Systematic bias in OCR experiments*

Daniel Lopresti Andrew Tomkins Jeffrey Zhou Jiangying Zhou

Matsushita Information Technology Laboratory
Panasonic Technologies, Inc.
Two Research Way
Princeton, NJ 08540-6628

ABSTRACT

In this paper, we examine the effects of systematic differences (*bias*) and sample size (*variance*) on computed OCR accuracy. We present results from large-scale experiments simulating several groups of researchers attempting to perform the same test, but using slightly different equipment and procedures. We first demonstrate that seemingly minor systematic differences between experiments can result in significant biases in the computed OCR accuracy. Then we show that while a relatively small number of pages is sufficient to obtain a precise estimate of accuracy in the case of "clean" input, real-world degradation can greatly increase the required sample size.

Keywords: optical character recognition, OCR accuracy, systematic bias, variance, sample size.

1 INTRODUCTION

As OCR software continues to improve, small differences in accuracy become more important in evaluating system performance. Current OCR packages can often achieve recognition rates exceeding 99% on "clean" input.[8] Under these circumstances, an increase from 99% to 99.4% is clearly significant in that the number of errors is reduced by 40%, yet it also seems quite modest when viewed in absolute terms. Such a change could be indicative of an appreciably better OCR process. However, it could also be brought about by differences in the equipment and procedures used in the experiments (*bias*), or by random chance if the sample space is not large enough (*variance*). Clearly, factors that introduce measurable bias and/or variance into OCR accuracy rates must be understood and, whenever possible, controlled for performance evaluations to be considered statistically significant.

In this paper we study this issue, using data from large-scale tests to analyze the degree to which bias and variance arise under certain real-world conditions. Our model is one of several groups of researchers attempting to perform the same experiment (as might be described in a published paper, for example), but using slightly different equipment and procedures. We first show that minor systematic differences between experiments result in significant biases in the computed accuracies. For "clean" input, we find that a relatively small number of pages is sufficient to obtain a precise estimate of OCR accuracy. That is, in these cases the sampled accuracies converge rapidly, but to different values. For degraded input (photocopies), the variances are high enough that

* Presented at *Document Recognition II (IS&T/SPIE Electronic Imaging)*, February 1995, San Jose, CA.

hundreds or even thousands of pages may have to be OCR'ed to compute a reliable accuracy figure, and, on top of this, the same systematic biases are still present. These results suggest that great care must be taken when using published recognition rates as the basis for comparing OCR algorithms and systems.

2 OCR EXPERIMENTS

Broadly speaking, OCR experiments can be conducted in either of two settings: "clean" or "real-world." Clean tests, using high-quality original pages, provide an upper-bound on performance, allowing focus on the core OCR process (character segmentation and classification) without having to worry about more complex document layouts, physical damage, etc. Real-world tests attempt to duplicate the sorts of formatting and degradation problems encountered in practice (*e.g.*, multiple columns, photocopies, faxes, skew). That these two scenarios should yield differing OCR accuracy rates is not surprising, however even more mundane matters such as the particular printer and scanner models used can have an impact, as shall be shown.

We begin by describing the parameters involved in a typical OCR experiment. While some of these are commonly specified in published studies, others are sometimes ignored (making it that much more difficult to reproduce the reported results). We then provide a detailed description of the particular experiments we performed in examining the amount of bias and variance in computed OCR accuracies.

The following parameters usually receive some mention in the literature and are relatively easy to reproduce:

Text It is well-known that a large percentage of OCR errors are context-dependent. There are a number of standard source texts currently available for use (*e.g.*, the Brown Corpus, famous novels in the public-domain such as *Moby-Dick*).

Logical Font Although many commercial packages employ omnifont technology, their performance varies significantly depending on the specific font used.

Page Quality As noted above, the quality of the input page has a big impact on OCR accuracy. This is a subjective measure, however, and sometimes difficult to control as the same printer or copier can produce variable-quality output depending on its toner level, state of repair, etc.

OCR Package Of fundamental importance, of course, is the OCR package used, including any version number and relevant user-defined settings.

