NISTIR 6101

Impact of Image Quality on Machine Print Optical Character Recognition

Michael D. Garris
Stanley Janet
William W. Klein

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Information Technology Laboratory
Information Access and
User Interfaces Division
Gaithersburg, MD 20899-0001

NIST

NISTIR 6101

Impact of Image Quality on Machine Print Optical Character Recognition

Michael D. Garris
Stanley Janet
William W. Klein

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Information Technology Laboratory
Information Access and
User Interfaces Division
Gaithersburg, MD 20899-0001

December 1997

U.S. DEPARTMENT OF COMMERCE
William M. Daley, Secretary

TECHNOLOGY ADMINISTRATION
Gary R. Bachula, Acting Under Secretary
for Technology

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
Raymond G. Kammer, Director

Impact of Image Quality on Machine Print Optical Character Recognition

Michael D. Garris, Stanley Janet, and William W. Klein

National Institute of Standards and Technology

ABSTRACT

The National Institute of Standards and Technology (NIST) is in the process of setting up a new series of conferences named the Metadata Text Retrieval Conferences (METTREC). They will focus on evaluating two critical technologies: document conversion using optical character recognition (OCR) and information retrieval (IR). Large collections of document images labeled with correct recognition and retrieval responses are needed to measure performance. Currently, the production of these materials is extremely expensive. NIST is developing a semi-automated truthing tool that will help reduce the cost of data preparation and enable evaluations to scale up. To accomplish this, current OCR technology is needed to produce an initial text to image alignment. This paper describes a small experiment in which three different vendor products (two Windows NT/95-based and one UNIX-based) are evaluated across three sets of document images containing progressively decreasing print and image quality. The evaluation images contain subjectively selected pages from the 1994 *Federal Register*. Results demonstrate the impact of degrading print and image quality with reported character recognition error rates ranging from 1% to as high as 74%.

Keywords: image quality, information retrieval, IR, machine print, METTREC, optical character recognition, OCR, page decomposition, technology evaluation

1.0 INTRODUCTION

The National Institute of Standards and Technology (NIST) is setting up a new series of conferences named the Metadata Text Retrieval Conferences (METTREC). They will focus on evaluating two critical technologies: document conversion using optical character recognition (OCR) and information retrieval (IR). Evaluations will be designed to investigate the impact of machine recognition errors on information retrieval and to determine what interfaces are appropriate to integrate the two technologies.

To support these evaluations, large training and testing sets of documents must be created. The *Federal Register* (FR) for 1994 has been chosen to be the initial source of documents for METTREC because it is: (1) a complete set of documents within the public domain; (2) a large collection containing over 250 issues consisting of over 67,000 pages of information; (3) a structured document set whose hierarchy contains metadata; (4) a collection of pages containing significant variations in print and image quality; and (5) a set of documents for which the text for the entire collection is stored in electronic files.

To conduct METTREC evaluations, each FR image page must be matched with its corresponding text to generate the "ground truth." The ground truth represents the correct text an OCR system should recognize and that an IR system should retrieve. Text for each day's issue of the FR has been provided by the Government Printing Office (GPO) and is stored in electronic files, but unfortunately the correspondence of the text to exact pages within an issue is not recorded. Ideally, we would like to know the image position of every word on a page.

We are currently working on a semi-automated process where the ground truth can be derived in an effective and efficient way. We have the images, and we have the correct text over a range of images. Our approach will use OCR to generate a "noisy" text to image correspondence. Dynamic string

alignment will then be used to match the correct GPO text to the noisy OCR text. For each page image, ground truth will be automatically produced where the OCR is sufficiently accurate. The correspondence of remaining passages of unaligned text will need to be assigned manually.

The higher the quality of recognition, the greater the yield of automatically generated ground truth. Techniques like this are needed to lower the per-page cost of generating test collections, which in turn permit evaluations to scale up. As a result, this approach places a premium on OCR accuracy. To move forward in the development of this approach, we needed an OCR technology capable of processing FR images with reasonably low rate of error.

In a prior effort that used NIST OCR technology to recognize FR page numbers[1], we observed a field error rate of 12%. Due to time constraints, adapting and developing our own in-house technology was too time-consuming and costly for recognizing the entire page. As a result, we decided to evaluate three commercially available OCR products. It has been stated that a 1% OCR error rate can only be attained by commercial OCR products whenever "a printed document is a fixed, typed original or a clean copy, in a simple paragraph format in a common typing font"[2]. Although this reference is from 1990, this still seems to be an accurate summary of the state of OCR. The pages of the FR certainly do not conform to these constraints, so we designed a small experiment to help determine the level of OCR performance that can be achieved. To accomplish this, three products were evaluated by focusing on their character level errors. Other sources of error, such as page decomposition and the processing of non-text items, were excluded from this evaluation.

The design of the evaluation is presented in Section 2.0; results are reported in Section 3.0, and conclusions are drawn in Section 4.0.

2.0 DESIGN OF THE EXPERIMENT

2.1 Image Scanning and Quality Verification

Before scanning the FR, its pages were cut from their bookbinding. This resulted in page sizes of approximately 20cm (8") by 28cm (11"). Image scanning was performed using a Kodak 923[1] scanner to output a compressed bitonal image in the tagged image format (TIFF™[3]). The scanner did not apply any special adaptive image enhancement to the grayscale image before converting it to a bitonal TIFF. Approximately 67,000 FR pages were scanned.

Reference [1] documents the process used to validate the entire collection of FR images. With the high-speed batch scanning of thousands of pages, a surprisingly large number of diversified errors were detected. Some errors appeared to be caused by the machinery and others by the operator. Images of pages were found to be missing, assigned to the wrong file, truncated, corrupted, skewed, scanned at the wrong density, etc. To ensure that each file contained a valid image, the following image quality checks were performed:

- Resolution = 15.75 pixels/mm (400 pixel/in)
- Compressed CCITT Group 4 image[4] file size ≥ 30 kilobytes (Kb)
- Width < 4000 pixels

---

[1] Specific hardware and software products identified in this paper were used in order to adequately support the evaluation of the technology described in this document. In no case does such identification imply recommendation or endorsement by NIST, nor does it imply that the equipment is necessarily the best available for the purpose.

