The Fourth Annual Test of OCR Accuracy

Stephen V. Rice, Frank R. Jenkins, and Thomas A. Nartker

1 Introduction

For four years, ISRI has conducted an annual test of optical character recognition (OCR) systems known as “page readers.” These systems accept as input a bitmapped image of any document page, and attempt to identify the machine-printed characters on the page. In the annual test, we measure the accuracy of this process by comparing the text that is produced as output with the correct text. The goals of the test include:

1. to provide a current, independent assessment of system performance,
2. to measure the advances in the technology from year to year,
3. to gain insight into the complex nature of OCR, and
4. to identify problems at the state-of-the-art.

The scope of the test has increased greatly over the past four years. In the first test [Rice 92], six OCR systems processed binary images of 132 pages containing a total of 278,000 characters. These pages were randomly selected from a U.S. Department of Energy (DOE) database of scientific and technical documents. In the second year, new measures of performance were introduced in evaluating eight OCR systems using a larger DOE sample (460 pages and 817,000 characters) [Rice 93a, Kanai 93, Nartker 94a]. The third annual test re-used this DOE sample and featured a 200-page sample of articles from popular U.S. magazines scanned at three different resolutions. Six OCR systems were thus tested on pages containing nearly 1.5 million characters [Rice 94, Nartker 94b].

In this report, we present the results of the fourth annual test, our largest and most comprehensive to date. The test samples contain more than three million characters from business letters, DOE documents, and articles from magazines and newspapers. Each page has been scanned four times to produce binary images at three different resolutions, plus one gray scale image. Furthermore, fax images at two different resolutions have been obtained for each business letter page. We introduce our first non-English sample, which is a collection of Spanish-language newspaper articles, and for the first time, we report the speed of the OCR systems.
1.1 Participants

Any organization may participate in the annual test provided:

1. it submits a version of an OCR system by the established deadline (December 15, 1994 for the fourth annual test),

2. the version runs on a PC or Sun SPARCstation, and

3. the version can process specific regions of a TIFF image in a fully automatic (non-interactive) way.

Furthermore, only one entry is allowed per organization.

There are many features of OCR systems that are evaluated in this test. Submitted versions need not support all of these features. For example, if a version does not support automatic zoning or Spanish OCR, then it will simply be excluded from that portion of the test.

Table 1 lists the eight organizations that participated in this year’s test, and the versions they submitted. Hewlett Packard Laboratories submitted a research prototype that operates on only an HP workstation. This was allowed because HP provided the hardware, and facilitated the interface, well in advance of the deadline.

1.2 Test Data

Five test samples were used in this year’s test.

1. The Business Letter Sample contains a variety of letters received by businesses and individuals and donated to ISRI.

2. The DOE Sample is the third and largest sample we have prepared by randomly selecting pages from a DOE collection of scientific and technical documents.

3. The Magazine Sample, which was used in the third annual test, consists of pages selected at random from the 100 U.S. magazines having the largest circulation.

4. The English Newspaper Sample contains articles selected at random from the 50 U.S. newspapers having the largest circulation.

5. The Spanish Newspaper Sample contains articles selected at random from 12 popular newspapers from Argentina, Mexico, and Spain.

For the newspaper samples, only articles from the first section of the newspaper were selected, and each article was clipped from the newspaper.

Each test page was placed manually on the platen of a Fujitsu M3096G scanner, and then digitized four times to produce binary images at 200, 300, and 400 dots per inch (dpi), and an 8-bit gray scale image at 300 dpi. A global threshold of 127 (out of 255) was used to create the binary images for the Business Letter, DOE, and Magazine Samples. A different threshold was chosen for the newspaper samples: 75 for the English articles, and 95 for the Spanish articles.
Table 1: Participating Organizations

<table>
<thead>
<tr>
<th>Organization</th>
<th>Version Name</th>
<th>Version No.</th>
<th>Platform</th>
<th>Version Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caere Corp. Los Gatos, California</td>
<td>Caere OCR</td>
<td>138.1</td>
<td>Sun SPARCstation</td>
<td>pre-release</td>
</tr>
<tr>
<td>Electronic Document Technology Pte. Ltd. Singapore</td>
<td>EDT ImageReader</td>
<td>3.0</td>
<td>PC DOS</td>
<td>commercial release</td>
</tr>
<tr>
<td>Hewlett Packard Laboratories Bristol, England</td>
<td>HP Labs OCR</td>
<td>7.0</td>
<td>HP workstation</td>
<td>research prototype</td>
</tr>
<tr>
<td>International Neural Machines Inc. Waterloo, Ontario</td>
<td>INM NeuroTalker</td>
<td>2.52</td>
<td>PC DOS</td>
<td>beta release</td>
</tr>
<tr>
<td>Ligature Ltd. Jerusalem, Israel</td>
<td>Ligature CharacterEyes Pro</td>
<td>2.6</td>
<td>PC Windows</td>
<td>beta release</td>
</tr>
<tr>
<td>MAXSOFT-OCRON, Inc. Fremont, California</td>
<td>MAXSOFT-OCRON Recore</td>
<td>3.2</td>
<td>PC Windows</td>
<td>beta release</td>
</tr>
<tr>
<td>Recognita Corp. Budapest, Hungary</td>
<td>Recognita OCR</td>
<td>3.0</td>
<td>PC Windows</td>
<td>beta release</td>
</tr>
<tr>
<td>Xerox Imaging Systems, Inc. Peabody, Massachusetts</td>
<td>XIS OCR Engine</td>
<td>10.5</td>
<td>Sun SPARCstation</td>
<td>beta release</td>
</tr>
</tbody>
</table>

