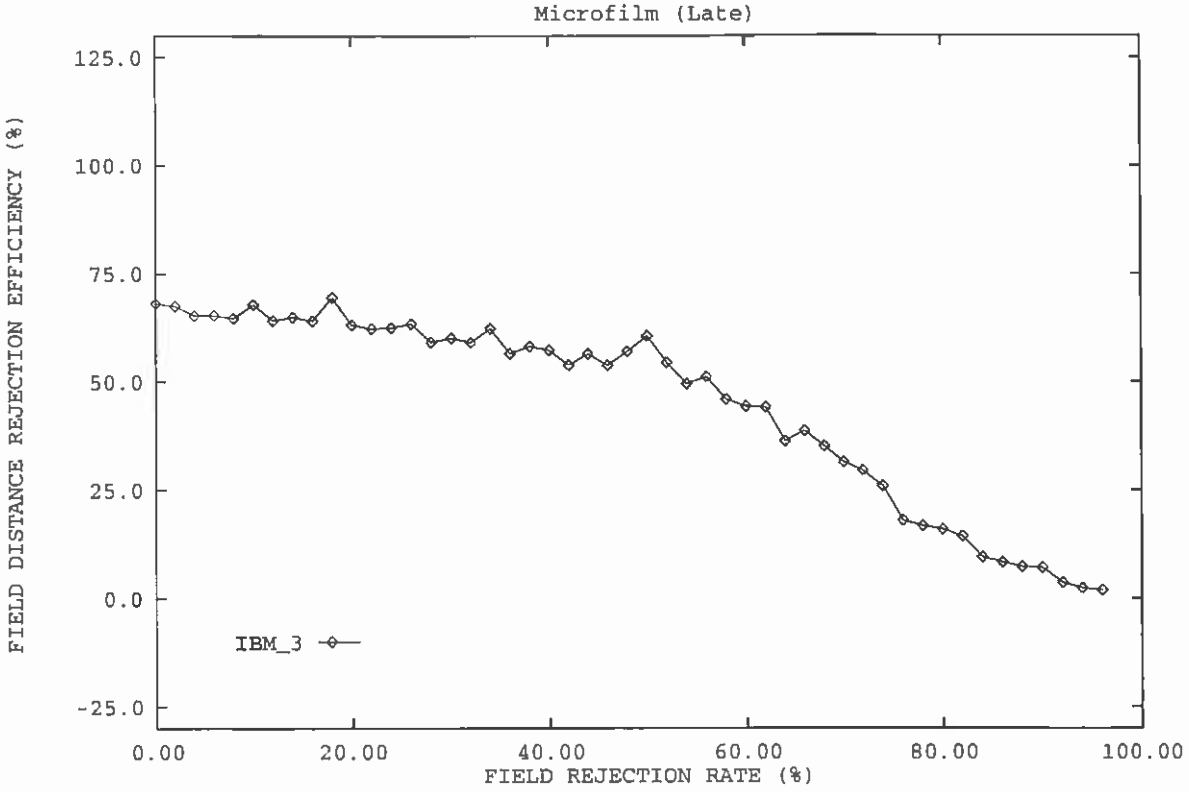

Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

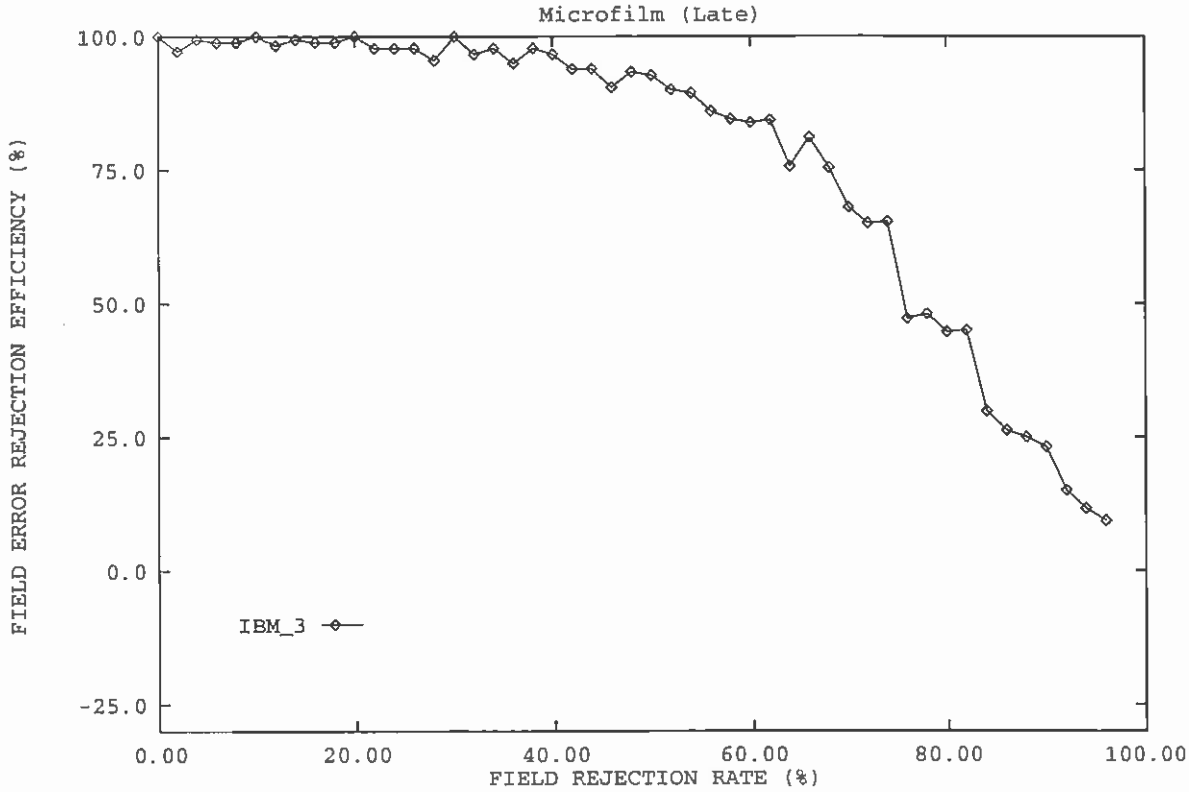

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Microfilm (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