- 4200 pixels < Height ≤ 4900 pixels

2.2 Federal Register Page Format

GPO prints a new issue of the FR each workday of the year. An issue is typically published in a single book and contains three distinct sections: a prefix, body, and appendix. Within the body section, "detail" pages elaborate and provide a record of the meeting notices, proposals, and transactions of the United States government for the day. Detail pages comprise 95% of the total FR page volume; therefore, recognition performance from this type of page was the focus for this experiment.

Detail pages like the one shown in Figure 1 are printed in mostly 9 point Vermilion font and contain a page heading that includes a text banner printed above two horizontal lines. The text banner contains information that identifies the document, the volume, the date, the topic, and a page number. All detail pages in this experiment contain three columns of information. Each page column may contain text, graphics, and/or tabular information that elaborate the transactions of the government. Since the primary focus of this experiment was to evaluate OCR character error rates, we excluded FR images containing graphics and tabular information.

2.3 Image Classification Criteria

The FR is printed on newspaper-quality recycled paper. The paper is light-weight and relatively absorbent so that printing ink frequently bleeds through the page. The quality of the typed print also fluctuates significantly. Patches of lightly printed characters and heavily smudged characters are often observed on the same page. Poor quality paper and high-speed printing contribute directly to varied image quality, which in turn directly impacts the rate of OCR errors.

To study the impact of these factors, three categories of image quality were defined for our evaluation: *good*, *bad*, and *ugly*. These are fairly subjective categorizations for which images were viewed on a 50.8cm (20") workstation display and judged based on the following criteria:

- **Good:** Illustrated in Figure 2, a *good* image contains a minimal amount noise, is easily readable, and has print quality that causes a minimal amount of characters to either touch or be broken.
- **Bad:** Illustrated in Figure 3, a *bad* image contains a moderate amount of noise, is readable by a human, and has print quality that causes many characters to either touch or be broken.
- **Ugly:** Illustrated in Figure 4, an *ugly* image contains an excessive amount of noise, contains sections that are illegible by a human, and has print quality that causes many characters to either touch or be broken.

Since we had to manually generate and verify the ground truth for each image, we limited our experiment to five representative pages for each class of image quality. A typical FR page contains 1100 to 1200 words totaling more than 6,000 characters. In all, 15 pages were used, providing over 70,000 characters from which OCR character error rates could be derived to compare the three OCR products.

specified assessment rate to cover such expenses will tend to effectuate the declared policy of the Act.

It is further found that good cause exists for not postponing the effective date of this action until 30 days after publication in the **Federal Register** [5 U.S.C. 553] because the Committee needs to have sufficient funds to pay its expenses which are incurred on a continuous basis. The 1994–95 fiscal year for the program began July 1, 1994. The marketing order requires that the rate of assessment apply to all assessable papayas handled during the fiscal year. In addition, handlers are aware of this action which was recommended by the Committee at a public meeting and published in the **Federal Register** as an interim final rule. No comments were received concerning the interim final rule that is adopted in this action as a final rule without change.

List of Subjects in 7 CFR Part 928

Marketing agreements, Papayas, Reporting and recordkeeping requirements.

For the reasons set forth in the preamble, 7 CFR part 928 is amended as follows:

PART 928—PAPAYAS GROWN IN HAWAII

Accordingly, the interim final rule amending 7 CFR part 928 which was published at 59 FR 33898 on July 1, 1994, is adopted as a final rule without change.

Dated: August 25, 1994.

Eric M. Forman,

Acting Deputy Director, Fruit and Vegetable Division.

[FR Doc. 94–21636 Filed 8–31–94; 8:45 am]

BILLING CODE 3410–02–P

---

7 CFR Part 947

[Docket No. FV94–947–2FIR]

Oregon-California Potatoes; Expenses and Assessment Rate

AGENCY: Agricultural Marketing Service, USDA.

ACTION: Final rule.

SUMMARY: The Department of Agriculture (Department) is adopting as a final rule, without change, the provisions of an interim final rule that authorized expenses and established an assessment rate that will generate funds to pay those expenses. Authorization of this budget enables the Oregon-California Potato Committee (Committee) to incur expenses that are reasonable and necessary to administer the program. Funds to administer this program are derived from assessments on handlers.

EFFECTIVE DATES: July 1, 1994, through June 30, 1995.

FOR FURTHER INFORMATION CONTACT: Martha Sue Clark, Marketing Order Administration Branch, Fruit and Vegetable Division, AMS, USDA, P.O. Box 96456, room 2523–S, Washington, DC 20090–6456, telephone 202–720–9918, or Teresa L. Hutchinson, Northwest Marketing Field Office, Fruit and Vegetable Division, AMS, USDA, Green-Wyatt Federal Building, room 369, 1220 Southwest Third Avenue, Portland, OR 97204, telephone 503–326–2724.

SUPPLEMENTARY INFORMATION: This rule is issued under Marketing Agreement No. 114 and Order No. 947, both as amended (7 CFR part 947), regulating the handling of Irish potatoes grown in Oregon-California. The marketing agreement and order are effective under the Agricultural Marketing Agreement Act of 1937, as amended (7 U.S.C. 601–674), hereinafter referred to as the Act.

The Department is issuing this rule in conformance with Executive Order 12866.

This rule has been reviewed under Executive Order 12778, Civil Justice Reform. Under the marketing order now in effect Oregon-California potato handlers are subject to assessments. Funds to administer the Oregon-California potato order are derived from such assessments. It is intended that the assessment rate as issued herein will be applicable to all assessable potatoes during the 1994–95 fiscal period, which began July 1, 1994, and ends June 30, 1995. This final rule will not preempt any State or local laws, regulations, or policies, unless they present an irreconcilable conflict with this rule.

The Act provides that administrative proceedings must be exhausted before parties may file suit in court. Under section 8c(15)(A) of the Act, any handler subject to an order may file with the Secretary a petition stating that the order, any provision of the order, or any obligation imposed in connection with the order is not in accordance with law and requesting a modification of the order or to be exempted therefrom. Such handler is afforded the opportunity for a hearing on the petition. After the hearing the Secretary would rule on the petition. The Act provides that the district court of the United States in any district in which the handler is an inhabitant, or has his or her principal place of business, has jurisdiction in equity to review the Secretary's ruling on the petition, provided a bill in equity is filed not later than 20 days after the date of the entry of the ruling.

Pursuant to the requirements set forth in the Regulatory Flexibility Act (RFA), the Administrator of the Agricultural Marketing Service (AMS) has considered the economic impact of this rule on small entities.