Table 2: Test Data

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pages</th>
<th>Zones</th>
<th>Words</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Letter Sample</td>
<td>200</td>
<td>1,419</td>
<td>51,460</td>
<td>319,756</td>
</tr>
<tr>
<td>DOE Sample</td>
<td>785</td>
<td>2,280</td>
<td>213,552</td>
<td>1,463,512</td>
</tr>
<tr>
<td>Magazine Sample</td>
<td>200</td>
<td>1,414</td>
<td>114,361</td>
<td>666,134</td>
</tr>
<tr>
<td>English Newspaper Sample</td>
<td>200</td>
<td>781</td>
<td>84,026</td>
<td>492,090</td>
</tr>
<tr>
<td>Spanish Newspaper Sample</td>
<td>144</td>
<td>558</td>
<td>57,670</td>
<td>348,091</td>
</tr>
<tr>
<td>Total</td>
<td>1,529</td>
<td>6,452</td>
<td>521,069</td>
<td>3,289,573</td>
</tr>
</tbody>
</table>
We created fax images of the Business Letter Sample by transmitting each page locally, using standard and fine modes, from a Xerox 7024 fax machine to a fax modem. The standard-mode fax images have a resolution of 204 dpi in the X-direction, and 98 dpi in the Y-direction. The fine-mode images are also 204 dpi in the X-direction, but are 196 dpi in the Y-direction.

We manually “zoned” each page, i.e., we delineated and ordered the text regions of the page. The OCR systems processed only these “zones.” Some text was deemed to be unsuitable for the test, and was excluded; examples include equations, advertisements, text that is part of a figure (such as the labelling of a graph or map), and text that is considered to be unreadable by humans.

We carefully prepared the correct text, or “ground-truth,” corresponding to each zone. To ensure the highest possible accuracy, the text for each zone was entered four times, by different typists working independently. The four versions were reconciled with the help of a difference algorithm.

Table 2 gives the number of pages, zones, words, and characters in each test sample.

1.3 Test Operation

Version 5.0 of the OCR Experimental Environment was used to conduct the fourth annual test. This is a suite of software tools developed by ISRI for large-scale, automated testing and experimental research in OCR. An earlier version of this software is described in [Rice 93b].

This software runs on Sun SPARCstations and provides remote control of PCs. Each OCR system is operated in a fully automatic manner, i.e., without human interaction. The comparison of OCR-generated text with correct text, and the tabulation of accuracy statistics, are performed entirely under computer control.

Pages were not re-scanned for each OCR system, nor were page images re-zoned; all OCR systems processed the same zoned portions of the same page images. Unless otherwise noted, tests were run using the 300 dpi binary and gray scale images. Exceptions are the tests involving the fax business letters, and the test of the effect of resolution, which also utilized the 200 and 400 dpi binary images.

Caere OCR and the XIS OCR Engine were operated under SunOS 4.1.3 on a single-processor Sun SPARCstation 10 with 64 megabytes of memory. The five PC-based OCR systems performed on identically-configured 486DX/33 machines with 8 megabytes of memory, running under MS-DOS 5.0, and for three of these, MS Windows 3.1. HP Labs OCR was operated under HP-UX A.09.01 on an HP 9000 Model 735 with 32 megabytes of memory.

Each machine was unburdened when timing figures were recorded, i.e., only the OCR system was running on the machine. Each OCR system processed one page image per invocation; thus, the timing figures include the modest overhead of initializing the OCR system for each image.

2 Character Accuracy

While there are many ways of quantifying the deviation between OCR-generated and correct text, in our most fundamental measure, we reflect the effort required by a human editor to correct the OCR-generated text. Specifically, we compute the minimum number of edit operations (character
insertions, deletions, and substitutions) needed to fully correct the text. We refer to this quantity as the number of \textit{errors} made by the OCR system. Expressing this as a percentage of the total number of characters, we obtain the \textit{character accuracy}:

\[
\frac{\text{#characters} - \text{#errors}}{\text{#characters}}.
\]

In the past, we utilized an algorithm that tended to over-estimate the minimum number of edit operations by 5 to 10\%. In the interest of precise reporting, this year we have switched to an algorithm that computes the minimum number exactly. It is an optimized version of an algorithm by Ukkonen \cite{Ukkonen85}.

Tables 3a-3f give the character accuracy results for each test sample. Some entries are missing due to unsupported features. Only Caere OCR and HP Labs OCR accept gray scale input. EDT ImageReader, HP Labs OCR, and INM NeuroTalker do not support Spanish OCR.

2.1 Failures

A failure is detected when an OCR system “crashes” or “hangs” when processing a page image, or when it returns an error status upon termination. An entry of \textit{none} in the \textit{Failures} column indicates that no failures were detected. Otherwise, the number of failed pages is specified, followed by the number of characters on those pages, expressed as a percentage of the total number of characters in the sample. If the latter exceeds one percent, then the failures are deemed to be excessive, and the accuracy results are not reported; otherwise, errors are charged equal to the number of characters on the failed pages.

Not all failures are detected. Bugs could cause extraneous characters to be output, or could prevent the generation of correct characters. Failures of this type cannot be distinguished in an automatic and reliable way from recognition errors. They go undetected, and errors are charged in proportion to the editing effort needed to correct the damage.

If the character accuracy of an OCR system is less than 90\% for a particular sample, then we note only that the accuracy falls below this threshold. We do not report further the performance of this system on this sample.

2.2 Confidence Intervals

Graphs 1a-1g show approximate 95\% confidence intervals for character accuracy. These intervals were computed using a technique from statistics known as the \textit{jackknife estimator} \cite{Dudewicz88}. In applying this technique, we have made the assumption that the pages within a sample are independent, but we have not assumed that the characters within a page are independent.