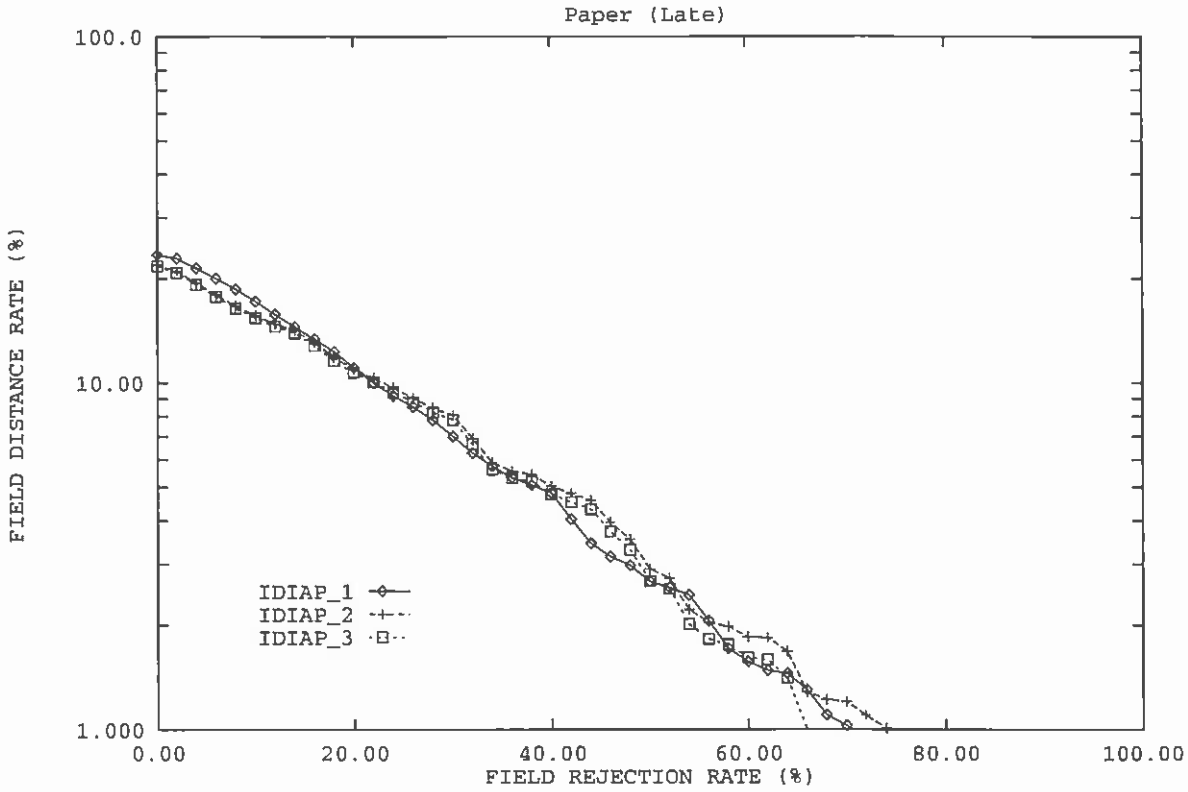

Microfilm (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

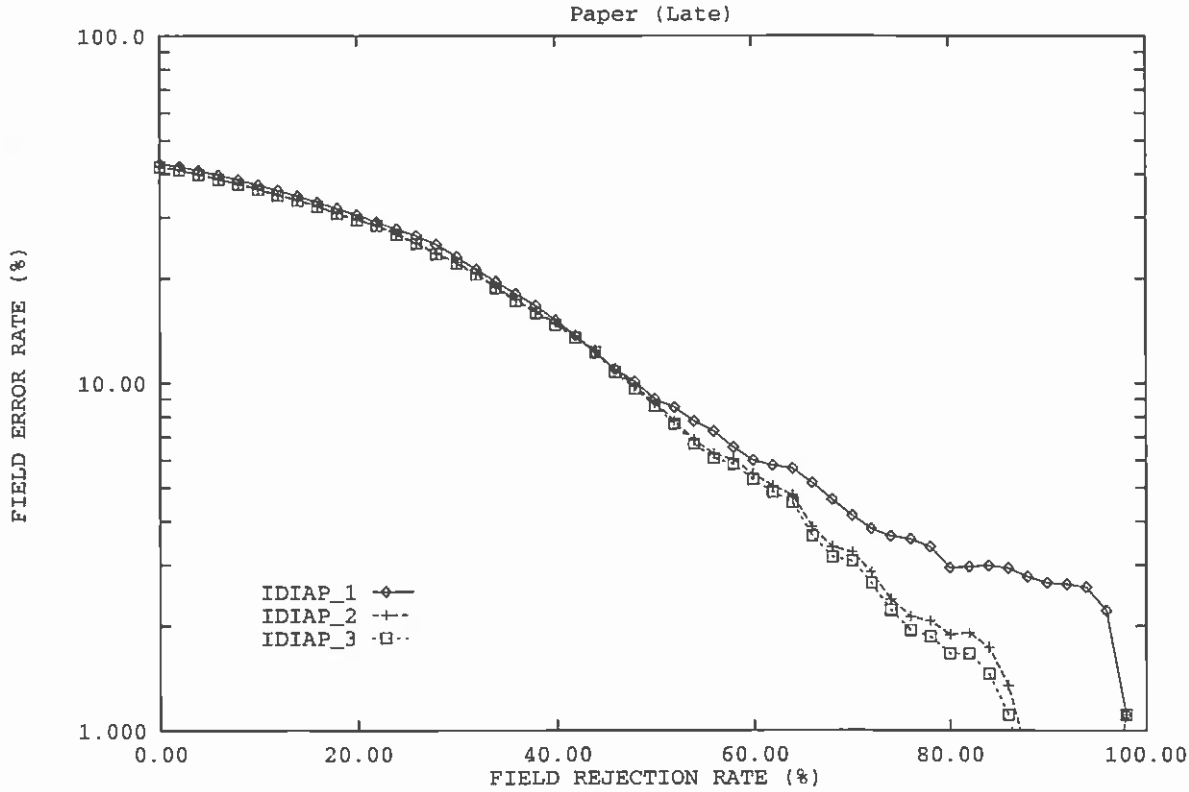

Microfilm (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
IBM_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