The purpose of the RFA is to fit regulatory actions to the scale of business subject to such actions in order that small businesses will not be unduly or disproportionately burdened. Marketing orders issued pursuant to the Act, and the rules issued thereunder, are unique in that they are brought about through group action of essentially small entities acting on their own behalf. Thus, both statutes have small entity orientation and compatibility.

There are approximately 550 producers of Oregon-California potatoes under this marketing order, and approximately 40 handlers. Small agricultural producers have been defined by the Small Business Administration (13 CFR 121.601) as those having annual receipts of less than $500,000, and small agricultural service firms are defined as those whose annual receipts are less than $5,000,000. The majority of Oregon-California potato producers and handlers may be classified as small entities.

The budget of expenses for the 1994–95 fiscal period was prepared by the Oregon-California Potato Committee, the agency responsible for local administration of the marketing order, and submitted to the Department for approval. The members of the Committee are producers and handlers of Oregon-California potatoes. They are familiar with the Committee's needs and with the costs of goods and services in their local area and are thus in a position to formulate an appropriate budget. The budget was formulated and discussed in a public meeting. Thus, all directly affected persons have had an opportunity to participate and provide input.

The assessment rate recommended by the Committee was derived by dividing anticipated expenses by expected shipments of Oregon-California potatoes. Because that rate will be applied to actual shipments, it must be established at a rate that will provide sufficient income to pay the Committee's expenses.

The Committee unanimously recommended a budget of $45,100, $1,500 more than last season. Increases in expenditures, which include $150 for the Committee's annual report, $50 for the Committee's audit, $1,000 for inspection fees, $500 for investigation

Figure 1. Detail page from the ***Federal Register***.

All States and Territories except Alabama, Connecticut, Hawaii, Alaska, Idaho, Kansas, Louisiana, Minnesota, Montana, Nebraska, Oklahoma, Oregon, Pennsylvania, South Dakota, Virginia, Washington, American Samoa and Palau have elected to participate in the Executive Order process and have established Single Points of Contact (SPOCs). Applicants from these 18 jurisdictions need take no action regarding Executive Order 12372. Applicants for projects to be administered by Federally-recognized Indian Tribes are also exempt from the requirements of E.O. 12372. Otherwise, applicants should contact their SPOCs as soon as possible to alert them of the prospective application and to receive any necessary instructions. Applicants must submit any required material to the SPOCs as soon as possible so that the program office can obtain and review SPOC comments as part of the award process. It is imperative that the applicant submit all required materials, if any, to the SPOC and indicate the date of this submittal (or the date of contact if no submittal is required) on the Standard Form 424, item 16a.

Under 45 CFR 100.8(a)(2), a SPOC has 60 days from the application deadline date to comment on proposed new or competing continuation awards.

SPOCs are encouraged to eliminate the submission of routine endorsements as official recommendations. Additionally, SPOCs are requested to differentiate clearly between mere advisory comments and those official State process recommendations which they intend to trigger the "accommodate or explain" rule.

When comments are submitted directly to ACF, they should be addressed to: Department of Health and Human Services, Administration for Children and Families, Division of Discretionary Grants, 6th Floor, OFM/DDG, 370 L'Enfant Promenade SW., Washington, DC 20447.

A list of Single Points of Contact for each State and Territory is included as appendix A of this announcement.

Applicable Regulations

Applicable HHS regulations will be provided to grantees upon awards.

Post-Award Requirements—Records and Reports

Grantees are required to file Financial Status (SF–269) on a semi-annual basis and Program Progress Reports on a quarterly basis. Funds shall be accounted for and reported upon separately from all other grant activities. Successful applicants for micro-enterprise development projects will be given specific instructions by ACF, following the award of the grant, for reporting grant performance and loan portfolio information.

The official receipt point for all reports and correspondence is the Division of Discretionary Grants. The original copy of each report shall be submitted to the Grants Management Specialist, Department of Health and Human Services, Administration for Children and Families, Division of Discretionary Grants, 6th Floor, OFM/DDG, 370 L'Enfant Promenade SW., Washington, DC 20447. A copy should be sent simultaneously to the Division of Operations, ORR. The mailing address is: Office of Refugee Resettlement, Division of Operations, Aerospace Building, Sixth Floor, 370 L'Enfant Promenade, SW., Washington, DC 20447.

The final Financial and Program Progress Reports shall be due 90 days after the project period expiration date or termination of grant support.

Although ORR does not expect the proposed components/projects to include evaluation activities, it does expect grantees to maintain adequate records to track and report on project outcomes and expenditures by budget line item.

The Catalog of Federal Domestic Assistance (CFDA) number assigned to this announcement is 93.576.

Dated: May 12, 1994.

Lavinia Limon,

Director, Office of Refugee Resettlement.

Appendix A

Executive Order 12372—State Single Points of Contact

Arizona

Mrs. Janice Dunn, ATTN: Arizona State Clearinghouse, 3800 N. Central Avenue, 14th floor, Phoenix, Arizona 85012, Telephone (602) 280–1315.

Arkansas

Ms. Tracie L. Copeland Manager, State Clearinghouse, Office of Intergovernmental Service, Department of Finance and Administration, P.O. Box 3278, Little Rock, Arkansas 72203, Telephone (501) 682–1074.

California

Mr. Glenn Stober, Grants Coordinator, Office of Planning and Research, 1400 Tenth Street, Sacramento, California 95814, Telephone (916) 323–7480.

Colorado

State Single Point of Contact, State Clearinghouse, Division of Local Government, 1313 Sherman Street, room 520, Denver, Colorado 80203, Telephone (303) 866–2156.

Delaware

Ms. Francine Booth, State Single Point of Contact, Executive Department, Thomas Collins Building, Dover, Delaware 19903, Telephone (302) 736–3326.

District of Columbia

Mr. Rodney T. Hallman, State Single Point of Contact, Office of Grants Mgmt. and Development, 717 14th Street NW., suite 500, Washington, D.C. 20005, Telephone (202) 727–6551.

Florida

Florida State Clearinghouse, Intergovernmental Affairs Policy Unit, Executive Office of the Governor, Office of Planning and Budgeting, The Capitol, Tallahassee, Florida 32399–0001, Telephone (904) 488–8114.

Georgia

Mr. Charles H. Badger, Administrator, Georgia State Clearinghouse, 254 Washington Street SW., room 534A, Atlanta, Georgia 30334, Telephone (404) 656–3855.