An OCR system that performs consistently within a sample is represented by a narrow interval, whereas a wide interval indicates considerable variability. When comparing the performance of two systems, non-overlapping intervals imply that there is a statistically significant difference between the systems.
### Character Accuracy

#### Table 3a: Original Business Letters

<table>
<thead>
<tr>
<th></th>
<th>300 dpi Binary</th>
<th></th>
<th></th>
<th>300 dpi 8-bit Gray Scale</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
<td>Failures</td>
<td>Errors</td>
<td>% Accuracy</td>
<td>Failures</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>4,459</td>
<td>98.61</td>
<td>none</td>
<td>3,102</td>
<td>99.03</td>
<td>none</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>13,162</td>
<td>95.88</td>
<td>1 / 0.30</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>5,959</td>
<td>98.14</td>
<td>none</td>
<td>4,850</td>
<td>98.48</td>
<td>none</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>none</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>---</td>
<td>---</td>
<td>1 / 1.07</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>8,377</td>
<td>97.38</td>
<td>none</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>11,280</td>
<td>96.47</td>
<td>none</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>5,473</td>
<td>98.29</td>
<td>none</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

#### Table 3b: Fax Business Letters

<table>
<thead>
<tr>
<th></th>
<th>Standard-mode Fax</th>
<th></th>
<th>Fine-mode Fax</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
<td>Failures</td>
<td>Errors</td>
<td>% Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>18,361</td>
<td>94.26</td>
<td>none</td>
<td>7,559</td>
<td>97.64</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>1 / 0.70</td>
<td>15,345</td>
<td>95.20</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>none</td>
<td>8,815</td>
<td>97.24</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>none</td>
<td>24,552</td>
<td>92.32</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>15,689</td>
<td>95.09</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>none</td>
<td>9,403</td>
<td>97.06</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>10,193</td>
<td>96.81</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>17,541</td>
<td>94.51</td>
<td>none</td>
<td>7,453</td>
<td>97.67</td>
</tr>
</tbody>
</table>

#### Table 3c: DOE Sample

<table>
<thead>
<tr>
<th></th>
<th>300 dpi Binary</th>
<th></th>
<th></th>
<th>300 dpi 8-bit Gray Scale</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
<td>Failures</td>
<td>Errors</td>
<td>% Accuracy</td>
<td>Failures</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>37,503</td>
<td>97.44</td>
<td>2 / 0.50</td>
<td>32,791</td>
<td>97.76</td>
<td>1 / 0.33</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>94,234</td>
<td>93.56</td>
<td>1 / 0.13</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>36,349</td>
<td>97.52</td>
<td>none</td>
<td>33,390</td>
<td>97.72</td>
<td>none</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>---</td>
<td>&lt; 90.00</td>
<td>none</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>---</td>
<td>---</td>
<td>7 / 1.28</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>56,746</td>
<td>96.12</td>
<td>none</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>57,713</td>
<td>96.06</td>
<td>none</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>34,644</td>
<td>97.63</td>
<td>none</td>
<td>--</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
### Table 3d: Magazine Sample

<table>
<thead>
<tr>
<th>OCR System</th>
<th>300 dpi Binary</th>
<th>300 dpi 8-bit Gray Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>14,483</td>
<td>97.83</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>15,043</td>
<td>97.74</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>---</td>
<td>&lt; 90.00</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>41,563</td>
<td>93.76</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>23,312</td>
<td>96.50</td>
</tr>
<tr>
<td>Recoginta OCR</td>
<td>26,474</td>
<td>96.03</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>16,784</td>
<td>97.48</td>
</tr>
</tbody>
</table>

### Table 3e: English Newspaper Sample

<table>
<thead>
<tr>
<th>OCR System</th>
<th>300 dpi Binary</th>
<th>300 dpi 8-bit Gray Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>5,079</td>
<td>99.97</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>6,432</td>
<td>98.69</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>47,773</td>
<td>90.29</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>11,230</td>
<td>97.72</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>7,002</td>
<td>98.58</td>
</tr>
<tr>
<td>Recoginta OCR</td>
<td>10,495</td>
<td>97.87</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>5,513</td>
<td>98.88</td>
</tr>
</tbody>
</table>

### Table 3f: Spanish Newspaper Sample

<table>
<thead>
<tr>
<th>OCR System</th>
<th>300 dpi Binary</th>
<th>300 dpi 8-bit Gray Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Errors</td>
<td>% Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>5,394</td>
<td>98.45</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>13,512</td>
<td>96.12</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>10,012</td>
<td>97.12</td>
</tr>
<tr>
<td>Recoginta OCR</td>
<td>8,929</td>
<td>97.43</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>7,213</td>
<td>97.93</td>
</tr>
</tbody>
</table>
2.3 Speed and Throughput

For most applications, speed is less important than accuracy. Indeed, there is little use for a fast OCR system that produces mostly gibberish as output. But given OCR systems of comparable accuracy, speed becomes an important factor.

We are generally opposed to reporting raw speed figures without consideration of accuracy. Thus, we introduce the following throughput function which reports speed while penalizing for errors:

\[
\frac{\text{characters} - P \times \text{errors}}{\text{seconds}}.
\]

\(P\) represents the penalty assigned to each error. When \(P = 0\), the function gives the raw speed in terms of characters per second. Some authors have defined throughput to mean “correct characters” per second, which corresponds to \(P = 1\). We do not feel that this is a sufficient penalty for errors; hence, in Graphs 2a-2g, we present throughput for \(P = 0\) to 10.

2.4 Accuracy by Character Class

In accuracy by character class, we divide the ground-truth characters into classes, and determine the percentage of characters in each class that were correctly recognized. The following classes were used.