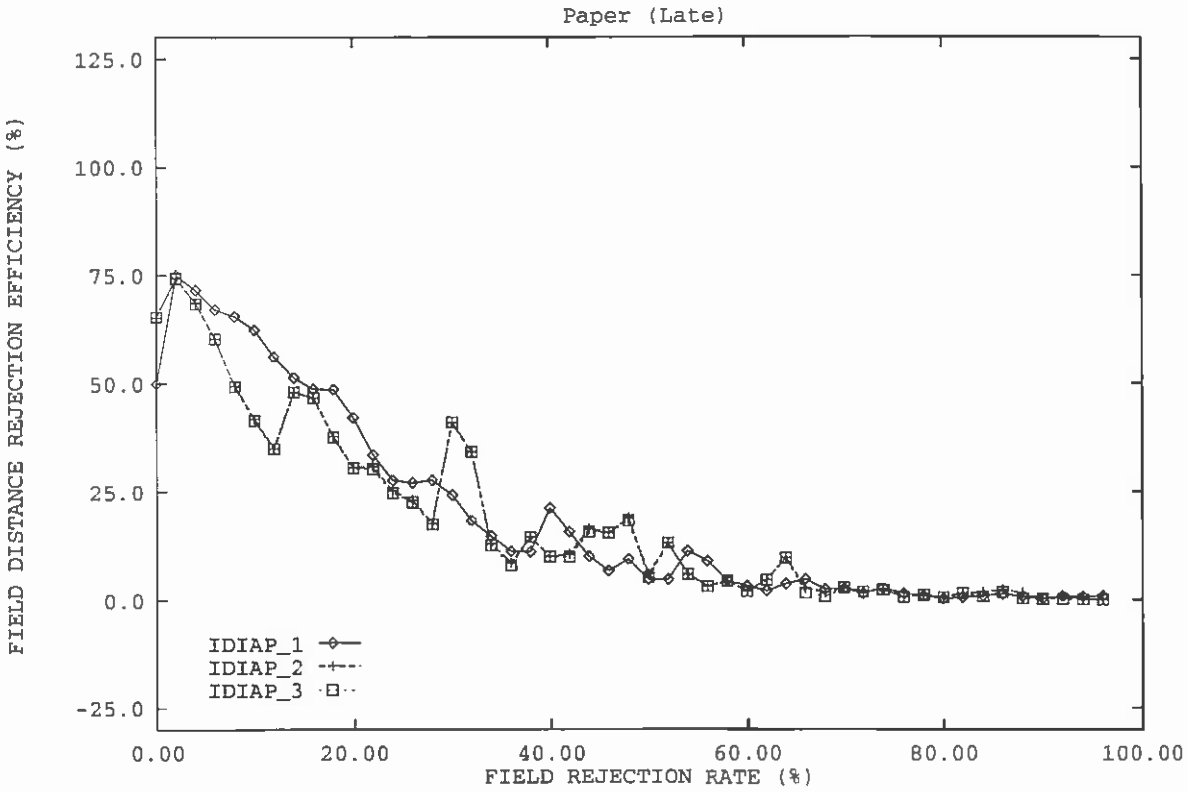
IDIAP

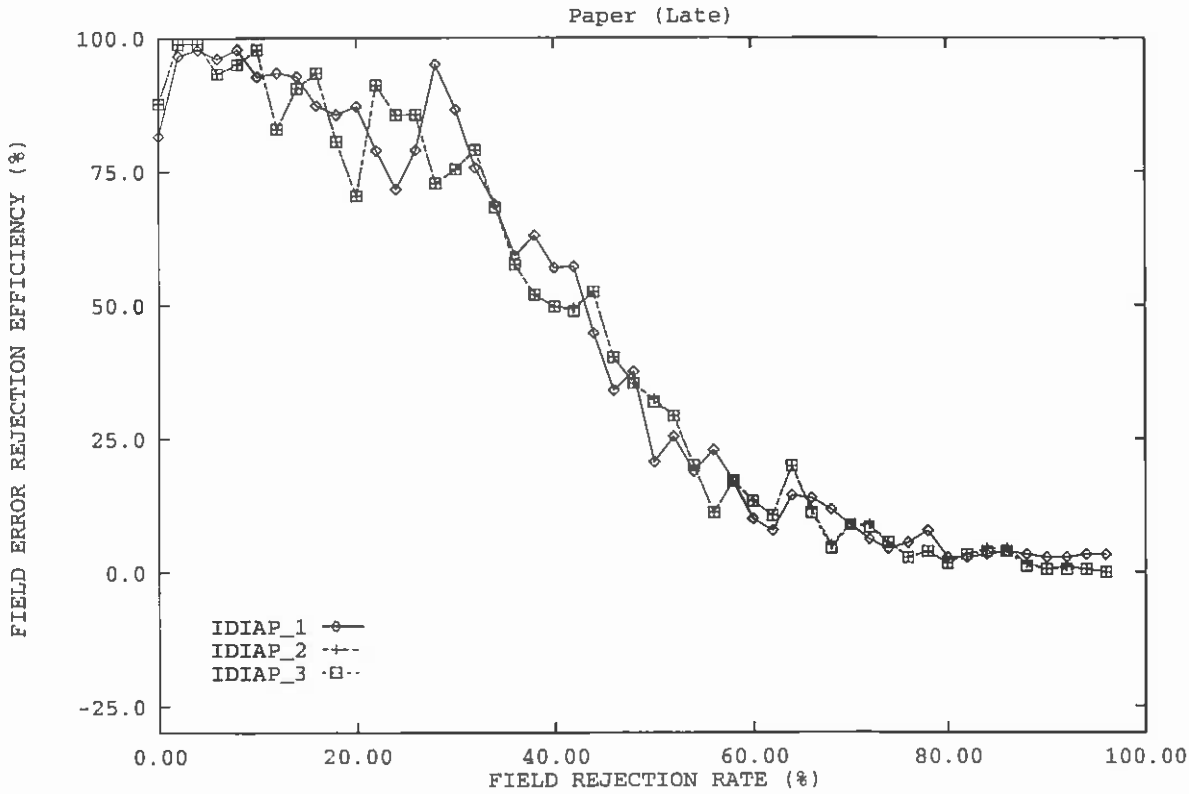
Late Submissions

See also: Summary for On-Time Submissions


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
IDIAP_1
IDIAP_2
IDIAP_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
IDIAP_1
IDIAP_2
IDIAP_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
IDIAP_1
IDIAP_2
IDIAP_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
IDIAP_1
IDIAP_2
IDIAP_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

MCC Experiments

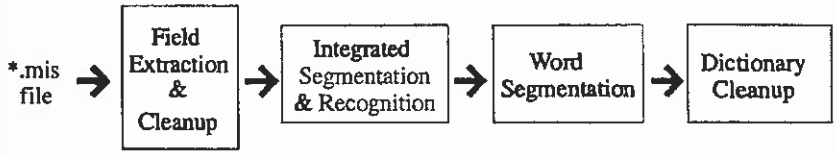
Gale Martin

Comments on the MCC Entries

The scoring for the MCC entry was considerably below what we were achieving on the test on the test set that we had specified from the original training materials (66% field error rate as compared to the 86% reported at the Conference). After the Conference, an analysis of the source of the drop led to the discovery that we had a bug in the field extraction code. The bug was introduced when we made relatively minor changes to enable the field extraction code to operate in the test mode, without the presence of reference files. The bug had the effect of capturing completely wrong portions of the mini-forms, and because it was not present on our training and internal testing operations, we didn't discover it prior to the NIST testing. When we corrected the bug, and reran the prior system, which was unchanged otherwise, on the Conference test files and scored it against the reference files subsequently sent out by NIST, we achieved considerably better results(i.e., 71% field error rate). Since that time, we have also been able to improve performance further, through some relatively minor changes in the neural network architecture, and in the integration of word-segmentation and word-level dictionary lookup. Our current field-level error rate stands at 58% on the Conference test set.

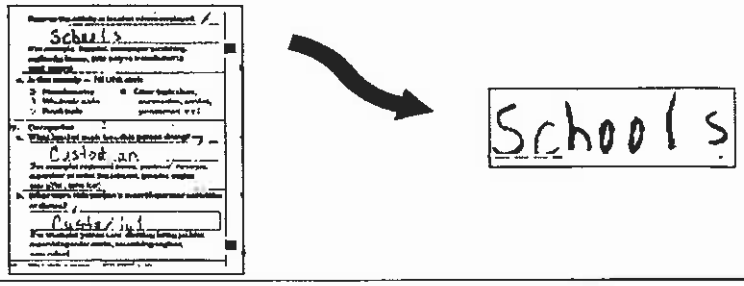
We are grateful to NIST and the U.S. Census Bureau for giving us the opportunity to test and extend our technology to handwritten phrases.

Processing Pipeline



Field Extraction & Cleanup

- Locate dashed line box around field and de–skew
- Include multiple lines below field for descenders
- Remove image noise
- Remove blank lines and columns surrounding text

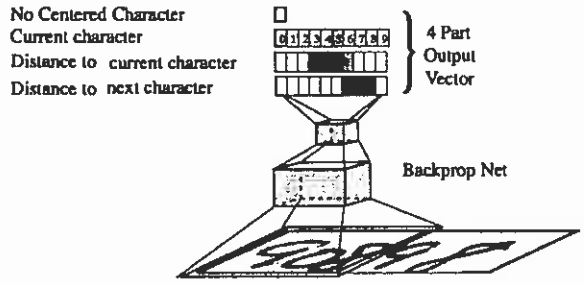


Integrated Segmentation & Recognition

- Saccade System
- Multi–Classifier Experiments

Saccade System

- Trained to navigate a field of text—
  center input window on characters,
  and then recognize them



- Developed for digit recognition

Multi–Classifier Experiments

- Boosting Algorithm (Drucker, Schapire & Simard, ATT)

  Automatically generates multiple classifiers

  1. Train first net on sample A
  2. Train second net on sample B

     Sample B: half come from new samples the first net got right
     half come from new samples the first net got wrong

     .....

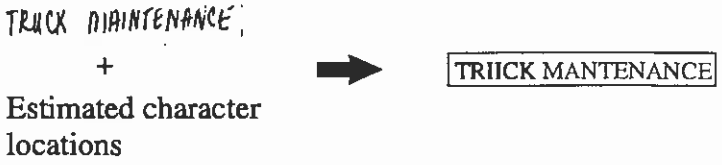
  Hurt rather than helped

- Instead used 2 classifiers

  Original integrated segmentation & recognition net
  Additional recognition net (used on centered characters only).