Illinois

Mr. Steve Klokkenga, State Single Point of Contact, Office of the Governor, 107 Stratton Building, Springfield, Illinois 62706, Telephone (217) 782–1671.

Indiana

Ms. Jean S. Blackwell, Budget Director, State Budget Agency, 212 State House, Indianapolis, Indiana 46204, Telephone (317) 232–5610.

Iowa

Mr. Steven R. McCann, Division of Community Progress, Iowa Department of Economic Development, 200 East Grand Avenue, Des Moines, Iowa 50309, Telephone (515) 281–3725.

Kentucky

Mr. Ronald W. Cook, Office of the Governor, Department of Local Government, 1024 Capitol Center Drive, Frankfort, Kentucky 40601, Telephone (502) 564–2382.

Maine

Ms. Joyce Benson, State Planning Office, State House Station #38, Augusta, Maine 04333, Telephone (207) 289–3261

Marylaod

Ms. Mary Abrams, Chief, Maryland State Clearinghouse, Department of State Planning, 301 West Preston Street, Baltimore, Maryland 21201–2365, Telephone (301) 225–4490

Massachusetts

Ms. Karen Arone, State Clearinghouse, Executive Office of Communities and Development, 100 Cambridge Street, room 1803, Boston, Massachusetts 02202, Telephone (617) 727–7001

Michigao

Mr. Richard S. Pastula, Director, Michigan Department of Commerce, Lansing, Michigan 48909, Telephone (517) 373–7356

Figure 2. "Good" Quality Image

to use the same species and strain as in other toxicological studies. Reasons for using rats as the predominant rodent species are practicality, comparability with other results obtained in this species and the large amount of background knowledge accumulated.

In embryotoxicity studies only, a second mammalian species traditionally has been required, the rabbit being the preferred choice as a "nonrodent." Reasons for using rabbits in embryotoxicity studies include the extensive background knowledge that has accumulated, as well as availability and practicality. Where the rabbit is unsuitable, an alternative nonrodent or a second rodent species may be acceptable and should be considered on a case-by-case basis (Note 5).

2.2 Other Test Systems

Other test systems are considered to be any developing mammalian and nonmammalian cell systems, tissues, organs, or organism cultures developing independently in vitro or in vivo. Integrated with whole animal studies either for priority selection within homologous series or as secondary investigations to elucidate mechanisms of action, these systems can provide invaluable information and, indirectly, reduce the numbers of animals used in experimentation. However, they lack the complexity of the developmental processes and the dynamic interchange between the maternal and the developing organisms. These systems cannot provide assurance of the absence of effect nor provide perspective in respect of risk/exposure. In short, there are no alternative test systems to whole animals currently available for reproduction toxicity testing with the aims set out in the introduction (Note 6).

3. General Recommendations Concerning Treatment

3.1 Dosages

Selection of dosages is one of the most critical issues in design of the reproductive toxicity study. The choice of the high dose should be based on data from all available studies (pharmacology, acute and chronic toxicity and kinetic studies, Note 7). A repeated dose toxicity study of about 2 to 4 weeks duration provides a close approximation to the duration of treatment in segmental designs of reproductive studies. When sufficient information is not available, preliminary studies are advisable (see Note 4).

Having determined the high dosage, lower dosages should be selected in a descending sequence, the intervals depending on kinetic and other toxicity factors. Whilst it is desirable to be able to determine a "no observed adverse effect level," priority should be given to setting dosage intervals close enough to reveal any dosage-related trends that may be present (Note 8).

3.2 Route and Frequency of Administration

In general the route or routes of administration should be similar to those intended for human usage. One route of substance administration may be acceptable if it can be shown that a similar distribution (kinetic profile) results from different routes (Note 9).

The usual frequency of administration is once daily but consideration should be given to use either more frequent or less frequent administration taking kinetic variables into account (see also Note 10).

3.3 Kinetics

It is preferable to have some information on kinetics before initiating reproduction studies since this may suggest the need to adjust choice of species, study design, and dosing schedules. At this time, the information need not be sophisticated nor derived from pregnant or lactating animals.

At the time of study evaluation, further information on kinetics in pregnant or lactating animals may be required according to the results obtained (Note 10).

3.4 Control Groups

It is recommended that control animals be dosed with the vehicle at the same rate as test group animals. When the vehicle may cause effects or affect the action of the test substance, a second (sham- or untreated) control group should be considered.

4. Proposed Study Designs—Combination of Studies

All available pharmacological, kinetic, and toxicological data for the test compound and similar substances should be considered in deciding the most appropriate strategy and choice of study design. It is anticipated that, initially, preference will be given to designs that do not differ too radically from those of established guidelines for medicinal products (the most probable option). For most medicinal products, the three-study design will usually be adequate. Other strategies, combinations of studies, and study designs could be as valid or more valid as the "most probable option" according to circumstances. The key factor is that, in total, they leave no gaps between stages and allow direct or indirect evaluation of all stages of the reproductive process (Note 11).

Designs should be justified.

4.1 The Most Probable Option

The most probable option can be equated to a combination of studies for effects on:

- Fertility and early embryonic development,
- Prenatal and postnatal development, including maternal function, and
- Embryo-fetal development.

4.1.1 Study of Fertility and Early Embryonic Development to Implantation

Aim

To test for toxic effects/disturbances resulting from treatment from before mating (males/females) through mating and implantation. This comprises evaluation of stages A and B of the reproductive process (see 1.2). For females this should detect effects on the oestrous cycle, tubal transport, implantation, and development of preimplantation stages of the embryo. For males it will permit detection of functional effects (e.g., on libido, epididymal sperm maturation) that may not be detected by histological examinations of the male reproductive organs (Note 12).

Assessment of

- Maturation of gametes,
- Mating behavior,
- Fertility,
- Preimplantation stages of the embryo, and
- Implantation.

Animals

At least one species, preferably rats.

Number of Animals

The number of animals per sex per group should be sufficient to allow meaningful interpretation of the data (Note 13).

Administration Period

The design assumes that, especially for effects on spermatogenesis, use will be made of data from repeated dose toxicity studies of at least 1-month duration. Provided no effects have been found that preclude this, a premating treatment interval of 2 weeks for females and 4 weeks for males can be used (Note 12). Selection of the length of the premating administration period should be stated and justified (see also 1.1, pointing out the need for research). Treatment should continue throughout mating to termination of males and at least through implantation for females. This will permit evaluation of functional effects on male fertility that cannot be detected by histologic examination in repeated dose toxicity studies and effects on mating behavior in both sexes. If data from other studies show there are effects on weight or histologic appearance of reproductive organs in males or females, or if the quality of examinations is dubious or if there are no data from other studies, then a more comprehensive study should be designed (Note 12).

