

Scanning for OCR Text Conversion

Background

Optical Character Recognition (OCR) is the recognition and interpretation of text in a digital image. Its main application is the translation of printed type into an editable and searchable document. Modern OCR software has become very accurate, however; in many ways it is still largely dependent on the quality of the scanned image. In particular, when dealing with bound print sources such as novels and thick text-books that, due to physical limitations, are not easily read by optical scanning devices, it is highly desirable to have a device that can produce high quality images, capture all the information and give consistent results for every page.

Aim

To measure the accuracy of OCR when a thick-spine book is scanned with a conventional flatbed face-down scanner versus BookDrive DIY and to show how the image quality and ultimately the scanning method is related to this.

Method

We compared the accuracy of OCR text conversion on pre-processed images from a conventional flatbed face-down scanner (typical in most office environments) and BookDrive DIY.



The sample book (shown to the left) is a hard-cover 1421 page text-book in good condition. The font type and size is quite standard, the page background is bright and white and there are no images. We scanned four pages from the center of the book (pages 700, 701, 702 and 703) at 200 and 300 dpi resolutions to produce color JPEG images without further manipulation or processing.

For the sake of simplicity we will give three different performance measures rather than attempting to apply a single, all-encompassing formula.

The first and most important measure is our OCR 'Accuracy' measure.

$$\text{Accuracy} = 100 - (\text{missed characters} + \text{uncertain characters}) * 100 / \text{actual total characters}$$

This measures the percentage of printed type that is recognized and confidently interpreted.

The second measure is our 'Image Quality' measure.

$$\text{Image Quality} = \text{interpreted characters} * 100 / \text{actual total characters}$$

This measures the percentage of printed type that failed to be recognized as text at all, regardless of whether characters were interpreted correctly or not.

The third measure is our OCR 'Confidence' measure.

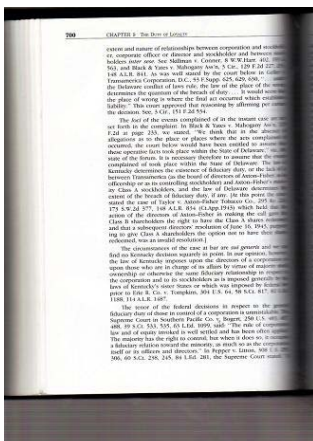
$$\text{Confidence} = 100 - \text{uncertain characters} * 100 / \text{interpreted characters}$$

This measures the percentage of interpreted characters that are questionable or cannot be interpreted with any real level of confidence. Image resolution may be an important factor here.

Observations

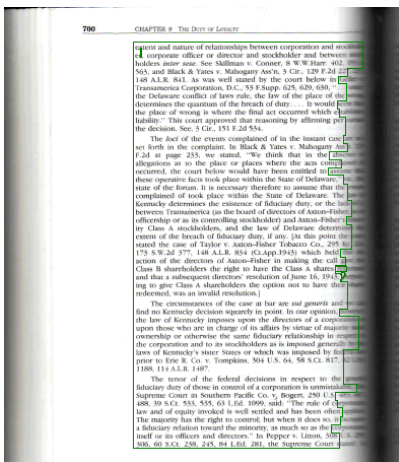
Before we look at the figures it is worth mentioning the difficulty we experienced in scanning this type of thick-spine book with a flatbed face-down scanner. In particular, a lot of handling was involved, not only in flipping the book over to turn the page but in repositioning the book before each scan so that each page lay within the boundary of the A4 scanning region. Furthermore, considerable force had to be applied to flatten the pages near the page binding in order to capture as much of the text as possible. In fact some pages had to be scanned two or three times to obtain an acceptable image but even then results were less than satisfactory. In terms of ease-of-use and speed or throughput, flatbed face-down scanners are clearly more cumbersome and much slower than BookDrive DIY and they rely heavily on the ability of the user to correctly and consistently position the book each time before scanning.

An image sample from the flatbed scanner



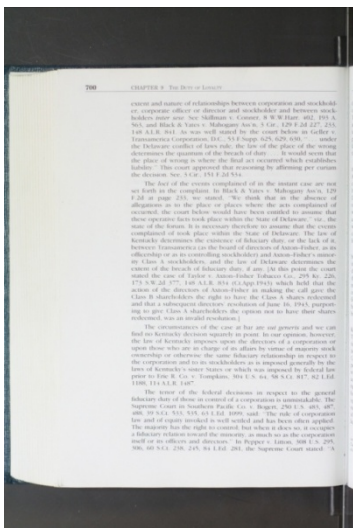
The image above shows page 700 of the text-book when scanned with the flatbed scanner. The image has not been processed or enhanced in any way.