Word Segmentation

- Density histogram
- Estimated character locations



Dictionary Cleanup

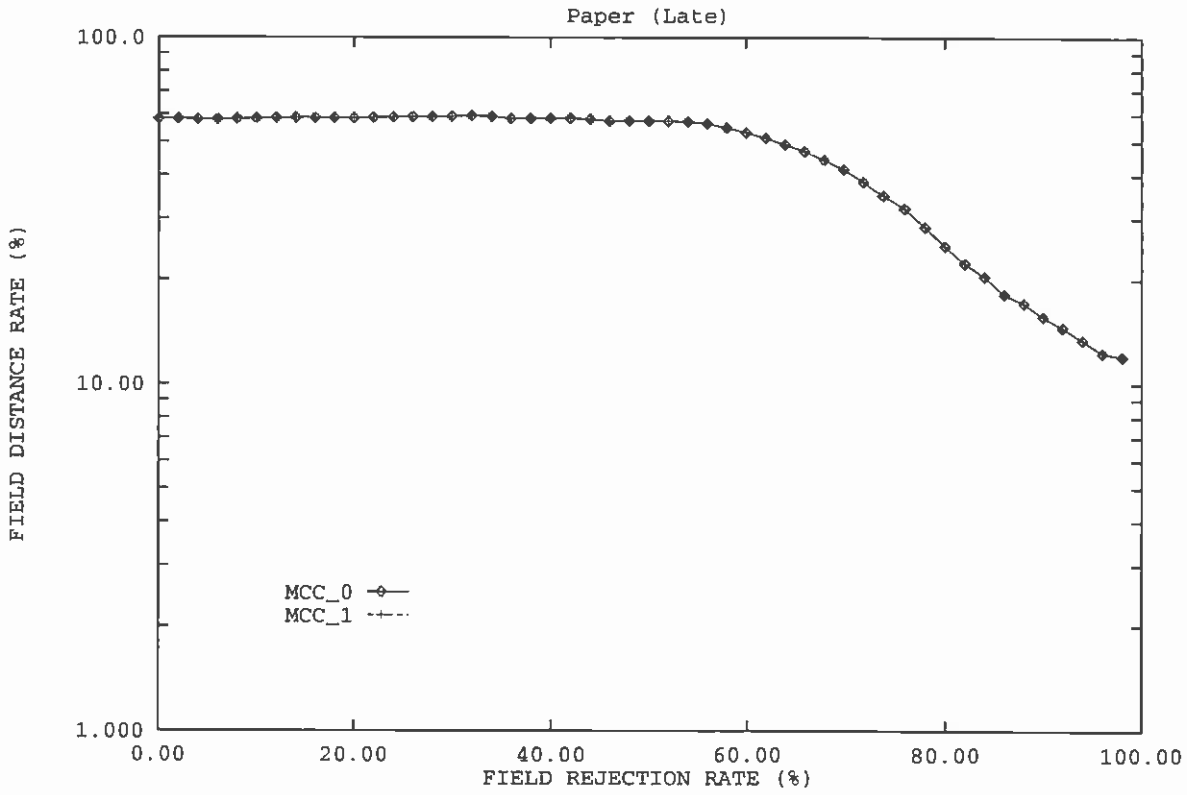
- First Pass:
  - Word–level Dynamic Programming
  - Frequent words (occurred 6 times or more in .ref files from database12 & database13
  - Dictionary sizes 800–1000 words
- Second Pass:
  - Long Phrase Dictionaries
  - Reject if not in the dictionary

