Image Advantage Solutions Inc.

Bringing Your Documents into the Digital Age

So What is OCR Anyway?

An overview of Optical Character Recognition

So What is OCR Anyway?

Technical Article

Written by: Image Advantage Solutions Inc.

Company: Image Advantage Solutions Inc.
1-1354 County Road #2
Mallorytown, Ontario, Canada
K0E 1R0

Phone: 1-613-659-4620

Email: sales@imageadvantage.com

Website: www.imageadvantage.com

Published: 2014-08-25

Learn About:

- What is Optical Character Recognition?
- How is it performed?
- Why is it beneficial?
- How do you do it?
- What are the barriers to a good OCR?
- Should you try it yourself?

Table of Contents

Overview

OCR stands for “Optical Character Recognition” which is the process of converting text on paper to editable text. This article will discuss the processes for performing Optical Character Recognition and its benefits.

What is OCR

The Merriam-Webster Dictionary defines OCR as:

> “Scanning and comparison technique intended to identify printed text or numerical data. It avoids the need to retype already printed material for data entry. OCR software attempts to identify characters by comparing shapes to those stored in the software library. The software tries to identify words using character proximity and will try to reconstruct the original page layout. High accuracy can be obtained by using sharp, clear scans of high-quality originals, but it decreases as the quality of the original declines.”

So in simple terms it is the process of converting text on paper to text that can be used in a computer word processor.

How is it Performed?

The first step in the OCR process is to scan the document. The best setting for OCR scanning is 300dpi “black and white”. Pretty much any scanner these days will give a good enough image for a good quality OCR.

The next step is to use computerized OCR software to OCR the image. So let’s get a little technical. The way this works (see figure 1) is that the OCR engine will divide the page up into a grid of very small boxes. It will then start at the top left corner of the page and look at each box going from left to right. Assuming that the very top of the page is all white it will keep on going down to the next line down and keep going until it sees a box that is not all white. The system will then follow that character around until it comes back to where it started. It will then cut that character out and compare it to a number of templates that will be character sets in a large number of type styles and sizes. Once it gets the best match it will know what the character is, the font and the font size. It will then compare the following characters to that template until it sees a different font or font size. After the OCR process it will usually save its results as either a word processor document or text file. More sophisticated OCR software programs will include formatting of the document and also include graphics, tables, charts, headers and footers.

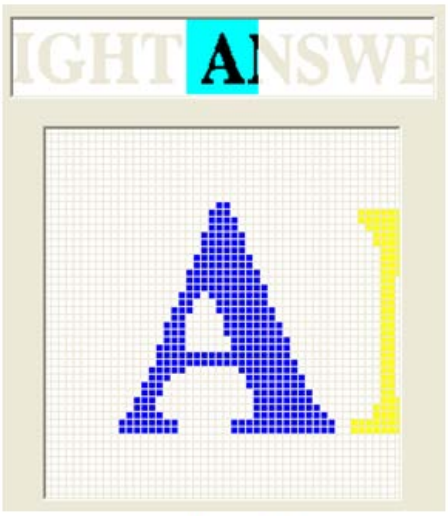


Figure 1

Why is it Beneficial?

OCR'ing is beneficial for a number of reasons:

1. **Make Document Editable** – It is much more efficient and cost effective to OCR a paper report, book, manual...etc. than to retype it.

2. **Full Text Search Capability –** OCR may be applied to textual Adobe Acrobat PDF files to make them full text searchable. When combined with Adobe Acrobat Catalogue searching across whole collections of multiple files becomes possible.

3. **Make Documents Ready for Readers for the Blind –** Technology is available for the blind that will read text from a document out loud. OCR'ing may assist in making paper documents textual so that this is possible.

How do you do it?

1. **Paper Documents to Editable Text** – One of the most popular OCR software solutions is Nuance OmniPage Ultimate. It is in the $130US range and is very powerful. There is a fairly steep learning curve to this software however and novice users may have difficulty. If you don't use this process regularly you may find yourself having to relearn it each time. It will allow OCR'ing from PDF or TIFF images and will allow you to capture text, images, tables and graphs. It will attempt to format your document but there is usually some manual formatting involved. The OCR accuracy is very high.