Mating

A mating ratio of 1:1 is advisable and procedures should allow identification of both parents of a litter (Note 14).

Terminal Sacrifice

Females may be sacrificed at any point after midpregnancy.

Males may be sacrificed at any time after mating but it is advisable to ensure successful induction of pregnancy before taking such an irrevocable step (Note 15).

Observations

During study:

- Signs and mortalities at least once daily;
- Body weight and body weight changes at least twice weekly (Note 16);
- Food intake at least once weekly (except during mating);
- Record vaginal smears daily, at least during the mating period, to determine whether there are effects on mating or precoital time; and
- Observations that have proved of value in other toxicity studies.

At terminal examination:

- Necropsy (macroscopic examination) of all adults;
- Preserve organs with macroscopic findings for possible histological evaluation; keep corresponding organs of sufficient controls for comparison;
- Preserve testes, epididymides, ovaries and uteri from all animals for possible histological examination and evaluation on a

Figure 3. "Bad" Quality Image

requirements of amended section 110(a)(2).

B. Part D Requirements

Before Winston-Salem/Forsyth County may be redesignated to attainment, it also must have fulfilled the applicable requirements of part D. Under part D, an area's classification indicates the requirements to which it will be subject. Subpart 1 of part D sets forth the basic nonattainment requirements applicable to all nonattainment areas, classified as well as nonclassifiable. Subpart 3 of part D establishes additional requirements for nonattainment areas classified under section 186(a). The Winston-Salem area was classified as moderate (See 40 CFR 81.334). Therefore, in order to be redesignated to attainment, the State must meet the applicable requirements of subpart 1 of part D, specifically sections 172(c) and 176, and the requirements of subpart 3 of part D, which became due on or before April 27, 1994, the date the State submitted a complete redesignation request. EPA interprets section 107(d)(3)(v) to mean that, for a redesignation request to be approved, the State must have met all requirements that become applicable to the subject area prior to or at time of the submission of the redesignation request. The area will become subject to the CAA that come due subsequent to the submission of the redesignation request until the request is approved (See section 175A(c)) and if the redesignation is disapproved, the State remains obligated to fulfill those requirements.

B1. Subpart 1 of Part D—Section 172(c) sets forth general requirements applicable to all nonattainment areas. Under section 172(b), the section 172(c) requirements are applicable as determined by the Administrator but no later than three years after an area is designated as nonattainment. Because Winston-Salem was designated as a new CO nonattainment area on June 6, 1992, the requirements are not due until June 6, 1995. Therefore, the submission of a New Source Review program and contingency measures required under 172(c) are not yet due. The Region is, however, in the process of approving the State's revised NSR regulation which includes CO nonattainment areas. Upon redesignation of these areas to attainment, the Prevent of Significant Deterioration (PSD) provisions contained in part C of title I are applicable. On June 12, 1975, December 30, 1976, June 19, 1978, August 7, 1980, February 23, 1982, and August 15, 1994, EPA approved revisions to the State of North Carolina's PSD program (See 40 FR 25004, 41 FR 56805, 43 FR 26388, 45 FR 52676, 47 FR 7836, 59 FR 41708).

B2. Subpart 1 of Part D—Section 176(c) of the CAA requires States to revise their SIPs to establish criteria and procedures to ensure that Federal actions, before they are taken, conform to the air quality planning goals in the applicable SIP. The requirement to determine conformity applies to transportation plans, programs and projects developed, funded or approved under Title 23 U.S.C. or the Federal Transit Act ("transportation conformity"). Section 176 further provides that the conformity revisions to be submitted by States must be consistent with Federal conformity regulations that the CAA required EPA to promulgate. Congress provided for the State revisions to be submitted one year after the date for promulgation of final EPA conformity regulations. When that date passed without such promulgation, EPA's General Preamble for the Implementation of Title I informed States that its conformity regulations would establish a submittal date (see 57 FR 13498, 13557 (April 16, 1992)).

EPA promulgated final conformity regulations on November 24, 1993 (58 FR 62188) and November 30, 1993 (58 FR 63214). These conformity rules require that the States adopt both transportation and general conformity provisions in the SIP for areas designated nonattainment or subject to a maintenance plan approved under CAA section 175A. Pursuant to § 51.396 of the transportation conformity rule and § 51.851 of the general conformity rule, the State of North Carolina is required to submit a SIP revision containing transportation conformity criteria and procedures consistent with those established in the Federal rule by November 25, 1994. Similarly, North Carolina is required to submit a SIP revision containing general conformity criteria and procedures consistent with those established in the Federal rule by December 1, 1994. Because the deadlines for these submittals have not yet come due, they are not applicable requirements under section 107(d)(3)(E)(v) and, thus, do not affect approval of this redesignation request.

B3. Subpart 3 of Part D—Under section 187(a) areas designated nonattainment for CO under the amended CAA and classified as moderate were required to meet several requirements by November 15, 1992. North Carolina was required to submit a 1990 Emission Inventory. EPA has reviewed and is approving in this notice North Carolina's 1990 Base Year Emission Inventory. The requirement to make I/M corrections are not applicable to Forsyth County since it was not a pre-enactment nonattainment area, and therefore did not have an existing program before the CAA. Section 211(m) further required North Carolina to submit an oxygenated fuels regulation for the Winston-Salem area. North Carolina submitted a complete Oxygenated Fuel SIP on November 20, 1992. The Oxygenated Fuel Program is fully adopted and has been approved by EPA (See 59 FR 33683 published on June 30, 1994). Therefore, all Subpart 3 requirements that were applicable at the time the State submitted its redesignation request have been met.

3. Fully Approved SIP Under Section 110(k) of the CAA

Based on EPA's approval of SIP revisions under the 1990 Amendments, EPA has determined that the Winston-Salem/Forsyth County area has a fully approved SIP under section 110(k), which also meets the applicable requirements of section 110 and Part D as discussed above.

4. Improvement in Air Quality Due to Permanent and Enforceable Measures

The control measures to which the emission reductions are attributed mostly to the Federal Motor Vehicle Control Program (FMVCP). The fleet turnover under the FMVCP produced annual CO emission reductions of 6 percent.