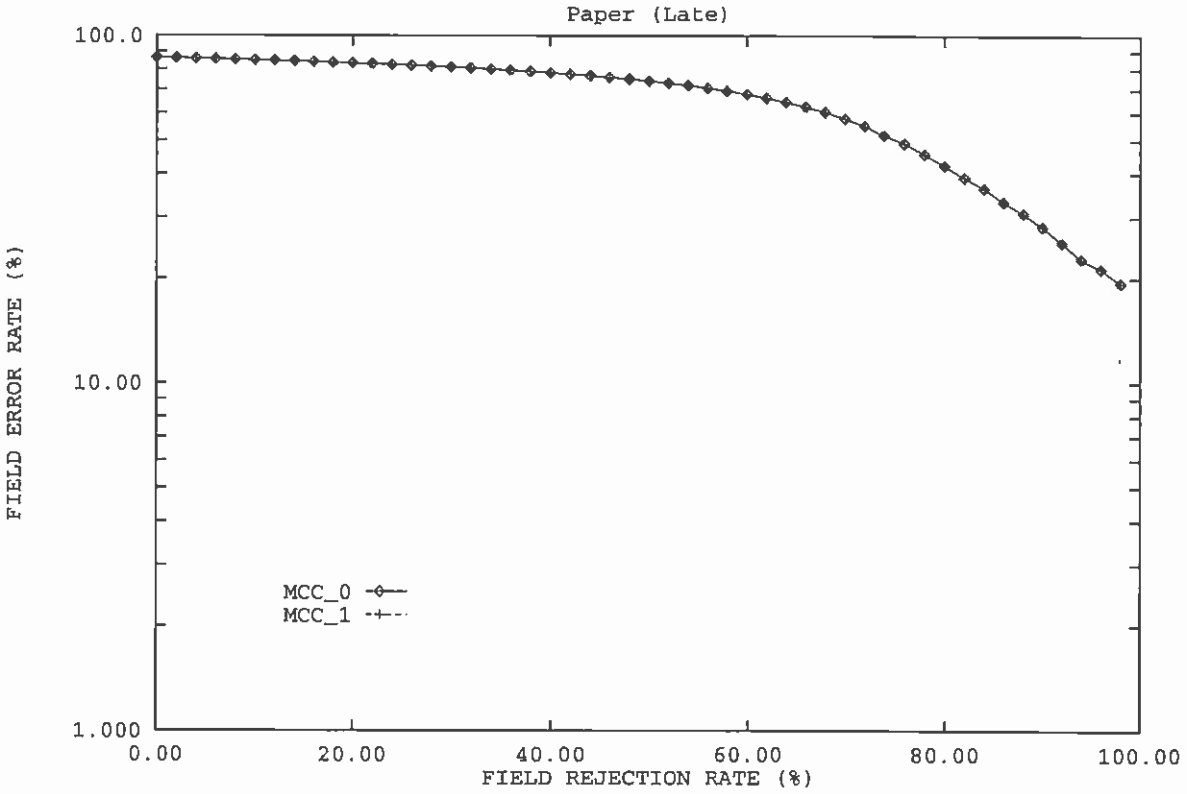
TRIICK MANTENANCE

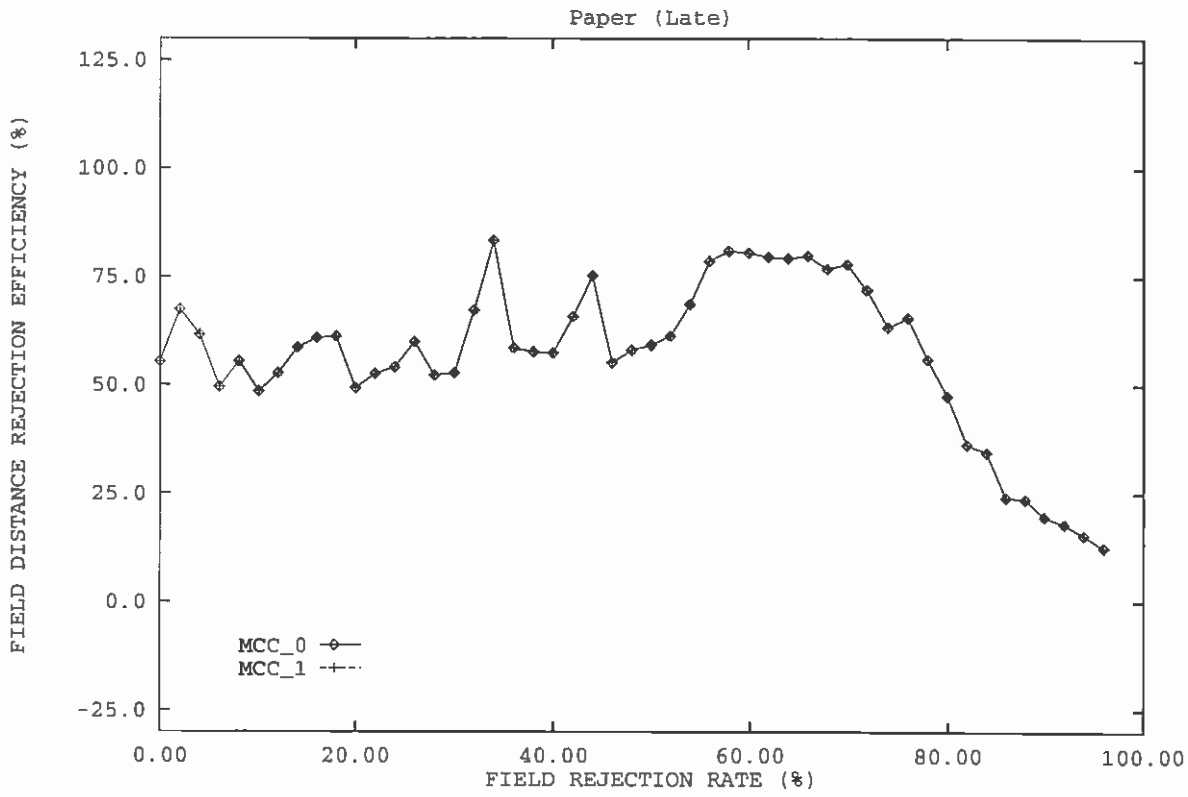
\+

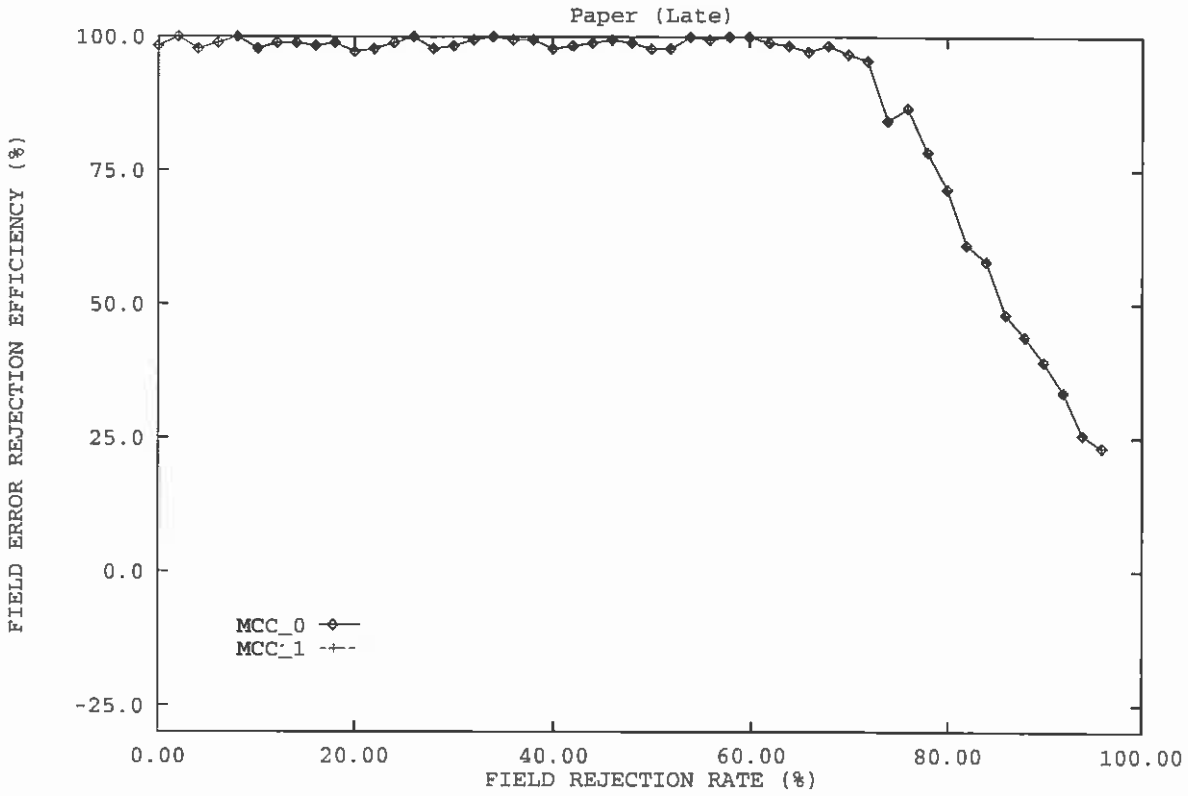
confidence values

TRUCK MAINTENANCE


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
MCC_0
MCC_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
MCC_0
MCC_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

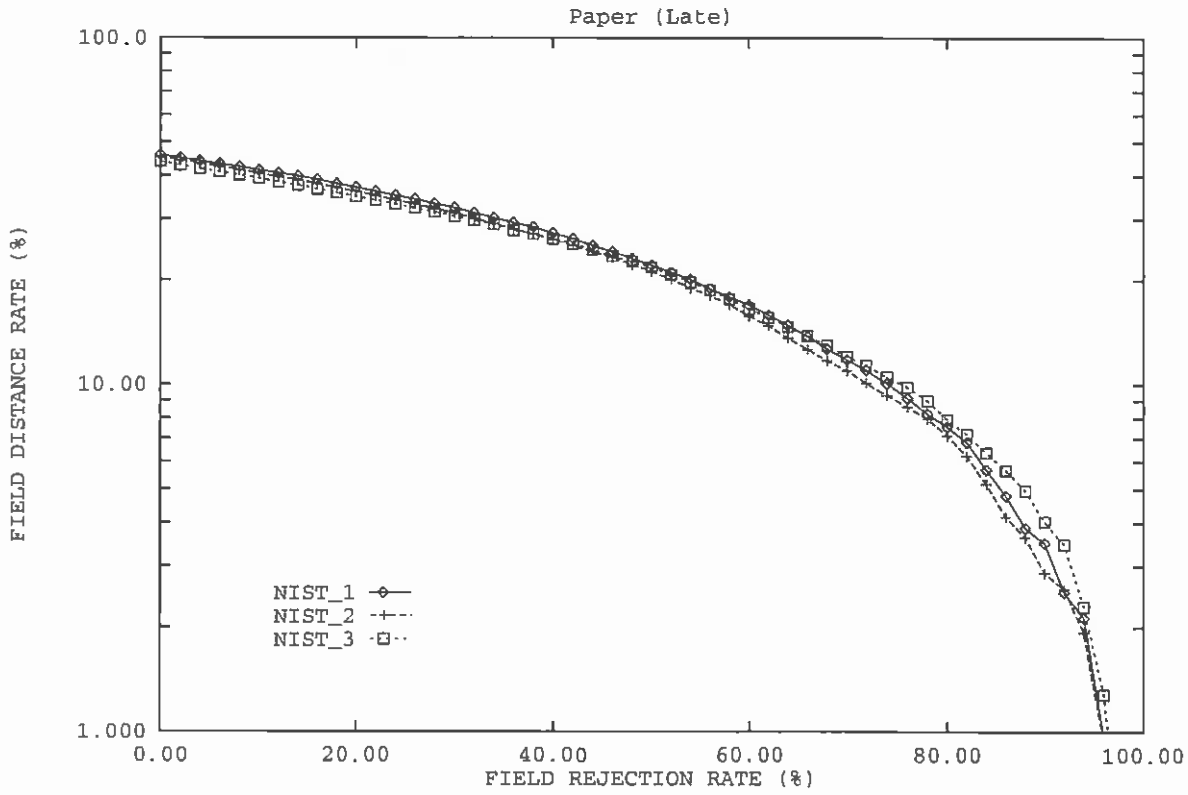

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
MCC_0
MCC_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