1. **Spacing**: the blank and end-of-line characters,
2. **a - z**: the lowercase letters,
3. **A - Z**: the uppercase letters,
4. **0 - 9**: the decimal digits, and
5. **Special**: punctuation and other special symbols.

For the Spanish Newspaper Sample, a Spanish class was added containing the Spanish accented letters and punctuation symbols. Graphs 3a-3g display the results for each test sample.

The largest of these classes is the **a - z** class; depending on the sample, 68 to 75% of the ground-truth characters belong to this class. The second largest class is the **Spacing** class, accounting for 16 to 17% of the characters.

The **A - Z**, **0 - 9**, and **Special** classes contain 3-7%, 1-6%, and 3-5% of the characters, respectively. The **Spanish** class contains 2% of the characters in the Spanish Newspaper Sample. The OCR systems were less accurate on these smaller classes.

2.5 Effect of Resolution

Graphs 4a-4e show how character accuracy is affected by decreasing the resolution of binary images from 300 to 200 dpi, and by increasing it to 400 dpi. Graph 4a also includes the fax images. If the data point for a particular resolution is missing, then the OCR system had difficulty processing the images scanned at this resolution: either it made excessive failures or its accuracy was less than 90%.

Decreasing the resolution from 300 to 200 dpi caused a substantial increase in the number of
errors: approximately a 50% increase for the Business Letter and DOE Samples, and a 75% increase for the Magazine and English Newspaper Samples. The number of errors jumped by 200% for the Spanish Newspaper Sample.

One would expect the number of errors to decrease by increasing the resolution from 300 to 400 dpi. In some cases it did by a small amount, but just as often, the number of errors increased. The higher resolution provided little or no advantage.

The fine-mode fax images have essentially the same resolution as the 200 dpi images. But when processing the former, the OCR systems made 5 to 15% fewer errors than they did on the latter. Upon inspection, we observed that the images created by the fax machine are “darker,” and contain fewer broken characters, than the images produced using the Fujitsu scanner.

The standard-mode fax images presented a very difficult test for the OCR systems. More than twice the number of errors were made on these images than on the fine-mode images, and only two systems, Caere OCR and the XIS OCR Engine, attained a character accuracy above 90%. Two organizations, Ligature and Recognita, chose not to participate in this test.

2.6 Page Quality Groups

If we process a given page using several OCR systems, and determine the character accuracy of each system on this page, then we can compute the median of these accuracies to obtain a measure of the quality, or “OCR difficulty,” of the page. We use this measure to divide the pages of each sample into five “Page Quality Groups” of approximately equal size. Group 1 contains the pages with the highest median accuracy (best page quality), and Group 5 contains the pages with the lowest median accuracy (worst page quality). In Graphs 5a-5g, the character accuracy within each group is plotted to show the effect of page quality.

A large percentage of the errors, about 50 to 60%, are made on the worst 20% of each sample, i.e., Group 5. For the DOE Sample, this percentage is even higher, roughly 70 to 80%. We can gain some insight into what makes OCR difficult by examining the images of pages belonging to Group 5. In Figures 1-5, we present snippets taken from these images. Each was reproduced from the 300 dpi binary image, and enlarged by 50% to make it easier to see the degradation.

Broken and touching characters, also known as “splits” and “joins,” are a very common source of error, and occur in each test sample. When processing the business letters, the OCR systems encountered some difficulty reading letterheads, which are often printed in a stylistic manner; also, creases in the hard copy affected the recognition of entire lines. The DOE Sample contains many challenging tables, and some pages with skewed text and/or curved baselines. Text printed on a shaded background is common in magazine articles, and is a significant source of error. Bleedthrough and other irregularities of newsprint caused some speckling of the newspaper images.

3 Word Accuracy

A popular application of OCR is to build a text database from a collection of hard-copy documents. Information retrieval techniques can then be applied to locate documents of interest. In this environment, the percentage of words that are correctly recognized, or word accuracy of the
Word Accuracy

Table 4a: Original Business Letters

<table>
<thead>
<tr>
<th></th>
<th>Misrec. Words</th>
<th>300 dpi Binary</th>
<th>300 dpi 8-bit Gray Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Word Accuracy</td>
<td>Stopword Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>1,144</td>
<td>97.78</td>
<td>98.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.46</td>
<td>99.35</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>3,654</td>
<td>92.90</td>
<td>95.72</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>1,631</td>
<td>96.83</td>
<td>98.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97.09</td>
<td>98.47</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>1,990</td>
<td>96.13</td>
<td>98.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.47</td>
<td>96.14</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>2,621</td>
<td>94.91</td>
<td>96.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97.09</td>
<td>98.47</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>1,578</td>
<td>96.93</td>
<td>98.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.47</td>
<td>96.14</td>
</tr>
</tbody>
</table>

Table 4b: Fax Business Letters

<table>
<thead>
<tr>
<th></th>
<th>Misrec. Words</th>
<th>Standard-mode Fax</th>
<th>Fine-mode Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Word Accuracy</td>
<td>Stopword Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>4,643</td>
<td>90.98</td>
<td>95.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96.12</td>
<td>98.11</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>4,909</td>
<td>90.46</td>
<td>94.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>94.47</td>
<td>98.23</td>
</tr>
</tbody>
</table>

