A Report on the Accuracy of OCR Devices

Stephen V. Rice, Junichi Kanai and Thomas A. Nartker
Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway
Las Vegas, NV 89154-4021

March 4, 1992

1 Introduction

The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV) has conducted an experiment to determine the accuracy of six commercially-available OCR devices:

Caere OmniPage Professional
Calera RS 9000
ExperVision TypeReader
Kurzweil 5200
Recognita Plus
Toshiba ExpressReader

In addition, the accuracy obtained by the ISRI Voting Algorithm, which makes use of these six devices, was computed. This algorithm is a fully-automated and extended version of the method described in [1].

This study evaluates just one aspect of the OCR products, i.e., their accuracy. There are other important features that were not evaluated, such as their user interface, and their automatic zoning capabilities. The selection of an OCR product should not be based on accuracy alone.

2 Test Data

Test data consisted of 240 pages that were selected at random from the GT1 database [2]. For each page in the GT1 database, there is a 300 dpi binary image file. These images were produced under conditions that are typical for a large-scale data conversion operation. The operators who performed the scanning had adequate training, but were not "experts."

In each page, regions containing "main body" text were manually zoned, and correct text corresponding to each zone was carefully prepared. Regions containing the following were neither zoned nor used in this experiment:

- tables, figures, and their captions
- footnotes
- page numbers, headers and footers
- mathematical equations

Due to the technical nature of the documents in the GT1 database, 108 of the selected pages contained no "main body" text. Therefore, the remaining 132 pages, containing a total of 242 zones and 278,786 characters, were used to test the accuracy of the OCR devices.

3 Methodology

3.1 Settings

Each device processed exactly the same zoned portions of the same images. All processing, including the determination and tabulation of errors, was carried out entirely under computer control, i.e., there was no human interaction with the devices during the experiment. The software tools that were used are part of the ISRI OCR experimental environment [3].

None of the devices received any special "training." No "de-columnization" was required since each zone contained only a single column. For each device supporting an optional system lexicon feature, this feature was enabled. These include the Calera RS 9000, the Kurzweil 5200, and the Toshiba ExpressReader.

The ISRI Voting Algorithm also processed exactly the same data without any human interaction. The voting algorithm is based on automatic synchronization of the outputs of the participating OCR devices by means of a string-matching algorithm.

3.2 Error Counting

Accuracy was determined on a character basis. Each character insertion, substitution, or deletion required to correct the generated text was counted as an error. Any "reject characters" that were generated were not treated specially, but were counted as errors.

Differences in the horizontal and vertical spacing produced by the devices, however, were first eliminated. All blank lines, and all leading and trailing blanks on a line, were discarded. In addition, consecutive blanks within a line were compressed to a single blank.

A total of 210 non-ASCII characters, 157 subscripts, and 127 superscripts, were found on the 132 pages. To avoid penalizing a device for failing to recognize these, the error-checking software allowed any characters to be generated for these without counting them as errors.

4 Results and Analysis

Table 1 shows the results for the entire 132-page sample containing 278,786 characters.

	# Errors	% Accuracy
Caere OmniPage Professional	8841	96.83
Calera RS 9000	3709	98.67
ExperVision TypeReader	6318	97.73
Kurzweil 5200	4716	98.31
Recognita Plus	11282	95.95
Toshiba ExpressReader	12169	95.64
ISRI Voting Algorithm	1867	99.33

Table 1. Accuracy Statistics for the Entire Sample

Given a page, and the accuracy obtained by each of the six devices for the page, the median accuracy can be determined. This is a good measure of the quality of a page-image, or at least its "OCR difficulty." The 132 pages were sorted by this measure, and divided into three groups containing approximately the same number of characters:

Best	-	39 pages containing 93,016 characters	(highest median accuracy)
Middle	-	40 pages containing 93,586 characters	
Worst	-	53 pages containing 92,184 characters	(lowest median accuracy)

Table 2 shows that a large proportion of the errors are made on a small number of pages. Figure 1 shows that there are significant differences in the accuracy of commercially-available OCR devices, especially when processing poor quality pages.

The worst pages were examined by the authors. These images were generated from poor quality photocopied pages. Figure 2 shows examples of poor quality characters on these pages. These results suggest that OCR research should focus on the recognition of poor quality characters.

