



P.O. Box 22219, Lincoln, NE 68542-2219
Phone: (402) 423-4782, Fax: (402) 423-5154
<http://www.braille2000.co>

A Primer on Scanning for OCR

Scanning is a process by which flat images are converted to digital form, what you might call a digital photograph. As a photograph, the complete image can be viewed on the screen or even printed. But you can't perform word-processing on the text and you can't automatically translate the text to braille, because you only have a picture of the text, not characters codes (not ASCII codes). OCR generates character codes to match the pictured text. If OCR works perfectly, the character codes are exactly the text you scanned. As with photography, scanning is part science and part art. The question is, what kinds of adjustments can and should be made to get the best overall results?

So, for better or for worse, to get the best results you have to know something about the nature of OCR. So in this primer we will start with OCR concepts and work backwards to derive recommendations for scanning.

Optical Character Recognition (OCR) is a process by which glyph images (a glyph is the visual image of a character) yield character codes. The glyph **A** is a picture that looks like an A. The character A is a specific computer code that stands for A (a code such that if you send it to a printer, it causes the printer to paint the glyph that looks like an A on the page). Glyphs (shapes) and characters (the enumeration of an alphabet) are two linked but distinct abstract concepts.

Here is an enlargement of the word "closet." scanned from a paperback book.

The picture shows the glyphs for a word spelled c l o s e t followed by a period, as printed in the middle of a sentence in the middle of some page in a book (this is a real example). The job of OCR is to figure out which dark patches go together to form individual glyphs, and then to match the shapes one at a time against a library of glyph drawings to find out which glyph in the library is most similar to each glyph scanned, and to assign the character code that goes with the most similar glyph and thus identify the shapes. When OCR works on the picture shown above, we would like it to identify the shapes for the c and the l and the o, etc., and spit out the character codes for c l o s e t and period. Given a picture of letters arranged as words, OCR is supposed to give back strings of character codes arranged as words. Looking at the example, which clearly says "closet." you would think this process would be simple. But think again. This time squint your eyes as you look at the example.

If you squint your eyes enough, the t and the period start to look very much like a capital L. OCR algorithms are aware that printed text is not perfect and that some letters may have "drop outs" which are places where the ink is lighter or missing. The little partial gap between the bottom of the t and the period could be interpreted as a drop out and then the two dark areas would belong together, and it would be reasonable to think it was a capital L. This "error" is a real one, and when this word was converted with OCR, the text produced was closeL (no period).

Now squint your eyes some more. The e tends to look more and more like a c. And the s tends to look more and more like an a. Those "errors" were not made by the OCR program I used on this example, but they are the kinds of errors made by OCR. Remember than humans understand recognize glyphs in the context of the discourse. "closeL" is thus nonsense and we instantly reconsider things and view it as "closet.". OCR tools can't understand the context of the discourse, and consider only individual letters, and can thus be fooled.

Here is the same word on the same page, however the scanning of the page is different. You should notice two differences: the background is not as white and the glyphs are smoother and less jagged. Squint your eyes again. The t and period don't look as much like an L do they? And you're right. The OCR conversion produced closet. this time rather than closetL as before. Why?

The first difference, the gray background, is one reason. No, gray is not a good color for scanning (and it need not be this gray). This second example was scanned in "gray scale" mode while the first one was

scanned in “black and white” mode. In gray scale, the individual dots of the digital image are represented by a number that varies from 0 for black to 255 for white with possible values in between for 254 various shades of gray. In black and white, each dot of the digital image can be only 0 for black or 1 for white, with no shades in between. In the black and white image of the word (the top illustration) the background is white although in reality it was light gray (the page scanned was from a paperback book whose pages were a little off-white). The black and white image looks really “crisp”, but in reality all the lighter grays have been set to “white” and all the darker grays have been set to “black” and this can remove some detail in the image, such as the tail of the t as it approaches the period. So Advice #1 is: set your scanner for gray scale images, not black and white (on some scanners, gray scale is called black and white photo mode and the black and white mode is called black and white drawing mode).

The second difference between the images is the smoothness of the glyph edges. When an image is scanned, the scanner measures the light intensity for thousands (or millions) of equally spaced locations along the image. These locations are called “dots” or “pixels” (picture elements), arranged as a grid covering the image. The scanner has both coarse grids and medium grids and fine grids, controlled by the “resolution” of the scan, as measured in dot locations per linear inch (dots per inch or dpi). The dpi number tells how fine the grid of measurements is. At 100 dpi (the first image) there are 100 picture intensity measurements made per linear inch. The grid is usually a square grid (same dpi vertically as horizontally) and if you consider one square inch of picture, at 100 dpi, there will be a total of 10,000 measurements taken. To the computer, this means 10,000 values, each being either 0 or 1 for a black and white scan, or each being a number between 0 and 255 for a gray scale scan.