Table 4c: DOE Sample

<table>
<thead>
<tr>
<th></th>
<th>Misrec. Words</th>
<th>300 dpi Binary</th>
<th>300 dpi 8-bit Gray Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Word Accuracy</td>
<td>Stopword Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>9,386</td>
<td>95.60</td>
<td>98.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96.11</td>
<td>98.61</td>
</tr>
<tr>
<td>EDT ImageReader</td>
<td>23,350</td>
<td>89.07</td>
<td>93.47</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>7,826</td>
<td>96.34</td>
<td>98.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96.52</td>
<td>99.09</td>
</tr>
<tr>
<td>MAXSOFT-OCR Recore</td>
<td>15,451</td>
<td>92.76</td>
<td>96.49</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>16,674</td>
<td>92.19</td>
<td>95.69</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>9,239</td>
<td>95.67</td>
<td>98.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.47</td>
<td>96.14</td>
</tr>
</tbody>
</table>
### Table 4d: Magazine Sample

<table>
<thead>
<tr>
<th></th>
<th>300 dpi Binary</th>
<th></th>
<th>300 dpi 8-bit Gray Scale</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Misrec. Words</td>
<td>Word Accuracy</td>
<td>Stopword Accuracy</td>
<td>Stopword Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>3,659</td>
<td>96.80</td>
<td>97.89</td>
<td>96.05</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>4,566</td>
<td>96.01</td>
<td>97.47</td>
<td>94.99</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>11,617</td>
<td>89.84</td>
<td>91.95</td>
<td>88.37</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>6,595</td>
<td>94.23</td>
<td>95.79</td>
<td>93.15</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>6,261</td>
<td>94.53</td>
<td>96.26</td>
<td>93.32</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>4,923</td>
<td>95.70</td>
<td>97.32</td>
<td>94.56</td>
</tr>
</tbody>
</table>

### Table 4e: English Newspaper Sample

<table>
<thead>
<tr>
<th></th>
<th>300 dpi Binary</th>
<th></th>
<th>300 dpi 8-bit Gray Scale</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Misrec. Words</td>
<td>Word Accuracy</td>
<td>Stopword Accuracy</td>
<td>Stopword Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>1,181</td>
<td>98.59</td>
<td>99.10</td>
<td>98.24</td>
</tr>
<tr>
<td>HP Labs OCR</td>
<td>1,946</td>
<td>97.68</td>
<td>98.73</td>
<td>96.94</td>
</tr>
<tr>
<td>INM NeuroTalker</td>
<td>13,989</td>
<td>83.35</td>
<td>88.47</td>
<td>79.71</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>3,646</td>
<td>95.66</td>
<td>96.80</td>
<td>94.85</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>2,219</td>
<td>97.36</td>
<td>98.22</td>
<td>96.74</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>2,948</td>
<td>96.49</td>
<td>97.88</td>
<td>95.50</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>1,892</td>
<td>97.75</td>
<td>98.64</td>
<td>97.11</td>
</tr>
</tbody>
</table>

### Table 4f: Spanish Newspaper Sample

<table>
<thead>
<tr>
<th></th>
<th>300 dpi Binary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Misrec. Words</td>
<td>Word Accuracy</td>
</tr>
<tr>
<td>Caere OCR</td>
<td>2,193</td>
<td>96.20</td>
</tr>
<tr>
<td>Ligature CharacterEyes Pro</td>
<td>5,785</td>
<td>89.97</td>
</tr>
<tr>
<td>MAXSOFT-OCRON Recore</td>
<td>5,015</td>
<td>91.30</td>
</tr>
<tr>
<td>Recognita OCR</td>
<td>3,400</td>
<td>94.10</td>
</tr>
<tr>
<td>XIS OCR Engine</td>
<td>2,966</td>
<td>94.86</td>
</tr>
</tbody>
</table>
OCR-generated text, is an important measure.

We define a word to be any sequence of one or more letters. A word is considered to be correctly recognized if all of its letters have been correctly identified. Since full-text searching is normally performed on a case-insensitive basis, a letter that is generated in the wrong case (e.g., C for c) is still considered to be correct.

Tables 4a-4f give the number of misrecognized words and the word accuracy for each test sample. Graphs 6a-6g show the word accuracy within each Page Quality Group.

3.1 Stopwords and Non-stopwords

Stopwords are common words such as the, of, and, in, etc. Since these words have little retrieval value, they are normally not indexed by text retrieval systems. Since users search only for non-stopwords, it is the percentage of non-stopwords that are correctly recognized, or non-stopword accuracy, that is especially of interest.

We utilize the default set of 110 stopwords from the BASISPLUS text retrieval system [IDI 90]. About 40% of the words in the test samples are stopwords. Tables 4a-4e show the stopword and non-stopword accuracy for each test sample. (These were not computed for the Spanish Newspaper Sample.)

3.2 Distinct Non-stopword Accuracy

Suppose a user wishes to find every document in the database that contains a specific term (non-stopword). If a document containing this term is to be located, then at least one occurrence of the term must have been correctly recognized by the OCR system. With this in mind, we introduce a measure called distinct non-stopword accuracy.

For a given page, we refer to each unique term occurring on the page as a distinct non-stopword, and we say that it is correctly recognized if at least one of its occurrences has been correctly identified. Distinct non-stopword accuracy is the percentage of distinct non-stopwords that are correctly recognized.

It has been argued that full-text searching is resilient to OCR errors because of the inherent redundancy of text. Because a search term may occur many times within a document, it is unlikely that the OCR system will misrecognize every occurrence. Under the assumption of independence, if OCR-generated text has an overall non-stopword accuracy of 90%, then for a term occurring \( n \) times, the probability that every occurrence has been misrecognized is \( 10^{-n} \). A counter-argument can be put forward that an image defect that causes the OCR system to misrecognize one occurrence may be present in every occurrence, making it considerably more likely that every occurrence will be missed.

In Graphs 7a-7f, we show how often at least one occurrence is correctly recognized as the number of occurrences increases from one to four. Contrast these graphs with Graph 7g, which illustrates the expected shape of the curves under the independence assumption.