	Best		Middle		Worst	
	# Errors	% Accuracy	# Errors	% Accuracy	# Errors	% Accuracy
Caere	310	99.67	895	99.04	7636	91.72
Calera	268	99.71	686	99.27	2755	97.01
ExperVision	262	99.72	1117	98.81	4939	94.64
Kurzweil	212	99.77	767	99.18	3737	95.95
Recognita	809	99.13	1521	98.37	8952	90.29
Toshiba	659	99.29	1679	98.21	9831	89.34
Voting	54	99.94	256	99.73	1557	98.31

Table 2. Accuracy Statistics Based on Page Quality

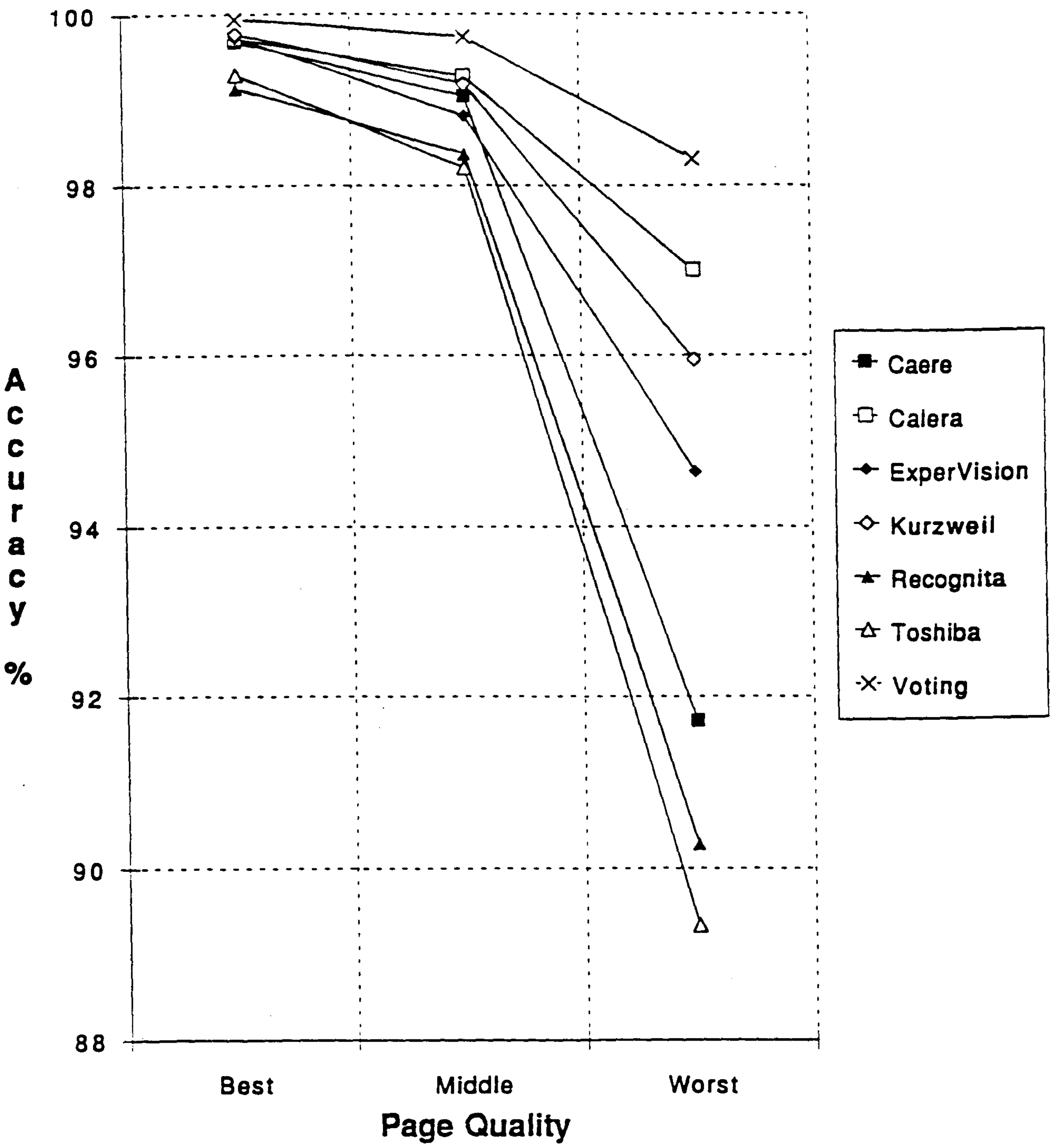
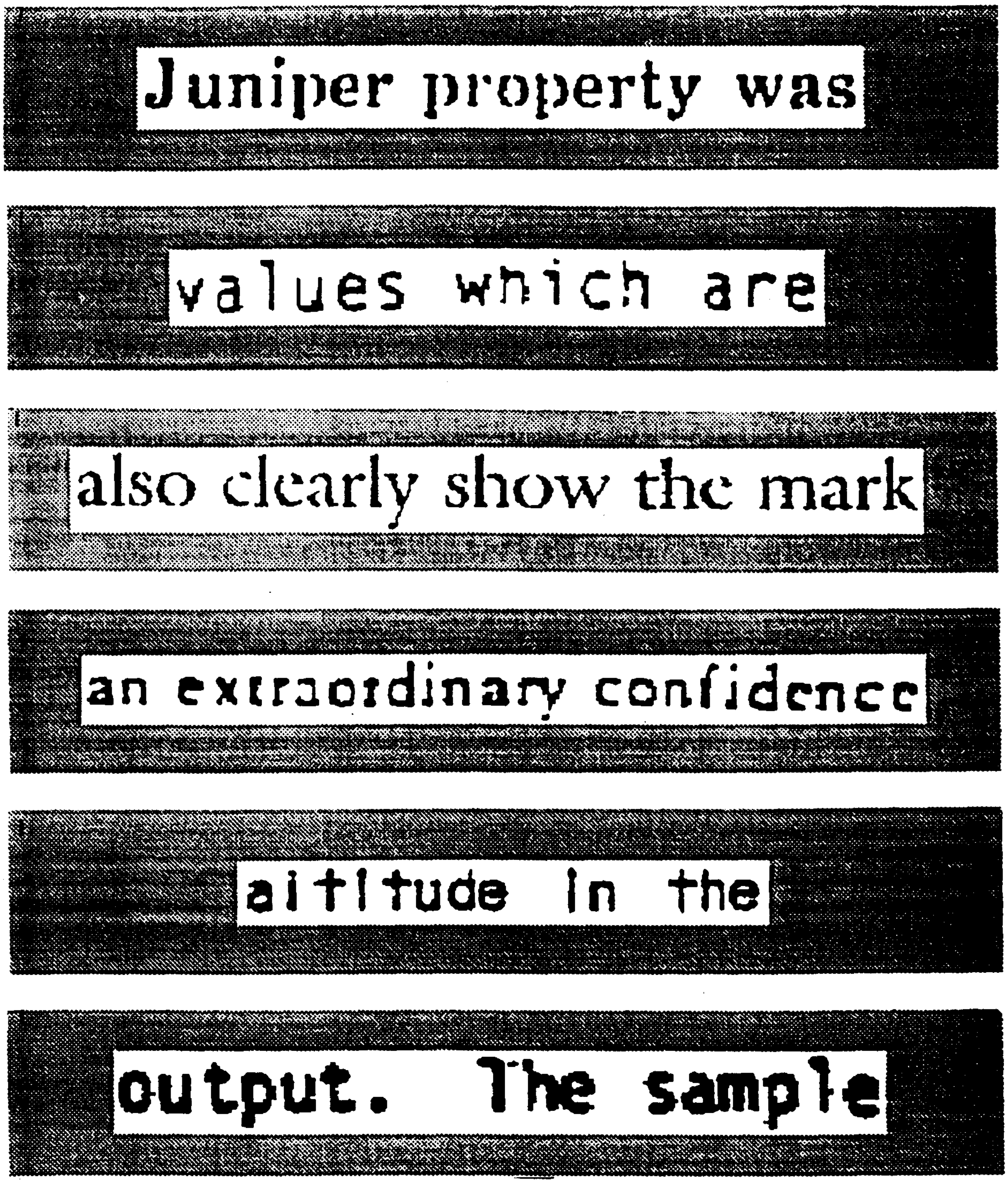