The OCR text recognition area



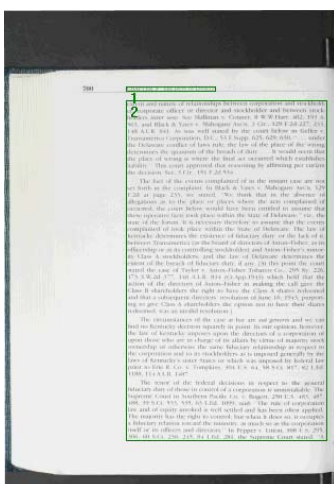
The image above shows the boundary of the recognizable text area as identified by our OCR software. The area close to the page binding is dark and fuzzy and the text is unreadable.

An image sample from BookDrive DIY



The image above shows page 700 of the text-book when scanned with BookDrive DIY. The image has not been processed or enhanced in any way.

The OCR text recognition area



The image above shows the boundary of the recognizable text area as identified by our OCR software. The boundary encloses all of the text on the page.

Results

OCR accuracy on test sample from flatbed face-down scanner

Page	Sample Page Image		OCR Accuracy	
	Total Characters	Resolution (dpi)	Interpreted Characters	Uncertain Characters
700	3260	200	2858	38
700	3260	300	3060	81
701	3312	200	2811	68
701	3312	300	2867	70
702	3260	200	2953	60
702	3260	300	2937	71
703	2895	200	2561	90
703	2895	300	2580	46

OCR accuracy on test sample from BookDrive DIY

Page	Sample Page Image		OCR Accuracy	
	Total Characters	Resolution (dpi)	Interpreted Characters	Uncertain Characters
700	3260	200	3260	2
700	3260	300	3260	2
701	3312	200	3311	9
701	3312	300	3312	5
702	3260	200	3259	5
702	3260	300	3260	1
703	2895	200	2894	5
703	2895	300	2895	4

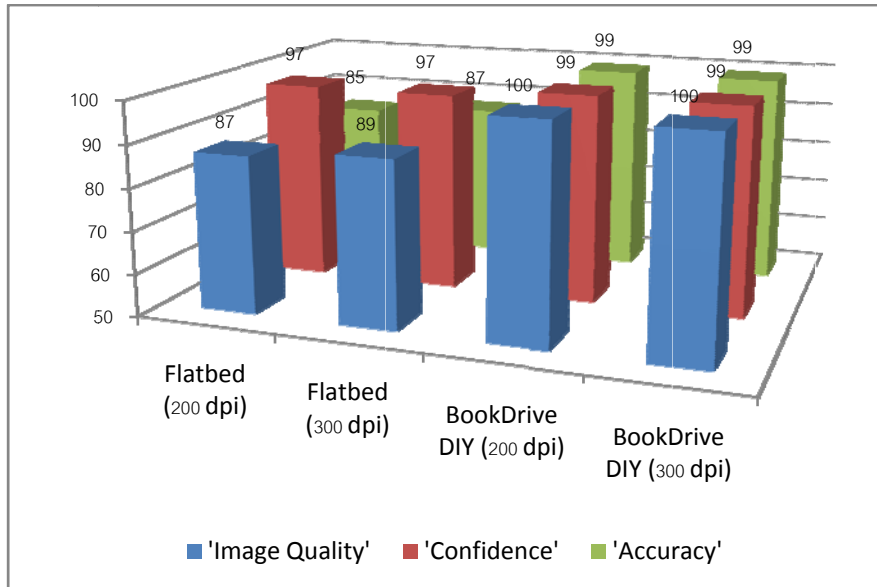
The above figures show the ratio of interpreted characters to uncertain characters for each page when OCR was performed on the test samples.

For samples from the flatbed scanner, the total number of interpreted characters was much less than the actual total number of characters on the page and the number of uncertain characters was very high. The results for flatbed scanners are inconsistent and there appears to be no direct correlation between image resolution and OCR accuracy. For some pages more characters were interpreted correctly with less uncertain characters using the higher resolution images while for others the opposite was true.

For samples from BookDrive DIY, there was only a small deviation in the total number of interpreted characters for the 200 and 300 dpi images and the results were more consistent. In general there were slightly less uncertain characters with the higher resolution (300 dpi) images and the number of interpreted characters was exactly equal to the actual number of characters on the page.

OCR Performance Measures for F

latbed Scanner vs BookDrive DIY



The chart above shows OCR performance measures based on images from the flatbed scanner versus BookDrive DIY. Note that figures represent percentage values rounded to 1 decimal place and floored to the nearest whole number.

Conclusion

Because the flatbed scanner produced poor images characterized by a dark fuzzy region near the page binding, OCR 'Accuracy' was very low and a high percentage of printed type (in this region) was missed entirely. This is reflected in the low 'Image Quality' rating. At a higher resolution (300dpi) some of this type was more recognizable but most of it was interpreted incorrectly and in real terms there was no overall improvement in 'Confidence'. BookDrive DIY produced high quality images without dark shady regions and OCR 'Accuracy' was extremely high on these samples. Increasing the image resolution had no real impact on either 'Confidence' or 'Accuracy', however; the 300dpi samples seemed to produce slightly better figures.