3.3 Phrase Accuracy

Users also search for documents containing specific phrases. We define a phrase of length \( n \) to be any sequence of \( n \) words. A phrase is correctly recognized if all of its words have been correctly
identified. *Phrase accuracy* is the percentage of phrases that are correctly recognized. Graphs 8a-8g show the accuracy for phrases of lengths one through eight. Note that the phrase accuracy for length one is equal to the word accuracy.

Phrase accuracy provides a useful measure of “error bunching.” Given two OCR systems with the same word accuracy, the one with the lower phrase accuracy has produced errors that are more widely dispersed throughout the text.

4 Marked Character Efficiency

Finding and correcting errors in OCR-generated text can be a tedious and expensive process for users. But an OCR system provides some assistance by flagging the generated characters that it believes are most likely in error. A *reject character* (~) is placed in the output when the OCR system is unable to recognize a character. Also, the system may place a *suspect marker* (∧) before each character that is generated with low confidence. We refer to reject characters, and characters marked as suspect, as *marked characters*.

This sentence contains reject characters and suspect markers.

In the above sentence, there are three marked errors: the two reject characters and the “l” in “contains.” The second “a” in “markers” is an unmarked error. The “d” in “and” is a false mark, which is a correctly-generated character that is marked as suspect.

In marked character efficiency, we measure the utility of the marked characters produced by an OCR system. Graphs 9a-9c display curves that show how the character accuracy of the OCR-generated text increases as a human editor examines more and more marked characters and corrects the marked errors. Initially, this process is very efficient as the editor corrects the errors identified by reject characters, and then the first level of suspect markers. But as the number of suspect markers increases, the curves flatten due to an increasing percentage of false marks.

Marked characters make it possible for an editor to inspect only one-half of one percent of the OCR-generated text yet correct 20 to 45% of the errors in the text. The editor may examine more characters than this, but the operation becomes considerably less efficient.

5 Automatic Zoning

In every test discussed so far, the OCR systems were given the coordinates of the text regions to process. In the test of automatic zoning, this information was not provided, and each system was asked to locate the text regions, and determine their correct reading order.

To measure how well this task is performed, we apply an algorithm that estimates the number of character insertions and block move operations needed to correct the automatic zoning errors. If an OCR system does not find a text region, then insertions are required to enter the missing text. If text blocks are out of order, then move operations are needed to re-order them. Using a conversion factor to express each move operation in terms of an equivalent number of insertions, the cost of correction is ultimately given solely in terms of insertions. Details of this method are presented in [Kanai 95].
Graphs 10a-10c present the results of this test. The cost of correction is plotted for a range of conversion factors, and has been normalized using the number of characters in each sample to allow for comparison across samples. HP Labs OCR and INM NeuroTalker do not support automatic zoning, and are thus missing from the graphs. The other curves that are missing are due to excessive failures. The XIS OCR Engine performed the best overall on this test.

The cost of correction was least for the English Newspaper Sample. Since each article was clipped from the newspaper, the OCR systems contended primarily with a headline above one or more columns of text. But the magazine articles were not clipped, and are part of a page layout that may be very creative (i.e., complicated). The DOE pages presented the challenge of distinguishing tables from multi-column text; the OCR system should “de-columnize” the latter, but not the former.

6 Comparison of OCR Systems: Accuracy and Speed

6.1 Binary Input

The best overall accuracy on binary images was achieved by Caere OCR, HP Labs OCR, and the XIS OCR Engine. We observed no significant differences in accuracy among these systems, with one exception: HP Labs OCR was out-performed by the other two systems on the standard-mode fax images.

But significant differences in speed were evident. The XIS OCR Engine was 2.3 to 4.4 times faster than Caere OCR, depending on the sample. Although Caere OCR and HP Labs OCR operated on different platforms, it is apparent that the former was much faster than the latter. When HP Labs OCR was submitted for this test, an HP Labs representative indicated that this version had not been optimized and was not expected to be competitive in terms of speed.

MAXSOFT-OCRON Recore and Recognita OCR comprise a second tier of systems. Although these two systems were comparable in terms of accuracy, Recognita OCR performed with roughly twice the speed.

A third tier of systems consists of EDT ImageReader, INM NeuroTalker, and Ligature CharacterEyes Pro. An interesting feature of INM NeuroTalker is the ability to adjust the trade-off of speed versus accuracy. At the request of INM, we began to test this system using the setting for highest accuracy (and slowest speed). But after encountering too many failures for the system to remain in the test, at INM’s request we changed the setting to one that decreased the accuracy, increased the speed, and circumvented the failures. (INM attributes the failures to a faulty DOS extender that was shipped with the system.)

Since the Magazine Sample was used in the third annual test, it can serve as a yardstick to measure progress in the past year. Four organizations participated in both the third and fourth annual tests: Caere, EDT, Recognita, and XIS. In this year’s test, the systems from Caere and XIS made 27 to 28% fewer errors on this sample than their predecessors made a year ago. The Recognita version made 6% more errors this year than last year. No comparison was possible for the EDT version because of failures.
6.2 Gray Scale Input

Caere OCR and HP Labs OCR utilize fundamentally different approaches when processing a gray scale image. HP Labs OCR first “binarizes” the image, i.e., creates a binary image from the gray scale image, and then recognizes the characters on the binary image. On the other hand, Caere OCR recognizes the characters directly from the gray scale image.

It is clear that gray scale input offers advantages over binary input in terms of accuracy. Depending on the sample, these systems produced 10 to 40% fewer errors when given gray scale input. But there was one exception: Caere OCR made substantially more errors when processing the gray scale images of the English Newspaper Sample. This may have been caused by bleedthrough, which is often visible in the gray scale image of a newspaper, but mostly disappears after the binarization process.