Figure 1. Accuracy Statistics Based on Page Quality



Juniper property was

values which are

also clearly show the mark

an extraordinary confidence

altitude in the

output. The sample

Figure 2. Examples of Poor Quality Characters (3x linear magnification)

Table 3 shows that the ISRI Voting Algorithm corrected a significant percentage of the errors made by the individual OCR devices. Its correction rate depends on the performance of the individual devices participating in the voting process.

Page Quality	Best	Middle	Worst
Errors made by the Voting	54	256	1557
Errors made by the Best Device	212	686	2755
Errors Corrected by the Voting	75%	63%	43%

Table 3. Errors Corrected by the ISRI Voting Algorithm

5 Conclusion

This study evaluated just one aspect of the OCR products, i.e., their accuracy. When these OCR devices processed good quality page-images, all products were able to recognize almost all characters correctly. However, there are significant differences in their accuracy on poor quality pages.

The ISRI Voting Algorithm was able to correct many of the errors made by the individual devices. Its correction rate is highest on high quality input.

ISRI plans to study further the accuracy of OCR devices as well as other aspects of these products. Additional experiments involving new devices (and new versions of these six devices, as they become available) will be undertaken using this test data and other data sets.

References

[1] R. Bradford and T. Nartker, "Error Correlation in Contemporary OCR Systems," *Proc. First International Conference on Document Analysis and Recognition,* St. Malo, France, Sept. 1991, pp. 516-523.

[2] T. Nartker, R. Bradford, and B. Cerny, "A PRELIMINARY REPORT ON UNLV/GT1: A Database for Ground-Truth Testing in Document Analysis and Character Recognition," *Proc. Symposium on Document Analysis and Information Retrieval,* Las Vegas, NV, March 1992, pp. 300-315.

[3] S. Rice, *The OCR Experimental Environment,* Technical Report, Information Science Research Institute, University of Nevada, Las Vegas, March 1992.