The second (better) image was done with a scan resolution of 300 dpi. The smoothness of the glyph outlines comes from measuring the light with a finer grid. There is more data and more detail. And more detail means better OCR. Scan resolution is adjustable through software to values from about 50 dpi up to 600 dpi or more (the maximum productive dpi is the so called “optical resolution” of the scanner; higher optical resolution costs more, but fortunately nearly all scanners have optical resolutions of 300 dpi or more, quite sufficient). So Advice #2 is: set your scanner for a scan resolution of about 300 dpi. Using 300 dpi will increase the size of image files and slow down individual scans, but it will yield OCR results with fewer errors, requiring less “fixing up”. At 300 dpi, a gray scale scan of a typical textbook page will yield roughly 1 to 3 megabytes of data (a significant size, but worth it).

There are other image adjustments that you can make, but they are of lesser importance. Try to put the page to be scanned on the scanner nicely aligned with the text running straight across and not at an angle. OCR tools have automatic image rotation capabilities, but it can impair recognition quality. You may also have adjustments for image brightness and image contrast and there may be an automatic setting for these things. The automatic setting usually does a good job. If your scanner has no automatic brightness and contrast mode, then adjust those aspects by doing a test scan and then viewing the image on the screen and zooming in to look at the image quality, particularly background color and “noise” (e.g., dark blotches). Choose a setting that gives a clear sharp image on white background without blotches.

A case in point... page 68 from Christopher Andersen’s New York Times Bestseller, Jack and Jackie. On the next page is the black and white scan at 100 dpi (left) and the OCR output (right). The OCR “errors” are numbered (by me) in the middle: 1. bold out of the blue; 2. L for t period (discussed above); 3. Hyphen instead of 1-em dash; 4. rn for m; 5. in for m. In general the OCR output is good, but why fix “errors” when you don’t have to.

The last page shows the scan image and OCR output using gray scale mode at 300 dpi. There is one new error, #6 (due to color of the paper and the gray scale scan being slightly darker than optimum) in which an ellipsis period becomes a comma. The length of the dash is still wrong and there is one bold out of the blue, but the character errors are gone, i.e., text of this quality should translate into very good braille.

Summary

1. use gray scale scanning
2. use 300 dpi, for most things
3. you may want to adjust brightness and/or contrast
4. make a few test scans and OCR conversions
5. expect near-perfect letter recognition (from good print sources not involving unusual decorative fonts)
6. be prepared to fix some punctuation, especially small punctuation including dashes, periods, and commas.

larger than a closet. To the right, twin beds. To the left, the door to a small bathroom with a white tile floor. Directly ahead, a single large window with a view of lawn, woods, river . . ." In the closet, Jackie found several white shirts with Vidal's name sewn in the collar. She wore them while horseback riding.

Summers were spent at Hammersmith Farm, with its twenty-eight rooms, fourteen fireplaces, seventeen bathrooms, and a resident staff of sixteen. Downstairs the hallways were carpeted in crimson. Already mindful of such things, Jackie decreed that henceforth all Auchincloss dogs should be black to provide suitable contrast. In time, their menagerie included a Scottie, a poodle, a cocker spaniel, and Caprice, Jackie's own Bouvier des Flandres—all black.

The focal point of the house was not the formal living room, but the hotel lobby-sized deck room with its dramatic green-tiled fireplace, wraparound view of the water, and nautical motif—including a stuffed pelican suspended from the ceiling and framed photographs of yachts once owned by the family.

The noted interior designer Elisabeth Draper, who had been hired by the peripatetic Janet to redo Merrywood and Hammersmith Farm, transformed Jackie's third-floor room overlooking Narragansett Bay into a light-filled aerie. The walls were pale yellow, and a classical frieze chosen by Jackie ran along the border of the oddly pitched ceiling. The furniture—two twin beds, mirrored dressing table, and writing desk—were all white.

At first Jackie wrote to her father saying she was lonely and missed New York. It was what Black Jack wanted to hear. For him, the marriage of Hughdie and Janet had been particularly traumatic: Black Jack continued paying for their schooling—Jackie now attended Holton-Arms, a Washington, D.C., day school—but he saw little of his cherished daughters.

Determined that she not be embarrassed by her gangly daughter, Janet followed the lead of Washington's finest families and enrolled Jackie in Miss Shippen's dance class. ("I

1,2: **larger than** a closet. To the right, twin beds. To the left, the door to a small bathroom with a white tile floor. Directly ahead, a single large window with a view of lawn, woods, river . . ." In the closet, Jackie found several white shirts with Vidal's name sewn in the collar. She wore them while horseback riding.

Summers were spent at Hammersmith Farm, with its twenty-eight rooms, fourteen fireplaces, seventeen bathrooms, and a resident staff of sixteen. Downstairs the hallways were carpeted in crimson. Already mindful of such things, Jackie decreed that henceforth all Auchincloss dogs should be black to provide suitable contrast. In time, their menagerie included a Scottie, a poodle, a cocker spaniel, and Caprice, Jackie's own Bouvier des Flandres—all black.