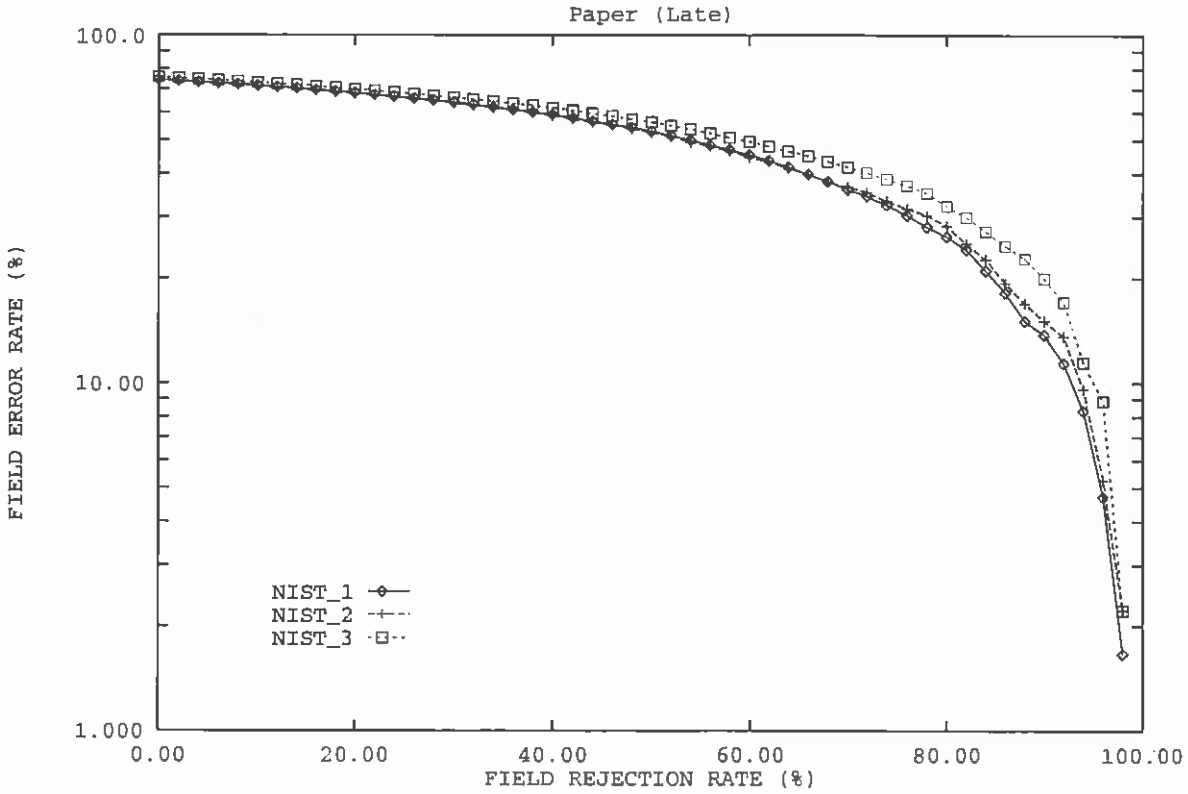

Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
MCC_0
MCC_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