It appears that gray scale input was of greatest value when recognizing text printed on a shaded background, which is commonly found in the Magazine Sample. The binary image of this text is usually speckled and problematic, but given a gray scale image, an OCR system has a much better chance of separating the text from its background.

Gray scale images require considerably more storage and take longer to process than binary images. On average, Caere OCR needed twice as much time, and HP Labs OCR needed 20% more time, to process the gray scale input.

7 Summary

Eight organizations submitted OCR systems for the fourth annual test. These systems processed bitmapped images of more than 1,500 pages of business letters, scientific documents, and magazine and newspaper articles. The analytic tools of the OCR Experimental Environment were used to compare the OCR-generated text with the correct text, and compute several measures of performance. These include character, word, non-stopword, and phrase accuracy, and three new measures: throughput, accuracy by character class, and distinct non-stopword accuracy. Also, the effect of page quality was observed, the utility of marked characters was gauged, and the cost of correcting automatic zoning errors was estimated.

The accuracy of the OCR systems declined dramatically when the resolution of the images was reduced from 300 to 200 dpi; however, little or no benefit was obtained by increasing the resolution to 400 dpi. Fax images, especially the standard-mode variety, presented a significant challenge. The improved accuracy obtained from gray scale input demonstrates that this is an important new direction.

With the assistance of confidence intervals, we partitioned the eight systems into three tiers based on the accuracy results for binary input. We then noted differences in speed within tiers. This “ranking” reflects the performance of these systems on this test; their relative performance may vary when processing other types of documents, or when processing similar documents under different test conditions. Finally, we wish to emphasize that ISRI does not endorse any particular OCR system or systems.
Acknowledgment

We acknowledge the help of Dr. Ashok Singh and Dr. George Nagy in establishing the confidence interval method for character accuracy. Also, “Accuracy by Character Class” was suggested by Junichi Kanai.

References


Figure 1: Examples from Page Quality Group 5, Business Letter Sample

EMBASSY OF COMMERCIAL CHICAG

quarter page adver your support again your convenience.

organization receiving expert information on research and maintains

Please forward them to returned to us together

I would highly rec organization. Sh

Don’t miss out on the latest fyi/im newsletter and the So

Incentive Award - A mid FAXed images with a sys

Service Representatives: sin

We have enclosed a C

P.S. This enroll by June 15, 1993 a

With Discover Card yo Insurance*. Every ti flight insurance auto charging your airline

$49.95! And this is not "display quality"--then Completely scalable, For

 month after your account is opened, and may waived first year; $40 each year thereafter for minimum, $25 maximum. Transaction fee for r transaction fees for two special Premium Acce
Operation of pressure transducer tensiometers in an infil-

Erosional cuts on streams; sand dunes at Kings Beach

LEWELLEN, W. S., "The of the Tornado Vortex," Pr.
Symposium on Tornadoes:

In reactor fuel elements a approximately unity is not un-
mixing is to be expected and.

Activity provides sit data that may have sit final design of repos

1057 37.4 119.0
401 33.7 118.1
230u 37.6 118.9

Figure 37.—Salt pool and collapse struc northwest of Badwater. Drawn t

Composition

engineered barriers must be desi

The ground water protection requirements (40 CFR 191.18) for the quality of any "special sourc

9. Precision

9.1 Criteria for judging the a
the maximum density and opti

Fynon-Juniper woodland is
Mountains or Slate Range; but th
7,000 feet in the Panamint Mount
Lake and in the northern Argus M
Lake. Associated with this zone

Acanthite, summary of thermodyna
data, 195
Activity of aqueous species, 36, 37,
93-96, 98, 114, 141, 143 145, 147

6 to 60 are shown, as well as the p
and permeabilities for samples of si
varying according to equation 26 to
sample volume of 1035 cm cubed i

12 3:48:56.12 CSNL
20 13:15:47.38
20 13:18:44.58
MAY 6 6:17:13.99

SE of Hawthorne, Nev. 6.5 38.3
Parkfield 6.0 36.9
Southern California 5.1 34.1
N of Bishop 5.0 37.5

would bring about dehydrati acid sites (i.e. surface alumini ons). The sites which chemi

consideration for the heterogeneities and 2 thermally induced sa
Figure 3: Examples from Page Quality Group 5, Magazine Sample

19

achieved the same results, but they do it?

Answer: In years past, the seeds were fit with a considerable a

GALLONS USED
DAILY BY A FAMILY

the cheapest plastic lenses with coating. The prices they were q
by as much as 75 percent from

PARKS
# Children under 3 admitted free. * Prices do not inclu
Walt Disney World Resort #
Lake Buena Vista, Florida; (407) 824-4321

years ago, in a group-th session, I heard a story
have never forgotten.

on scientific research, said Dr.
ic McDuffie, former senior vid
ent for medical affairs for the tis Foundation.

independent front and rear suspension, tric damping, and speed-sensitive rack-and
steering. These underpinnings allow Cad

Taste buds take note. Tray Gour
Your Own Chef in the College C
(Lake Isle Press, $10.95) is the book
cooked on the stovetop kitchen comfortably coo

ter. Still, much remains mysterious: Vin
cent? The patient had no French ance
vented from his hometown of Worc

Style. Liquid (L) or powder (P).
Calories. Per eight-fluid-ounce
Carbohydrates. Percent by we

Summer's fresh bar saucers and marina

TOP 10 TAPE RENTALS
UNFORGIVEN Clint Eastwood, Warner........
THE BODYGUARD Kevin Costner, Warner....