The focal point of the house was not the formal living room, but the hotel lobby-sized deck room with its dramatic green-tiled fireplace, wraparound view of the water, and nautical motif—including a stuffed pelican suspended from the ceiling and framed photographs of yachts once owned by the family.

The noted interior designer Elisabeth Draper, who had been hired by the peripatetic Janet to redo Merrywood and Hammersmith Farm, transformed Jackie's third-floor room overlooking Narragansett Bay into a light-filled aerie. The walls were pale yellow, and a classical frieze chosen by Jackie

4,3: ran along the border of the oddly pitched ceiling. The furniture—two twin beds, mirrored dressing table, and writing desk—were all white.

1: At first Jackie wrote to her father saying she was lonely and missed New York. It was what Black Jack wanted to hear. **For him, the marriage** of Hughdie and Janet had been particularly traumatic: Black Jack continued

3: paying for their schooling—Jackie now attended Holton-Arms, a

3: Washington, D.C., day school—but he saw little of his cherished daughters.

5: Determined that she not be embarrassed by her gangly daughter, Janet followed the lead of Washington's finest families and enrolled Jackie in Miss Shippen's dance class. ("I

larger than a closet. To the right, twin beds. To the left, the door to a small bathroom with a white tile floor. Directly ahead, a single large window with a view of lawn, woods, river . . ." In the closet, Jackie found several white shirts with Vidal's name sewn in the collar. She wore them while horseback riding.

Summers were spent at Hammersmith Farm, with its twenty-eight rooms, fourteen fireplaces, seventeen bathrooms, and a resident staff of sixteen. Downstairs the hallways were carpeted in crimson. Already mindful of such things, Jackie decreed that henceforth all Auchincloss dogs should be black to provide suitable contrast. In time, their menagerie included a Scottie, a poodle, a cocker spaniel, and Caprice, Jackie's own Bouvier des Flandres—all black.

The focal point of the house was not the formal living room, but the hotel lobby-sized deck room with its dramatic green-tiled fireplace, wraparound view of the water, and nautical motif—including a stuffed pelican suspended from the ceiling and framed photographs of yachts once owned by the family.

The noted interior designer Elisabeth Draper, who had been hired by the peripatetic Janet to redo Merrywood and Hammersmith Farm, transformed Jackie's third-floor room overlooking Narragansett Bay into a light-filled aerie. The walls were pale yellow, and a classical frieze chosen by Jackie ran along the border of the oddly pitched ceiling. The furniture—two twin beds, mirrored dressing table, and writing desk—were all white.

At first Jackie wrote to her father saying she was lonely and missed New York. It was what Black Jack wanted to hear. For him, the marriage of Hughdie and Janet had been particularly traumatic: Black Jack continued paying for their schooling—Jackie now attended Holton-Arms, a Washington, D.C., day school—but he saw little of his cherished daughters.

Determined that she not be embarrassed by her gangly daughter, Janet followed the lead of Washington's finest families and enrolled Jackie in Miss Shippen's dance class. ("I

6: larger than a closet. To the right, twin beds. To the left, the door to a small bathroom with a white tile floor. Directly ahead, a single large window with a view of lawn, woods, river. , ." In the closet, Jackie found several white shirts with Vidal's name sewn in the collar. She wore them while horseback riding.

3: Summers were spent at Hammersmith Farm, with its twenty-eight rooms, fourteen fireplaces, seventeen bathrooms, and a resident staff of sixteen. Downstairs the hallways were carpeted in crimson. Already mindful of such things, Jackie decreed that henceforth all Auchincloss dogs should be black to provide suitable contrast. In time, their menagerie included a Scottie, a poodle, a cocker spaniel, and Caprice, Jackie's own Bouvier des Flandres—all black.

1: The focal point of the house was not the formal living **room, but the hotel** lobby-sized deck room with its dramatic green-tiled fireplace, wraparound view of the water, and nautical motif—including a stuffed pelican suspended from the ceiling and framed photographs of yachts once owned by the family.

3: The noted interior designer Elisabeth Draper, who had been hired by the peripatetic Janet to redo Merrywood and Hammersmith Farm, transformed Jackie's third-floor room overlooking Narragansett Bay into a light-filled aerie. The walls were pale yellow, and a classical frieze chosen by Jackie
3: ran along the border of the oddly pitched ceiling. The furniture—two twin
3: beds, mirrored dressing table, and writing desk—were all white.

3: At first Jackie wrote to her father saying she was lonely and missed New York. It was what Black Jack wanted to hear. For him, the marriage of Hughdie and Janet had been particularly traumatic: Black Jack continued
3: paying for their schooling—Jackie now attended Holton-Arms, a
3: Washington, D.C., day school—but he saw little of his cherished daughters.

Determined that she not be embarrassed by her gangly daughter, Janet followed the lead of Washington's finest families and enrolled Jackie in Miss Shippen's dance class. ("I