On the other hand, certain other parameters are often neglected in published papers, and even when specified are likely to be ignored for reasons of cost and expediency (*e.g.*, it may be difficult to justify buying a new scanner solely for the purpose of exactly duplicating the hardware set-up used by another research group):

Formatting While the text itself may be standardized, there could be differences in the way it is formatted. For example, one site might choose to start each chapter of a novel on a new page, while another might not. Section headings might be placed on separate lines, or integrated with the running text. Line breaks, inter-character spacing (kerning), the use of italics and boldface, and the number and placement of text columns can all vary from test to test.

Physical Font All versions of a font with the same name are not necessarily equivalent. The source of the font family being used could differ depending on the host system, text formatting software, and printer. For example, text typeset in Times font using LaTeX on a Unix system looks appreciably different from that using a word processor on a personal computer.

Printer The choice of the printer for preparing the test data (*i.e.*, the document pages) can determine to a large degree the quality of text. Factors ranging from the resolution (*e.g.*, 600 dpi versus 300 dpi) to the state of the toner cartridge (new to almost depleted) may all vary. Figure 1 shows the same word printed on four different laserprinters and scanned at a resolution of 1,600 dpi. In a published study, all of these might be considered "clean, original-quality output." There are, however, obvious visible differences that will have an impact on OCR accuracy, even between the two 300 dpi printers. Moreover, these variations are not limited to the "jagginess" of the character outlines, but also seem to involve the shapes of the characters as well (*e.g.*, the curved stroke in the 'h' strikes the vertical stroke further up in LaserJet III output than in the other cases).

Scanner Scanners also vary from manufacturer to manufacturer. Some factors that may influence the result of OCR experiments include the resolution (both user-specified and hardware-determined), binarization threshold, and operator attention to skew and misfeeds. In addition, the quality of the autofeeder can also have a significant effect on the error behavior, as we have observed previously.[3] Even leaving a page flat on the glass and scanning it twice in rapid succession can result in different error distributions, a phenomenon that can be exploited to improve classifier performance in some cases.[6,11]

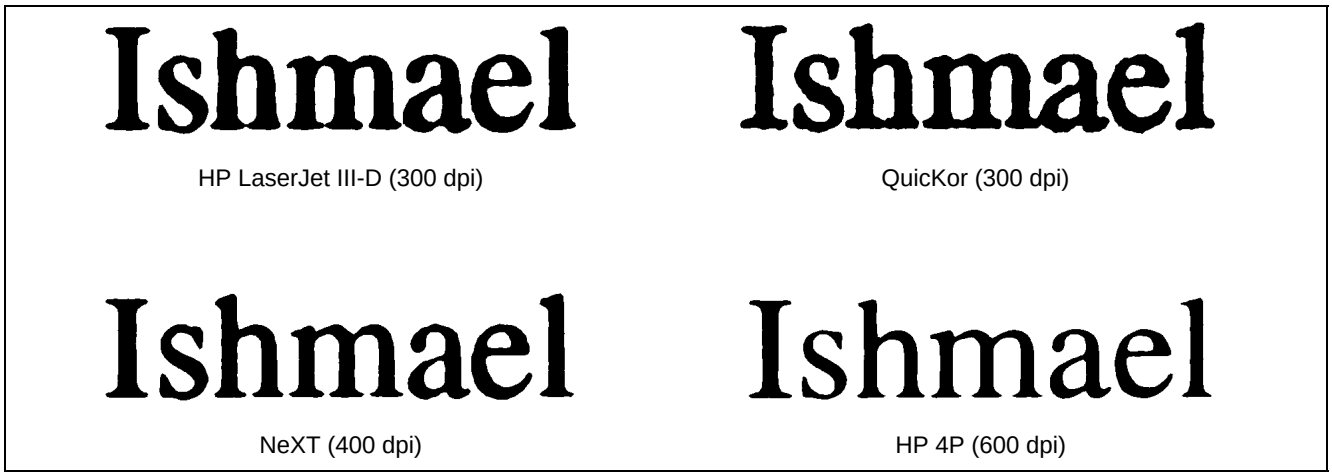

Figure 1: Output variation across the printers used in the experiments.