In association with its emission inventory discussed below, the State of North Carolina has demonstrated that actual enforceable emission reductions are responsible for the air quality improvement and that the CO emissions in the base year are not artificially low due to local economic downturn. EPA finds that the combination of certain existing EPA-approved SIP and federal measures contribute to the permanence and enforceability of reduction in ambient CO levels that have allowed the area to attain the NAAQS.

5. Fully Approved Maintenance Plan Under Section 175A

Section 175A of the CAA sets forth the elements of a maintenance plan for areas seeking redesignation from nonattainment to attainment. The plan must demonstrate continued attainment of the applicable NAAQS for at least ten years after the Administrator approves a redesignation to attainment. Eight years after the redesignation, the state must submit a revised maintenance plan which demonstrates attainment for the ten years following the initial ten-year period. To provide for the possibility of future NAAQS violations, the

Figure 4. "Ugly" Quality Image

2.4 OCR Products Evaluated

For this experiment, we chose three products that were commercially available. All three products have a history of extensive participation in the University of Nevada at Las Vegas Information Science Research Institute's annual competition that evaluated and assessed recognition accuracy for machine printed documents[5]. Two of the products execute on a Windows 95/NT™[1] personal computer (referred to as PC products A and B) and one product executes on a UNIX™[1] workstation.

2.5 Scoring OCR Results

Each of the three OCR products classified character segments as:

- **Accepted:** An output character classification confidence value equaled or exceeded an OCR product's, user definable, threshold confidence value for character acceptance. This type of classification is not highlighted or marked and would not be presented to a reject repair operator for adjudication.
- **Rejected:** An output character classification confidence value was below the OCR product's, user definable, threshold value for character acceptance. This type of classification is highlighted and presented in context to a reject repair operator for adjudication.
- **Unrecognized:** An OCR product cannot classify the segmented area with enough confidence to output an ASCII representation of it. Instead, it outputs a user definable, unrecognizable character symbol; the unrecognizable character symbol is usually a "~" or a "?". Typically, the segment associated with this type of classification is highlighted and presented in context to a reject repair operator for adjudication.

All the products recognized and output OCR character results using the ISO 8859/1 character set. We observed that the two PC products correctly recognized the "?" and "~" (characters that are used to denote an unrecognized character). The UNIX product did not recognize the "~" character at all.

This evaluation focused on raw character classification and did not use confidence thresholds. Every character reported was scored without any rejection. As a result, character classifications were either scored as correct or as an error.

A modified version of the University of Washington Scoring Package[6] written by Su Chen was used. The primary enhancement to this software was the addition of word level scoring (not reported here). The scoring package dynamically aligns *n* output OCR result line strings with *m* reference line strings; and then, for each pair of matching OCR and reference lines, it aligns the characters within the lines and scores the results. It reports character, word, and line accuracy measurements.

We scored OCR results using truth data sets that were in reading order. As stated earlier, these files had to be prepared manually. The electronic files generated by GPO contain a tagged text representation from which the print copy of each FR book is typeset; however, specific page number identifiers and boundaries are not included. A truth file for each page was manually generated and cross-checked by viewing the page's image and extracting the corresponding text from the GPO file. The text was then edited to correspond, line for line, with the image page content. This process was manually intense, requiring approximately 20 minutes per page to prepare. It took a clerk/typist 50 minutes to type an entire page from scratch, so starting from the GPO file was less expensive. Due to time constraints, we included only 15 pages in the evaluation set.

In future METTREC evaluations, we anticipate relying heavily on word-level scores. However, for this small evaluation, we only compare OCR error rate where:

$$ErrorRate = 1 - \frac{\#correct}{\#objects}$$

and *#correct* is the total number of correctly recognized characters and *#objects* is the total number of characters to be recognized.

3.0 EXPERIMENTAL RESULTS

3.1 Page Decomposition

Before presenting the character recognition results, a brief discussion of page decomposition is in order. Pages of the FR are printed with three text columns. At times, graphical and tabular information spans multiple columns, creating relatively dynamic page layouts. Our goal is to produce ground truth in "reading" order, therefore automatic detection of the page layout is critical. A critical aspect is accurate decolumnization of each page.

Of the three OCR products tested, two have automatic page decomposition capabilities. They are PC product B and the UNIX product. The one that did not, PC product A, merely reports OCR character results in a top-to-bottom left-to-right order across the entire page. Of the two that did decomposition, PC product B failed to correctly decompose 3 of the 15 pages, whereas the UNIX product only failed to decompose 1 of the pages correctly. Note that the pages in this evaluation were comprised strictly of 3 text columns. Correct decomposition of more elaborate FR pages is a concern with all of the products tested.

Since all three products had problems decomposing one or more FR pages, we chose to compute character recognition scores using manually defined zones. In our application of generating ground truth, we realize that manual zoning each image is too time consuming and not practical.

3.2 OCR Character Error Rates

This section reports the character recognition error rates measured from three OCR products across three categories of image quality.

Figure 5 plots the character recognition error rates measured on the five FR pages of good quality. In the good collection, the pages contain 6130, 6739, 6483, 6346, and 6189 characters respectively, totaling 31,887 characters. All scores are within a 1.5% interval, ranging from just over 2% to 0.5%. PC product B performs best on all but the last page, but the separation in these scores is so small that the differences are likely to be statistically insignificant. (We have not run statistical tests of significance in this experiment, but this assertion is supported by the statistical limits reported in Reference [5]. Tests of statistical significance will be reported in future METTREC evaluations, but they are not currently conducted in the UW Scoring Package.)