Opry, Garth Brooks met backstag with cancer pa
tient Libby Sharp of Gatlin-

WICKENBURG, ARIZONA
The Meadows
A leading treatment center for
TALK OF THE DAY:
“Names change for political, geographical or

Most likely to be stolen:
Volkswagen Cabriolet
Ford Mustang convertible
Cadillac DeVille two-door

keeper predicted yesterday he
will be indicted on embezzlement
charges, signaling a glitch in

because of who he wants them,
The distinctive marks turn of
spective employers, threatenin
t life he plans away from the st

and conversations
with world leaders

cholera in the crowded Kwan

camps of eastern Zaire is go
t United Nations said yesterday

But as the threat from ch

Bellize, a small Central Amer
country, use plant medicines,
losses of forests, and a lack of
est among young people in bel

ing a “large number of casua

spokesman Maj. Dacre Hollow

In response, U.N. officials w

By SUSAN MILLIGAN
News Washington Bureau

WASHINGTON — With Vi

Previous positions: Chairman of
Cisneros Asset Management Co.
1989-92; Mayor of San Antonio,
1981-89; San Antonio City Council,

The inquirer wants its news report
be fair and correct in every respect
you have a question or comment
about news coverage, write to

Washington Post

THIMPU, Bhutan — Its cit
seldom write letters and there

neighbors and then heading the
the hill toward a new life.

Federal and state officials say
him a lot

herefore unlabeled wines: Ch
C
eau Margaux; Chateau Lafite; Re
ert Mondavi cabernet sauvignon

The village must rely on
the McHenry County Sheri
Department in Wood
tock to enforce the limit.

unemployment rate decreased to
the work force in March from 8.3
uary, but rose from 7.1% in Ma
accord to SCB, the national
nández exclamó indignado que la dor bonaerense "es volver al 49", constitutional sancionada ese año, dente Juan Domingo Perón que lo

**Pensiones sociales**

La Caja de Pensiones So-
les-Ley 5110, delegación Rosa

-¿Declararán cuando se
llame a indagatorias?
-No. Existe el derecho a caln
eso no tiene nada que ver con la

**IXTAPALUCA** (Notimex),
enfrentamiento entre un grupo
muneros y agentes de la Dir-
General de Seguridad Públ

-De acuerdo con el plan de
la compañía, que aún no se de
totalidad, la intención es dive
servicios con la construcción

**BARCELONA.** (Agencias.)
-dirigentes de CiU y del PP
cambiaron ayer duras críticas

**El histórico via**

---

Yemen del Sur, aprovechan
una breve "tregua" para
fumar un cigarrillo. (Reuter)

El extenso debate por
Constituyente, ganad
la posición conjunt

En declaraciones radiales
senador porteño señaló que el
informe elaborado en el am
de la Auditoría General de

dospéchos. Desde esferas socia
políticas se consideró que el a
había sido Guillermo Luque, hijo
entonces diputado nacional por e

Un, el doctor Jorge Carpizo
a. Porfirio Muñoz Ledo,
croncos del PRD por el

Presidencia de la República
Zedillo, en materia económica
está siendo difundida por a

---

Y TIEMPO LIBRE

El famoso paseo a pie entre
Martí d'Empúries que cantó el
sep Pla es lo que debía ser. un p
1 Character Accuracy

1a: Original Business Letters

1b: Standard-mode Fax Business Letters

1c: Fine-mode Fax Business Letters

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- INM NeuroTalker
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine
2 Throughput

2a: Original Business Letters

2b: Standard-mode Fax Business Letters

2c: Fine-mode Fax Business Letters
3 Accuracy by Character Class

3a: Original Business Letters

3b: Standard-mode Fax Business Letters

3c: Fine-mode Fax Business Letters

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- INM NeuroTalker
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine
4 Effect of Resolution

![Graph showing effect of resolution on OCR accuracy for different OCR tools.](image)

Legend:
- Caere OCR
- EDT ImageReader
- HP Labs OCR
- INM NeuroTalker
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine

4a: Business Letter Sample

**% Character Accuracy**

- 90
- 91
- 92
- 93
- 94
- 95
- 96
- 97
- 98
- 99
- 100

**Resolution**
- Standard-mode Fax
- Fine-mode Fax
- 200 dpi Original
- 300 dpi Original
- 400 dpi Original
5 Character Accuracy vs. Page Quality

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- INM NeuroTalker
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine

5a: Original Business Letters

5b: Standard-mode Fax Business Letters

5c: Fine-mode Fax Business Letters
6 Word Accuracy vs. Page Quality

6a: Original Business Letters

6b: Standard-mode Fax Business Letters

6c: Fine-mode Fax Business Letters
7 Distinct Non-stopword Accuracy

7a: Original Business Letters

7b: Standard-mode Fax Business Letters

7c: Fine-mode Fax Business Letters

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- INM NeuroTalker
- Ligature CharacterEyes Pro
- MAXSOFT-OCR Oron Recore
- Recognia OCR
- XIS OCR Engine
8 Phrase Accuracy

8a: Original Business Letters

8b: Standard-mode Fax Business Letters

8c: Fine-mode Fax Business Letters
9 Marked Character Efficiency

- Caere OCR
- Caere OCR (gray scale)
- EDT ImageReader
- HP Labs OCR
- HP Labs OCR (gray scale)
- Ligature CharacterEyes Pro
- MAXSOFT-OCRON Recore
- Recognita OCR
- XIS OCR Engine

9a: Original Business Letters

9b: Magazine Sample

9c: English Newspaper Sample
10 Automatic Zoning

10a: DOE Sample

10b: Magazine Sample

10c: English Newspaper Sample