So that we could study bias and variance in the context of OCR accuracy, we conducted a number of experiments using different parameter settings. For each test, we printed, scanned, and OCR'ed the Hendricks House edition of Herman Melville's famous novel, *Moby-Dick*. This text totals 1,179,194 characters, and requires just over 300 pages when typeset in 10-point Times. The OCR package we employed was OCRServant, version 2.03, release 2.1, running on a NeXT workstation.

In our initial tests, we focused on character recognition errors in a clean source document. To insure some measure of consistency, we examined each laserprinted page for gross defects such as streaking, smudging, toner too low, etc. When necessary, pages were reprinted until the output was deemed to be of acceptable quality. Several factors were held constant across all of these experiments, as indicated in Table 1. As noted above, these are meant to correspond to the kinds of parameters reported in published studies (and hence reproducible by other researchers attempting to duplicate the results).

We also introduced several seemingly minor differences between the three clean-document tests. These are listed in Table 2, and correspond to parameters that might be overlooked in the reporting of an experiment, or are difficult to reproduce precisely for reasons discussed earlier.

We also carried out a similar series of experiments using degraded documents. For each of these tests we printed, degraded, scanned, and OCR'ed the same edition of *Moby-Dick*. While controlled degradation could

Experiment	Text	Font	Page Quality	OCR Package
$E1$, $E2$, $E3$	*Moby-Dick*	10-point Times	Original	OCRServant, ver. 2.03

Table 1: Factors held constant across the three clean-document experiments.

Experiment	Text Format	Printer	Scanner
$E1$	WriteNow	NeXT 400 dpi	Ricoh IS60/UX 300 dpi
$E2$	LaTeX	QuicKor 300 dpi	Microtek MRS-600ZS 300 dpi
$E3$	PostScript strings	NeXT 400 dpi	HSD SX-600 300 dpi

Table 2: Factors varied between the three clean-document experiments.

have been accomplished using a document image defect model,[1] we felt it was more appropriate for the purposes of our study to turn to an authentic source, in this case a Panasonic FP6070 photocopier. By lowering the copier's contrast setting and using its auto-feeder, we were able to introduce a relatively uniform degree of real-world damage across all of the pages. As before, certain factors were held constant across the three experiments involving this data, as shown in Table 3.

Experiment	Text	Font	Page Quality	OCR Package
$E4$, $E5$, $E6$	*Moby-Dick*	10-point Times	Uniformly Degraded	OCRServant, ver. 2.03

Table 3: Factors held constant across the three degraded-document experiments.

Again, we varied the text-formatting software, the printer, and the scanner, as indicated in Table 4.

Finally, in preparation for these last experiments, we inadvertently generated a set of pages that were not uniformly degraded. Because the toner cartridge in the copier was fairly old, damage was distributed across the pages in an almost random-looking way (certain lines and characters came out much lighter than others). As might be expected, this effect became more pronounced as time progressed. Our inability to control the degradation across the hundreds of copies we needed for our three separate tests made it impossible to consider using this data for direct comparison. However, in some sense it provides an even better representation of a "difficult," real-world document. With regards to the text format, printer, and scanner, this experiment was otherwise identical to $E4$, and hence we refer to it as $E4'$.

3 ANALYSIS

In all of our tests, we classified the resulting OCR errors using an approach based on the dynamic programming algorithm for string edit distance.[2,3] To review briefly, if $S = s_1 \ldots s_m$ is the source (original) string, $R = r_1 \ldots r_n$ is the recognized (OCR) string, and $d_{i,j}$ is the distance between the first i characters of S and the first j characters of T, then the traditional recurrence is:

$$d_{i,j} = \min \begin{cases} d_{i-1,j} + c_{del}(s_i) \\ d_{i,j-1} + c_{ins}(r_j) \\ d_{i-1,j-1} + c_{sub}(s_i, r_j) \end{cases} \qquad 1 \le i \le m,\ 1 \le j \le n \tag{1}$$

Experiment	Text Format	Printer	Scanner
$E4$	PostScript strings	NeXT 400 dpi	Microtek MRS-600Z 300 dpi
$E5$	LaTeX	HP 4P 600 dpi	Ricoh IS60/UX 300 dpi
$E6$	WordPefect	HP LaserJet III-D 300 dpi	HP ScanJet IIcx 300 dpi

Table 4: Factors varied between the three degraded-document experiments.