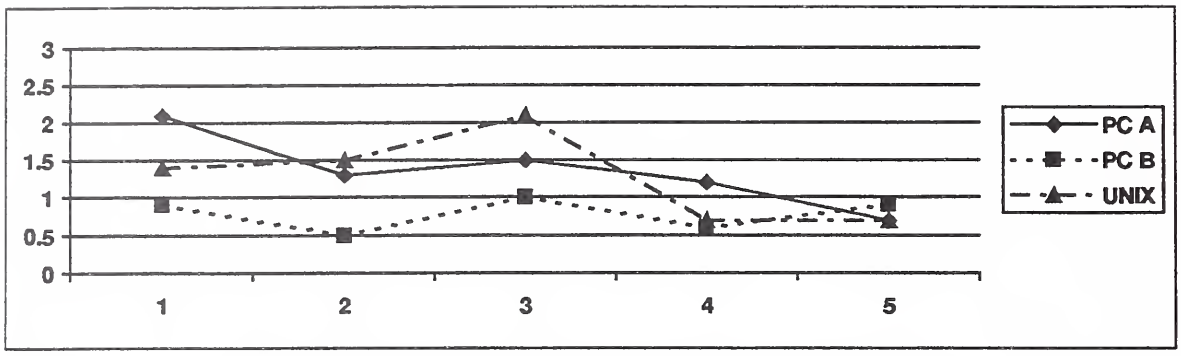

Figure 5. OCR Error Rate on Good Quality Images

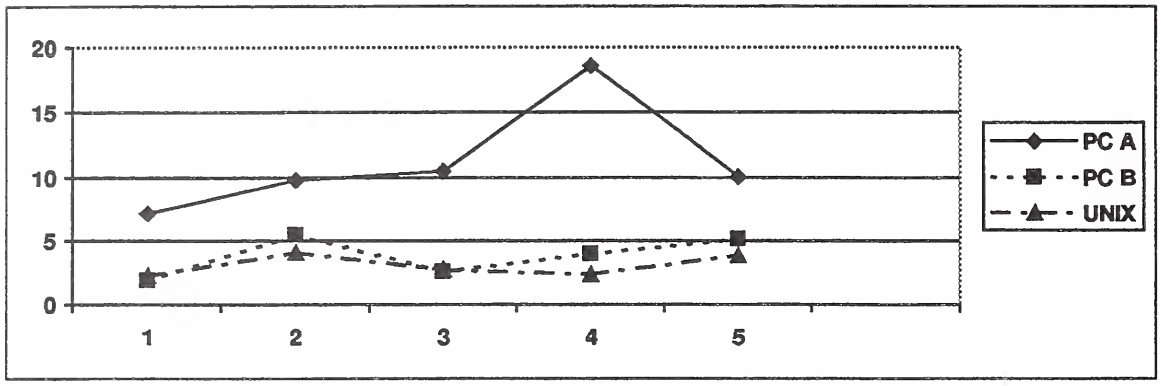

Figure 6. OCR Error Rate on Bad Quality Images

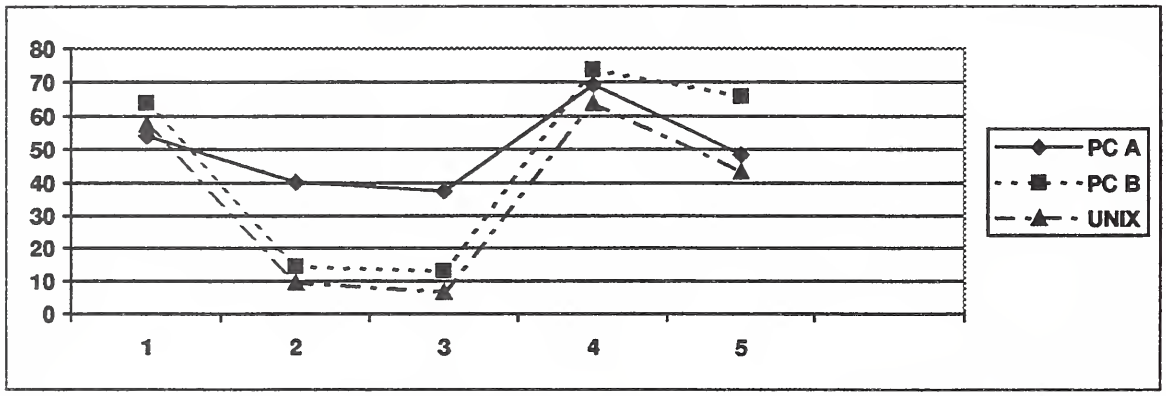

Figure 7. OCR Error Rate on Ugly Quality of Images

Figure 6 plots the character recognition error rates measured on the five FR pages of bad quality. In the bad collection, the pages contain 6639, 6290, 6588, 7836, 7892 characters respectively, totaling 35,245 characters. Unlike the results on the good pages, here there is significant separation between the products. PC product B and the UNIX product are tightly grouped (ranging between 2% and 5%), whereas PC product A performs consistently much worse (at least 5% worse on every page).

The results are a little more mixed in Figure 7. In the ugly collection, the pages contain 7285, 7123, 7080, 6203, and 7937 characters respectively, totaling 35,628 characters. As can be observed, the performance has fallen off dramatically with error rates reaching as high as 74%. PC product A actually performs best on the first page, but last on pages 2 and 3. PC product B tracks the UNIX product within 6% on the first three pages with performance falling off on the last two pages. The UNIX product is within 3% of PC product A on the first page, and then scores the best on the remaining 4 pages.

An interesting pattern can be observed in the scores plotted in Figure 7. All three products score best in the ugly set on pages 2 and 3, with PC product B having an error rate below 15% and the UNIX product below 10%. The lowest error rate measured for any system on the other 3 pages is over 40%. From these observations, there appear to be two types of pages represented in the ugly collection. Upon closer inspection of the images, this was confirmed. Ugly pages 2 and 3 contain a significant amount of "pepper" noise caused by ink bleed-through. Figure 8 shows a subimage containing this type of noise. Pages 1, 4, and 5 contain a different source of image degradation. In these pages, the printed characters are smudged due to a problem in the printing process. As can be seen in Figure 9, the characters appear to have been typed twice (once dark and once light) with a small translational offset. It should be noted that it is easier for a human to read the text in Figure 8 than the text illustrated in Figure 9. The latter requires a greater amount of word level context (as opposed to single character context) for a human to correctly identify a word.