2. **Adobe Acrobat PDF Files** – Scanned images of paper documents stored as Adobe Acrobat PDF files may be OCR'ed using the Adobe Acobat OCR functionality. With the full version of Adobe Acrobat Professional, about $380US, you would select from the menu bar:

   <Document><OCR Text Recognition><Recognize Text Using OCR>

   If you choose the <Searchable Image Exact> option it will still show the scanned image but will put an OCR'ed textual document behind the scanned image. This will allow you to cut and paste text from the image and will also make the image "full text searchable" so you can use the <find> tool to search for text right inside the page.

   The Adobe Acrobat Catalogue will allow you to create a search index that will allow you to search across multiple PDF files. You can create an Adobe Acrobat Catalogue with Adobe Acrobat Professional by clicking on the menu bar:

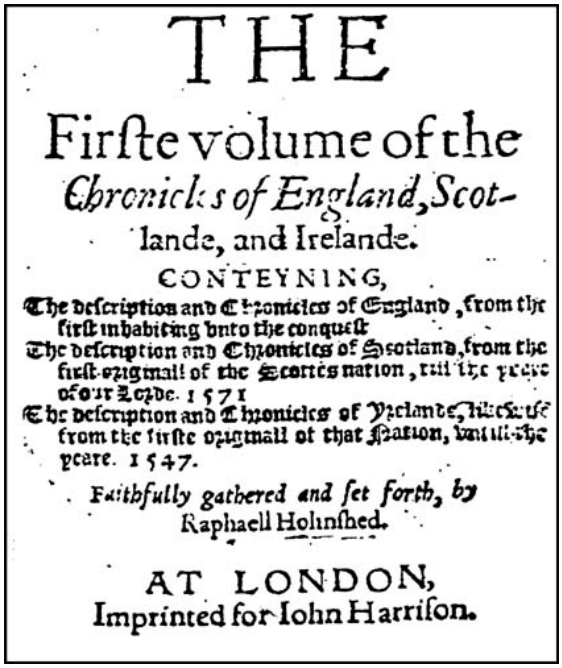
   <Advanced><Document Processing><Full Text Index with Catalogue>

What are the Barriers to a Good OCR

The best results of the OCR process will be obtained from a 300dpi resolution scan in Black and White. The text should be printed or typed and on white paper. If there is a good quality image of the text then the OCR process should be relatively accurate. Handwritten text will not work unless you have specific handwritten OCR software (another topic for another whitepaper).

The barriers to good quality OCR results are:

- Text in shaded boxes
- Poor quality scan
- Highlighter
- Coloured paper
- Broken characters
- Characters touching



THE

Firſte volume of the

Chronicles of England, Scot-

lande, and Irelande.

CONTEYNING,

The deſcription and Chronicles of England, from the firſt inhabiting vnto the conqueſt

The deſcription and Chronicles of Scotland, from the firſt originall of the Scottes nation, till the yeare of our Lorde. 1571

The deſcription and Chronicles of Yrelande, likewiſe from the firſte originall of that nation, vntill the yeare. 1547.

Faithfully gathered and ſet forth, by
Raphaell Holinſhed.

AT LONDON,
Imprinted for Iohn Harriſon.

The image to the left is not of a high enough quality to offer a good OCR. Many of the characters are touching each other so the OCR engine will have a hard time separating characters. There are also background dots on the page that the OCR engine will try to make into text. Some of the characters are also broken.

Should you Try it Yourself

If you have enough OCR'ing processes to complete that it justifies buying the software then feel free to try it yourself. The Adobe Acrobat OCR process is pretty straight forward. If you want to convert paper documents into formatted word processor files it gets more complicated. We would not recommend that you try this process unless:

- You are fairly computer literate
- You have the patience to learn a fairly complex process
- You are performing OCR on a regular basis

Summary

So as you can see there is a lot more to OCR'ing than most people think. Go ahead and try it for yourself. If, however, you find it too complicated then a professional scanning organization such as Image Advantage Solutions Inc. will ensure that the job is done accurately and professionally.

For a free consultation call Image Advantage Solutions Inc. at:

Telephone: (613) 659-4620

or

Email: sales@imageadvantage.com