When Equation 1 is used as the inner-loop step in an implementation, the time required is $O(mn)$ where m and n are the lengths of the two strings. If, in addition, the choices that lead to the minimums above (*i.e.*, the optimal decisions) are also recorded, the resulting trace-back table provides a sequence of operations that perform the transformation in question. For the OCR problem, these edits can be equated with the errors in the OCR string.

In our application, we use a slightly more general form of the algorithm that allows for arbitrary $p:q$ substitutions. This allows us to handle OCR segmentation errors in a natural way. The modified recurrence is:

$$d_{i,j} = \min\left\{d_{i-g,j-h} + c_{sub_{g:h}}(s_i \ldots s_{i+g-1}, r_j \ldots r_{j+h-1}) \;\mid\; 0 \leq g \leq p, 0 \leq h \leq q\right\} \tag{2}$$

This new computation requires time $O(mnpq)$ where m and n are the lengths of the two strings and p and q determine the maximum allowable size of a multi-sub. For the studies reported in this paper, we used $p = q = 4$.

Since this approach to classifying errors treats any $p:q$ substitution as a single event, we must also define a measure of the damage done to a given string by an OCR error:

$$damage(sub_{p:q}) \equiv \max(p, q) \tag{3}$$

If S is the original string and R is the OCR string, we can extend $damage(S, R)$ in the obvious way to be the sum of the damages for all the errors determined when editing S into R. We can then compute OCR character recognition accuracy as:

$$accuracy(S, R) \equiv \frac{(|S| - damage(S, R))}{|S|} \tag{4}$$

This definition yields accuracy rates consistent with the figures published by other researchers. As an example, an original line with 100 characters that undergoes one deletion and one 2:2 substitution is said to be recognized with 97% accuracy.

For our experiments, we calculated accuracy rates on a page-by-page basis. We performed our analysis in terms of both non-space errors (*i.e.*, errors involving only printing characters) and total errors. We regarded per-page accuracy as a random variable denoted by x, and computed its mean and variance for each experiment. Assuming the distribution is normal, the mean and variance can be estimated using N sample points (*i.e.*, N pages of *Moby-Dick*) using the standard estimators:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{6}$$

where x_i is an observation (sample) of x. The results of these estimates are given in Tables 5 and 6 for clean and degraded documents, respectively. Note that space errors can account for a significant percentage of the total in some cases.

In Table 5 we compare the OCR accuracies for the three clean-document experiments. For non-space errors, the accuracy for $E1$ is 99.8%, while for $E2$ it is 99.2%, a significant difference. In the case of total errors, the

rates also differ by over 0.6%. From another perspective, $E2$ exhibited four times as many OCR errors as $E1$. This result demonstrates that large biases in computed accuracies can be due to seemingly minor differences in the equipment and procedures used to perform the experiments (even when the text, font, and OCR package are held constant).

	Non-Space Errors		All Errors	
Experiment	Mean	Variance	Mean	Variance
$E1$	0.998077	1.122514e-06	0.996989	1.209464e-05
$E2$	0.991556	5.958679e-06	0.990707	6.739605e-06
$E3$	0.996092	3.993925e-06	0.994438	1.407912e-05

Table 5: Accuracies and variances for the three clean-document experiments.

A comparison of the OCR accuracies for the three experiments using degraded documents is given in Table 6. For non-space errors, the accuracy for $E5$ is 99.44%, while for $E6$ it is 97.58%. In the case of total errors, the rates also vary by over 1.5%. As before, this is due solely to the differences in text formatting, printer resolution, and scanner hardware used in the experiments.

	Non-Space Errors		All Errors	
Experiment	Mean	Variance	Mean	Variance
$E4$	0.982676	1.841861e-05	0.981011	2.130920e-05
$E5$	0.994354	7.788323e-06	0.991583	2.093035e-05
$E6$	0.975761	3.775062e-05	0.974281	5.027643e-05

Table 6: Accuracies and variances for the three degraded-document experiments.