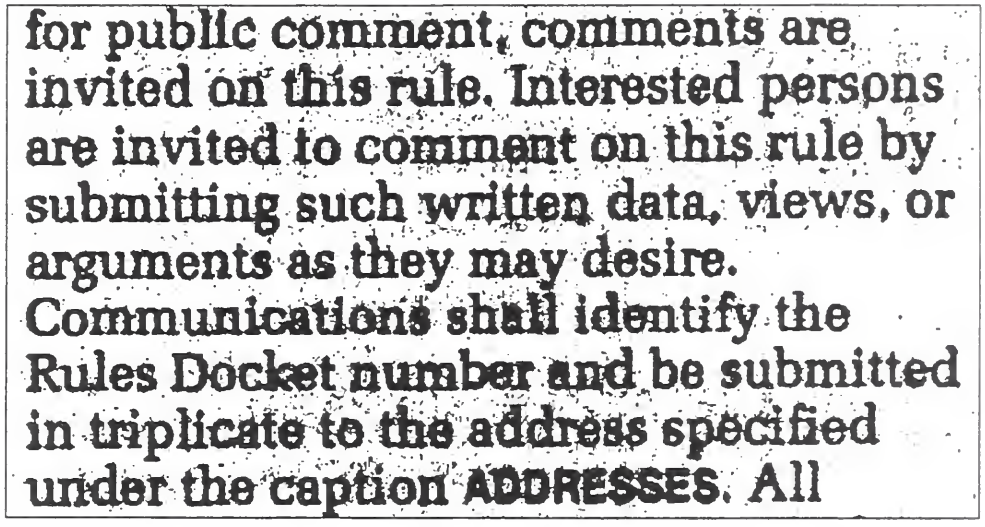

Figure 8. Pepper Noise from Ugly Page 2

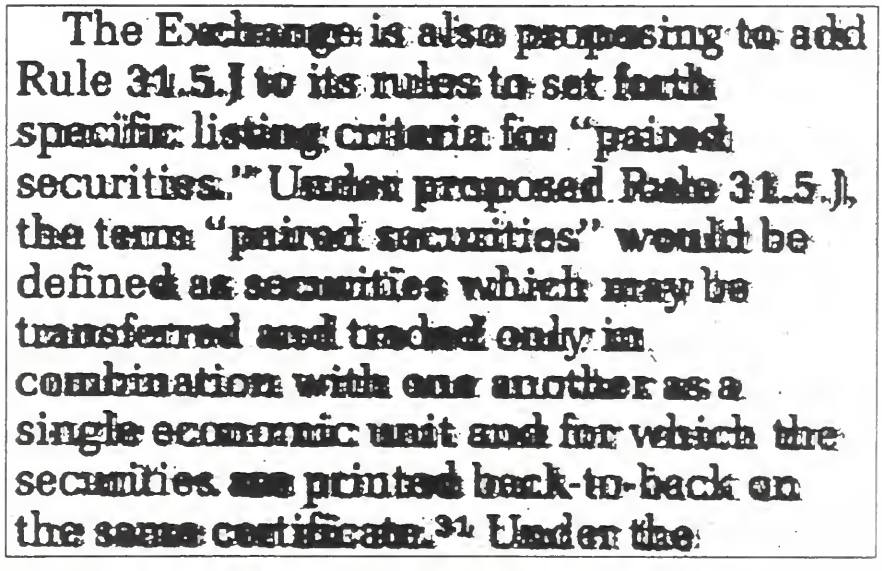

Figure 9. Smudged Characters from Ugly Page 5

Based on these scores, it appears that the vendors of PC product B and the UNIX product have reasonably good techniques for dealing with the presence of pepper noise; however, their error rates are significantly higher (by about 8%) than those on the good FR pages. In contrast, all the products performed poorly on

the pages with smudged characters. Perhaps this source of image degradation is unique to the FR and/or its publication process.

3.3 Timing Results

The UNIX product apparently detects poor image quality and applies additional processing resources to obtain a better segmentation and classification. Good quality images were optically recognized in less than 60 seconds (s), averaging 45s. Bad quality images were optically recognized in 60s to 120s, averaging 90s. Ugly quality images were optically recognized in over 120s, averaging 160s.

For the two PC products, the elapsed time to OCR an entire FR page was invariant with the quality of the image and averaged between 35s-40s. We conclude that the PC products are engineered primarily with speed in mind so that a somewhat linear/homogeneous solution is applied regardless of the quality of the image.

4.0 CONCLUSIONS

We have presented the results of a small OCR evaluation in which three different vendor products (two Windows NT/95-based and one UNIX-based) were tested. The purpose of the evaluation was to determine the state of commercial OCR technology with respect to processing pages of the *Federal Register* (FR). NIST must use this technology in order to produce initial text to image alignments for generating ground truth in future METTEC evaluations. This semi-automated truthing process will lower the cost of preparing testing materials and will permit experiments to scale up. Due to time constraints and the current cost of manually preparing ground truth for documents, fifteen FR pages were evaluated. Images from five different pages were visually and subjectively selected to represent each of three categories of progressively worse print and image quality. Though a small number, these pages contained over 70,000 characters. As a result, a number of interesting conclusions can be made.

Working with the products, we conclude that page decomposition is a very fragile technology, even on well-formed multi-column pages. Users should expect a relatively high error rate on more complex page layouts. Results suggest that OCR can produce good recognition results (error rates less than 1%) from high quality document images. On the other hand, current OCR technology produces dismal results (40% and higher) from document images that contain poor print quality and/or a high amount of image degradation. We did observe better performance on documents degraded with pepper noise than those degraded with smudged characters. By measuring execution times, we conclude that PC-based products are engineered primarily for speed and use a static algorithmic solution regardless of image quality. In contrast, the UNIX product exhibited the ability to detect low quality image and poor recognition conditions and alter its solution strategy to compensate. This enables a more adaptive and potentially more robust solution under difficult conditions. Based on all these factors, we selected the UNIX product for generating ground truth for METTREC.

When we searched for commercially available OCR, we found only a couple of UNIX-based products on the market. As companies migrate and develop OCR technology for PC's, their products are being targeted towards GUI-based, small-office automation applications that process relatively high quality document images. In the end, this will not serve the needs of corporations and government agencies that require the processing of low image quality documents in a centralized, high-speed, batch-oriented environment.

5.0 REFERENCES

[1] Michael D. Garris and William W. Klein, "Creating and Validating a Large Image Database for METTREC," NIST Internal Report 6090, November 1997.

[2] Richard G. Casey and Swam Y. Yong, *Image Analysis Applications*, Marcel Decker, Inc., Chapter 1, pp. 1-36, 1990.

[3] "TIFF™, Revision 6.0," June 3, 1992, Alludes Corporation, Seattle, WA 98104-2871.

[4] "Facsimile Coding Schemes and Coding Control Functions for Group 4 Fascicle VII.3 – Rec. T,6," 1984, CCITT.

[5] Steven V. Rice, Frank R. Jenkins, and Thomas Nartker, "The Fifth Annual Test of OCR Accuracy," Information Science Research Institute, TR-96-01, UNLV, April 1996.

[6] "UW-II English/Japanese Document Image Database," CD ROM, University of Washington, 1995.