

# OCR (Optical Character Recognition): Converting paper documents to text



Michael Hogan  
John W. Shipman

2008-01-02 20:00

## Abstract

Instructions for scanning text documents to produce PDF files using optical character recognition.

This publication is available in Web form<sup>1</sup> and also as a PDF document<sup>2</sup>. Please forward any comments to [tcc-doc@nmt.edu](mailto:tcc-doc@nmt.edu).

## Table of Contents

1. Overview .....	1
2. Creating the PDF file .....	2
3. Character recognition .....	2

## 1. Overview

This document describes techniques for converting textual content from a paper document into machine-readable form. The computer actually attempts to “read” or recognize the characters on your page, through a technique called *OCR*, for Optical Character Recognition.

If all you want to do is capture an exact image of a flat original, see *Using the flatbed scanner*<sup>3</sup>. Use OCR if you want to extract the textual content. OCR is definitely indicated if you want to modify the text.

### Warning

Please do not expect miracles from this process. For best results, you will need an original document that is very crisply printed in a common font. Results will be poor or useless for originals with complex layouts, strange fonts, and stray marks. If your original has multiple columns, the result may mix text from the columns together; single-column originals work best.

For this process, you will need to use one of the PC workstations that has a flatbed scanner attached. Most of these workstations are in Speare 5; ask the User Consultant there to help you find an appropriate system.

<sup>1</sup> <http://www.nmt.edu/tcc/help/pubs/ocr/>

<sup>2</sup> <http://www.nmt.edu/tcc/help/pubs/ocr/ocr.pdf>

<sup>3</sup> <http://www.nmt.edu/tcc/help/pubs/flatscan/>

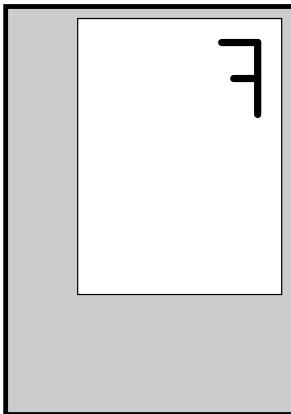
The necessary software package, *Adobe Acrobat Professional 6.0*, is available on all scanner-equipped systems. In general, there are two overall parts of the process. First you will scan the document and convert it into an Adobe PDF (Page Description Format) file. In the second step, this program attempts to recognize the text and attach it to the document. If this part succeeds, you can then save the textual content in other formats such as Microsoft Word and ordinary text.

## 2. Creating the PDF file

---

This section tells you how to scan your document and make it into a PDF file. This file can be useful by itself: it contains a pictorial representation of the original, and can be used to print reasonable paper copies.

1. Log in under Windows XP.
2. From the *Start* menu, select *All Programs* → *Adobe CS* → *Adobe Acrobat 6.0 Professional*.
3. Open the scanner and place your copy face down, with the top left corner of the page aligned with the top right corner of the scanner, like this:



4. In the *Adobe Acrobat Professional* window, select *File* → *Create PDF* → *From Scanner*.  
This brings up the *Create PDF From Scanner* window.
5. You may want to drag the horizontal slider at the bottom of this window toward “Higher Compression” to get smaller files, or toward “Higher Quality” for larger, cleaner-looking files. If in doubt, use the default setting.
6. Click the *Scan* button. You will see a progress bar. This part will take a minute or so.
7. When the scan is completed, you will get a popup menu entitled *Acrobat Scan Plug-In*. If you have another page of copy, load it into the scanner and click *Next*. Otherwise, click *Done*.

Once you have scanned all the pages of a particular document, leave the Adobe Acrobat window open and proceed to the next section.

## 3. Character recognition

---

The recognition process starts with a PDF file from the displayed in the Adobe Acrobat Professional window.

1. Pull down *Document* → *Paper Capture* → *Start Capture*. This brings up the *Paper Capture* window.

2. Under *Pages*, select the radiobutton for *All pages*, or *Current page*, or a range of pages by page number.
3. Under *Settings*, you will see a short list, and the second line of that list will display the “PDF Output Style”. If this element does not say “**PDF Output Style: Formatted Text & Graphics**”, click the *Edit...* button to bring up the *Paper Capture Settings Menu*; click the pulldown menu *PDF Output Style*; select “**Formatted Text & Graphics**”; and click *OK*.
4. Click *OK* to start the recognition process. This will take at least a few seconds. Once it is complete, your page will be redisplayed.

At this point, depending on how well the software has coped with your original, it may have recognized all the text, some of the text, or no text at all. When the software cannot recognize an area of the original, its fallback strategy is to represent that area as an image, instead of text. So in the general case what you now have is a mixture of text and images. You will find out how well it worked after you have saved the document in your desired format.

5. Click *File* → *Save As*. This brings up the *Save As* window.
6. At the top of this window is the usual “*Save in:*” fixture. Click the drop-down menu and navigate to the **U:** drive to save the file in your TCC account.
7. In the *File name:* field, enter the name of the file you want to save, with the customary suffix such as *.doc* for MS-Word files, *.txt* for text files, and so on.
8. In the *Save as type:* pulldown menu, select the file format you want to save in. Recommend choices include:

```
HTML 3.2 (*.htm)
HTML 4.01 with CSS 1.00 (*.htm)
Microsoft Word Document (*.doc)
Rich Text Format (*.rtf)
Text (Plain) (*.txt)
```

Check the file produced by this process, comparing the result to the original to see how well it worked. There will probably be some cleanup necessary. For example, if you exported the file to MS-Word, you may see some parts of the document represented as images. You'll need to retype those. Also, MS-Word representations of the text may be full of strange paragraph formatting and spurious changes to text fonts, styles and size.

If all else fails, you may need to retype the document. Unless your original is quite clean, the author's experience has been that most decent typists can retype it in less time than it will take you to scan the document and clean it up.