We now calculate the maximum discrepancy possible in our estimates for a 90% confidence interval. That is, we require that

$$\Pr\left[|\mu_x - \overline{x}| > \varepsilon\right] < 2\alpha \tag{7}$$

where μ_x is the true mean of x and $\alpha = 0.05$. According to the Theory of Confidence Estimation,[9] the confidence interval ϵ of $\overline{x}$ for a sample set of size N with unknown variance can be written as:

$$\epsilon = \frac{s_x t_{N-1;\alpha}}{\sqrt{N}} \tag{8}$$

where s_x is the square root of our variance estimate from above, and $t_{N-1;\alpha}$ corresponds to the interval of a t-distribution $T \sim t(N-1)$ such that:

$$Pr\left[|T| > t_{N-1;\alpha}\right] = 2\alpha \tag{9}$$

In other words, given $\alpha < 0.05$, we can make the following statement about our estimate of the mean:

$$\Pr\left[|\mu_x - \overline{x}| < \frac{s_x t_{N-1;\alpha}}{\sqrt{N}}\right] = 1 - 2\alpha \tag{10}$$

We can tabulate confidence intervals for various numbers of pages in the sample set. These values are given in Tables 7 and 8 for the clean- and degraded-document experiments, respectively.

Table 7 shows how much the sampled accuracy can differ from the true accuracy for various dataset sizes when using clean, first-generation originals. The entries can be interpreted as follows: with 90% probability,

the accuracy of the sample differs from the true accuracy by no more than the indicated amount. As the table demonstrates, to estimate accuracy to within 0.1% with 90% probability, 20 pages are sufficient in the case of non-space errors, and 40 pages are enough in general. Because these particular experiments are based on clean input, however, these figures should be viewed as lower bounds.

	Non-Space Errors			All Errors		
Pages	$E1$	$E2$	$E3$	$E1$	$E2$	$E3$
5	**0.096**	0.220	0.180	0.313	0.234	0.338
10	0.061	0.140	0.115	0.199	0.149	0.215
20	0.041	**0.094**	**0.077**	0.134	**0.100**	0.145
30	0.032	0.076	0.062	0.108	0.080	0.116
40	0.028	0.065	0.053	**0.093**	0.069	**0.100**

Table 7: 90% confidence intervals for the three clean-document experiments (measured in percent).

As Table 8 shows, to estimate accuracy to within 0.1% with 90% probability in the case of degraded documents, more pages are required. In our experiments, the number of pages needed ranges from 30 to 120 for non-space errors, and from 60 to 150 in general. When combined with the results presented in Table 6, we can see that the lower the recognition accuracy, the larger the sample space must be.

	Non-Space Errors			All Errors		
Pages	$E4$	$E5$	$E6$	$E4$	$E5$	$E6$
5	0.387	0.251	0.554	0.416	0.412	0.639
10	0.246	0.160	0.352	0.265	0.262	0.406
20	0.166	0.108	0.237	0.178	0.176	0.273
30	0.133	**0.086**	0.190	0.143	0.142	0.220
40	0.114	0.074	0.164	0.123	0.122	0.189
60	**0.093**	0.060	0.133	**0.100**	**0.099**	0.153
120	0.065	0.042	**0.093**	0.070	0.069	0.107
150	0.058	0.037	0.083	0.062	0.061	**0.095**

Table 8: 90% confidence intervals for the three degraded-document experiments (measured in percent).

Tables 9 and 10 present our results for experiment $E4'$. Not only are the accuracies lower than for $E4$, but the variances are also much higher (by several orders of magnitude). Based on our estimates, over a thousand pages would be required to obtain a reliable OCR accuracy figure in this case.

	Non-Space Errors		All Errors	
Experiment	Mean	Variance	Mean	Variance
$E4'$	0.9891	2.699512e-03	0.9757	3.032921e-03

Table 9: Mean accuracy and variance for experiment $E4'$.

Pages	Non-Space Errors	All Errors
5	4.682	4.963
40	1.383	1.466
60	1.121	1.188
120	0.780	0.827
200	0.604	0.641
500	0.382	0.405
1,000	0.270	0.286

Table 10: 90% confidence intervals for experiment $E4'$.