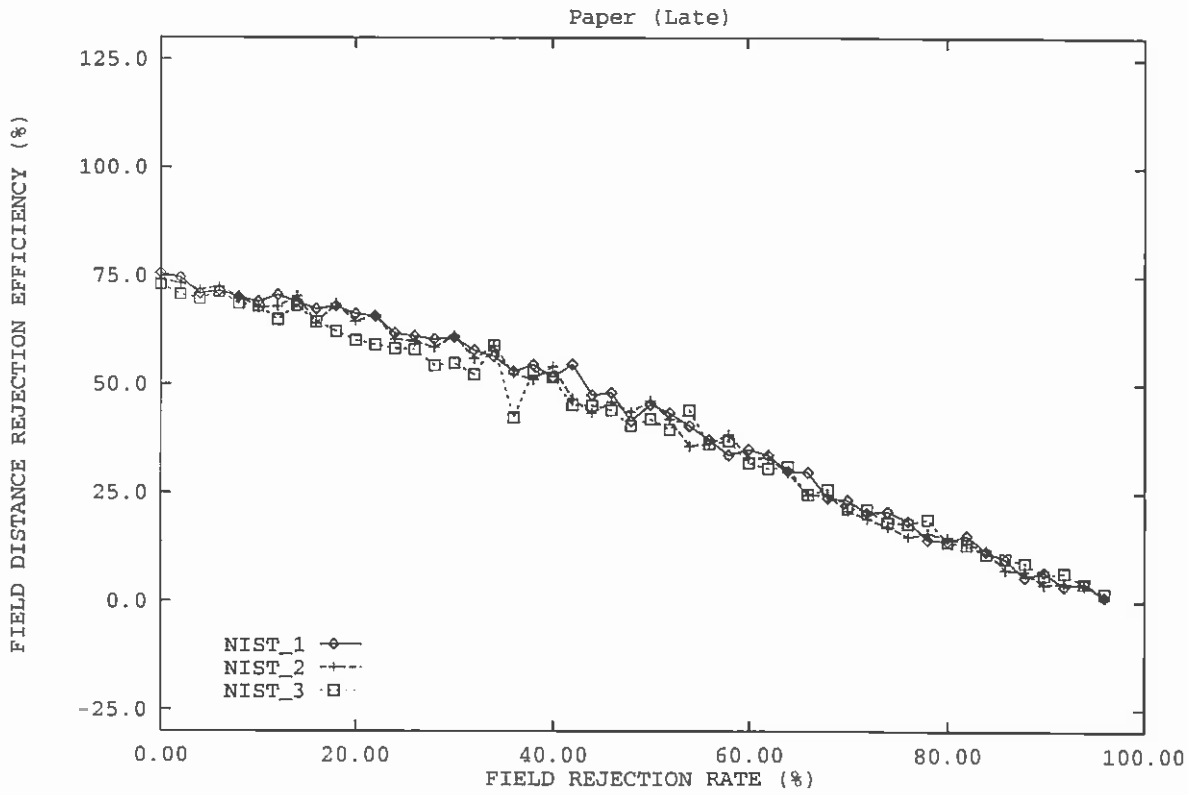
NIST

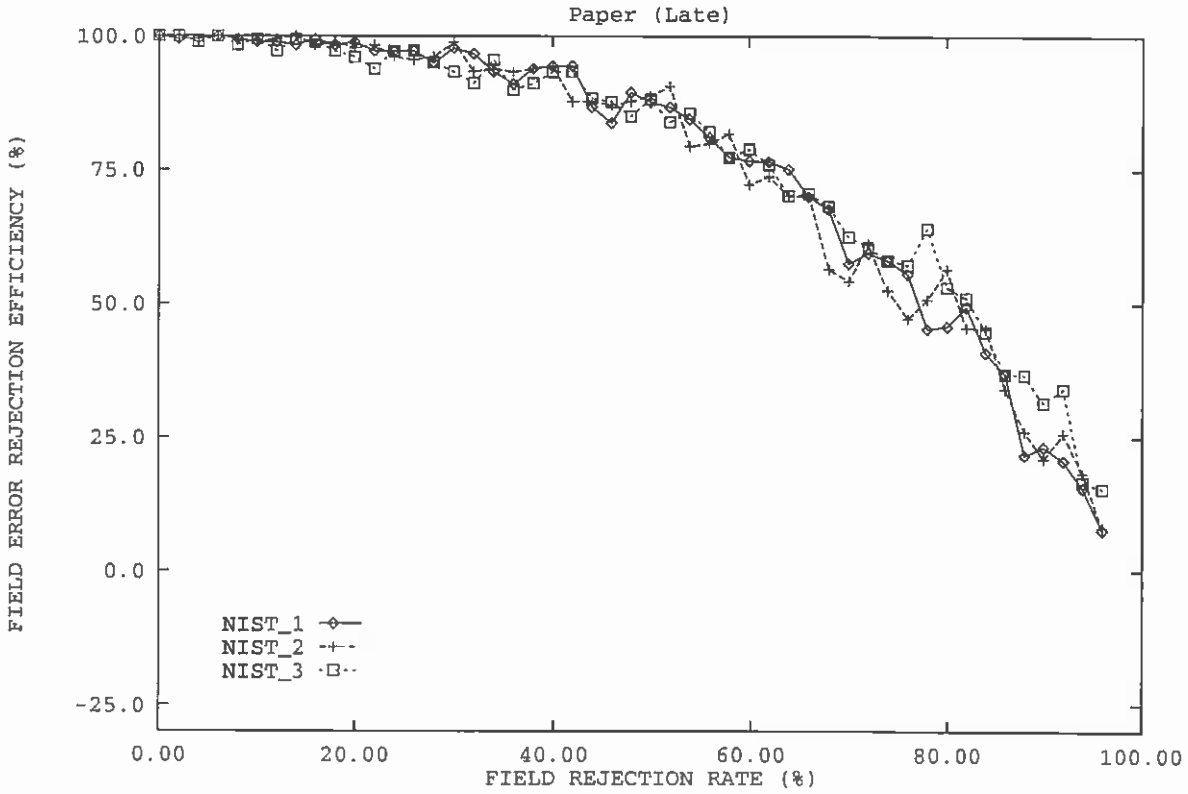
Late Submissions

See also: Summary for On-Time Submissions


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
NIST_1
NIST_2
NIST_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
NIST_1
NIST_2
NIST_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

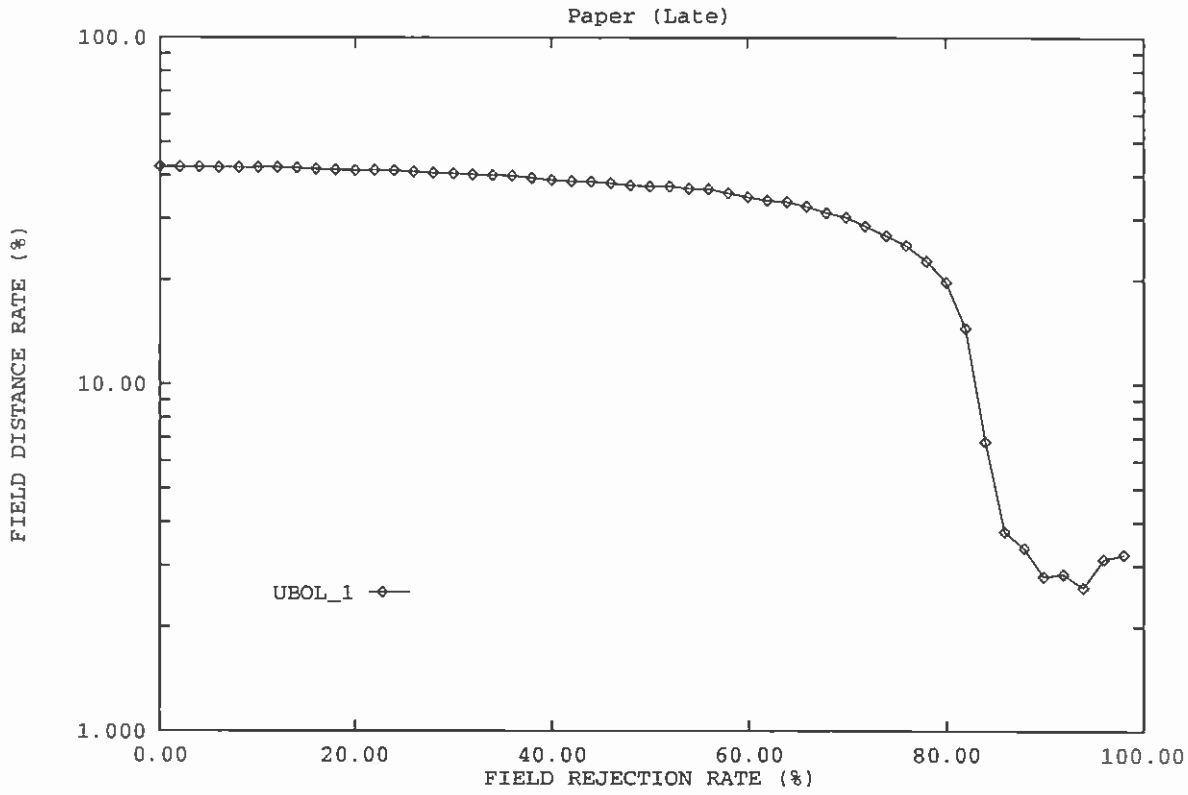

Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_1
NIST_2
NIST_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

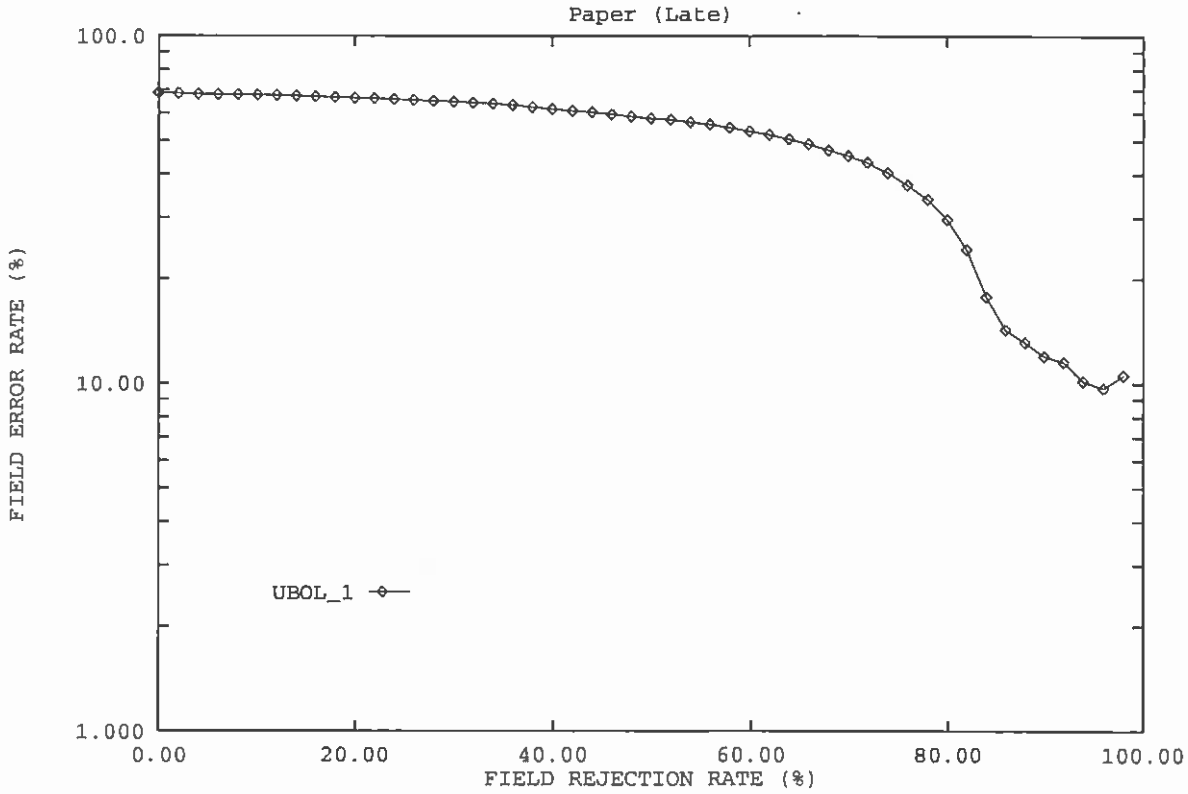

Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
NIST_1
NIST_2
NIST_3
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)