4 CONCLUSIONS

As the data in Table 7 demonstrate, OCR experiments using clean documents can yield reliable accuracy estimates using a relatively small number of pages. On the other hand, biases in error rates induced by differences in the equipment and procedures used to perform otherwise identical experiments can be quite high. Consider, for example, experiments $E1$ and $E3$. The only differences between these two tests were a minor change in the text formatting, and the use of a Ricoh scanner in one case and an HSD scanner in the other. Still, this was sufficient to double the computed error rate. This result suggests that error rates for similar experiments should not be considered significant beyond the first decimal place unless an extremely careful description of the hardware and software is made. Even so, the "human factor" – differences in operator precision – may well be enough to introduce biases of this degree.

Our results show that as OCR accuracies continue to improve, we will have to be increasingly careful about associated biases. The admittedly artificial experiments described in this paper were carefully controlled, much more so than in practice where it appears that accuracy figures are often compared without proper regard for the concerns we have discussed in this paper. It is now becoming known, for example, that even the procedure used for counting errors can introduce a significant variance into the measure.[10] In conclusion, aspects of OCR experiments which in the past have been considered unimportant can, in fact, make a significant difference.

The use of standard document image databases distributed on CD-ROM is one way to address the issue of systematic bias.[7] Note, however, that this approach has its own disadvantages: dependence on a small, predefined set of images, lack of control (on the experimenter's part) over the equipment used to generate the test data, etc.

In this paper, we have examined the problem of systematic bias in OCR experiments. As we have shown, this is an important issue that must be considered when evaluating results reported in the literature. The role of sampling in experimental computer vision has been discussed from a conceptual standpoint elsewhere.[4,5] Our paper quantifies the magnitude of real-world effects in the case of OCR.

5 ACKNOWLEDGEMENTS

The authors would like to thank Yuh-Lin Chang, Jeffrey Esakov, and Joshua Tauber for their helpful comments and assistance in gathering some of the data for this paper. The *Moby-Dick* text we used in our experiments was obtained from the Guttenberg Project at the University of Illinois, as prepared by E. F. Irey from the Hendricks

House edition. The trademarks mentioned in this paper are the property of their respective companies.

6 REFERENCES

[1] H. S. Baird. Document image defect models. In H. S. Baird, H. Bunke, and K. Yamamoto, editors, *Structured Document Analysis*, pages 546–556. Springer-Verlag, New York, NY, 1992.

[2] J. Esakov, D. P. Lopresti, and J. S. Sandberg. Classification and distribution of optical character recognition errors. In *Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging*, volume 2181, pages 204–216, San Jose, CA, February 1994.

[3] J. Esakov, D. P. Lopresti, J. S. Sandberg, and J. Zhou. Issues in automatic OCR error classification. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 401–412, Las Vegas, NV, April 1994.

[4] R. Fenrich and J. J. Hull. Concerns in creation of image databases. In *Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition*, pages 112–121, Buffalo, NY, May 1993.

[5] R. M. Haralick. Methodology for experimental computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 437–438, San Diego, CA, June 1989.

[6] D. P. Lopresti and J. Zhou. Using consensus sequence voting to correct OCR errors. In *Proceedings of IAPR Workshop on Document Analysis Systems*, pages 191–202, Kaiserslautern, Germany, October 1994.

[7] I. T. Phillips, S. Chen, and R. M. Haralick. CD-ROM document database standard. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 478–483, Tsukuba Science City, Japan, October 1993.

[8] S. V. Rice, J. Kanai, and T. A. Nartker. The third annual test of OCR accuracy. In *UNLV Information Science Research Institute Annual Report*, pages 11–38. University of Las Vegas, Las Vegas, NV, 1994.

[9] V. K. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley and Sons, Inc., 1976.

[10] J. S. Sandberg. Counting OCR errors in typeset text. Technical Report 85-93, Matsushita Information Technology Laboratory, January 1994.

[11] J. Zhou and D. Lopresti. Repeated sampling to improve classifier accuracy. In *Proceedings of the IAPR Workshop on Machine Vision Applications*, pages 346–351, Kawasaki, Japan, December 1994.