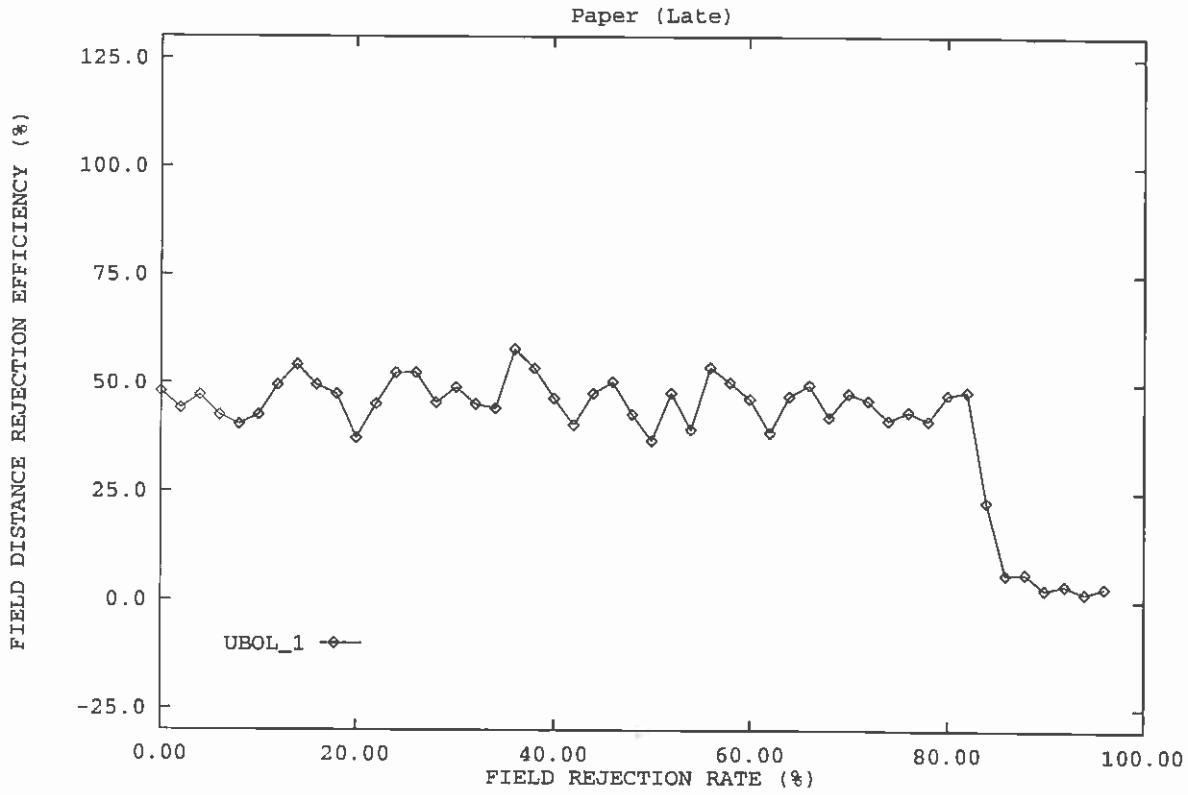
UBOL

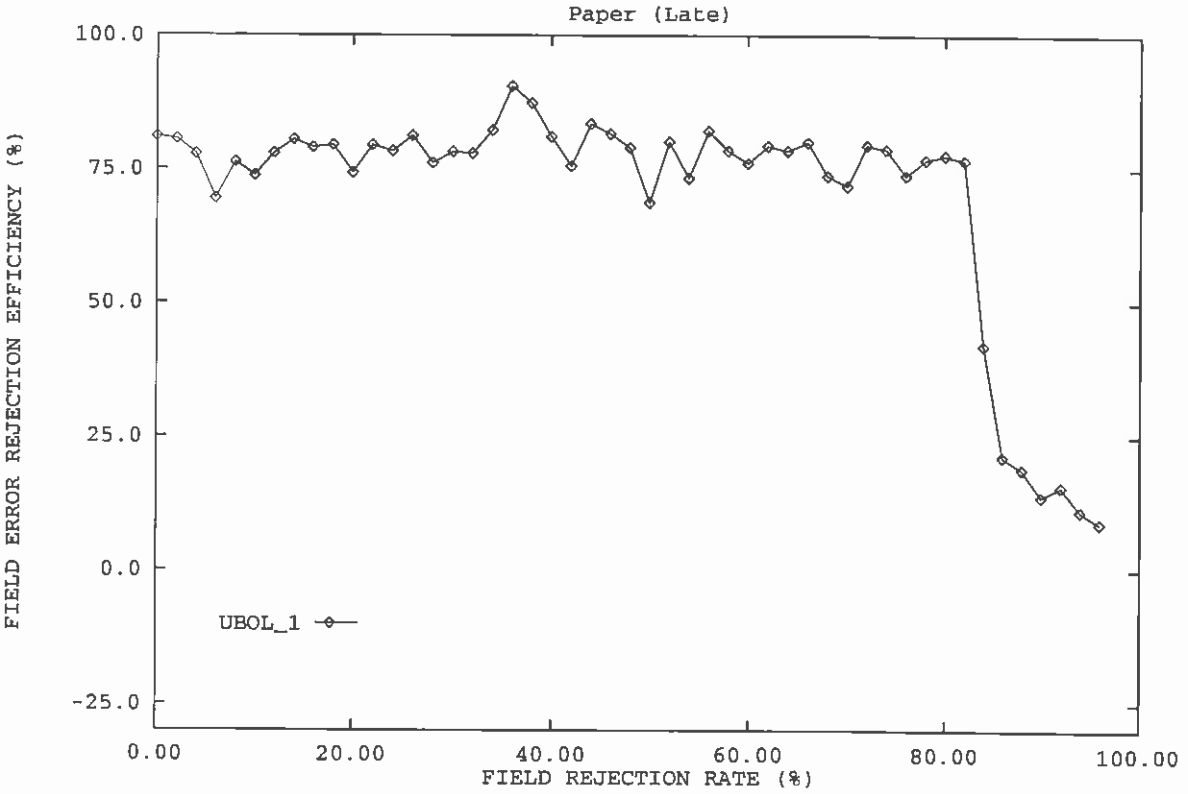
Late Submissions

See also: Summary for On-Time Submissions


Paper (Late)
100.0
10.00
1.000
FIELD DISTANCE RATE (%)
UBOL_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
100.0
10.00
1.000
FIELD ERROR RATE (%)
UBOL_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD DISTANCE REJECTION EFFICIENCY (%)
125.0
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)


Paper (Late)
FIELD ERROR REJECTION EFFICIENCY (%)
100.0
75.0
50.0
25.0
0.0
-25.0
UBOL_1
0.00
20.00
40.00
60.00
80.00
100.00
FIELD REJECTION RATE (%)