Optical Character Recognition

IMPACT Briefing Paper

IMPACT project
Niall Anderson, British Library

Table of Contents

Briefing Paper on Optical Character Recognition

Introduction

Optical Character Recognition (OCR) is the electronic translation of text-based images into machine-readable and editable text. Usually performed by software devices as part of a digitisation workflow, OCR works by performing a layout analysis of a digital image and breaking that image into smaller structural components to find zones of textual content. These zones will include the overall area of the page that features text. Within that zone the OCR software will identify individual lines of text, and within those lines will identify individual characters and words.

Many OCR software suites are available for many types of use, and each runs to slightly different standards and methodology. At its simplest, however, all OCR software follows the same basic principle: once the software engine has identified a single character, it runs that character's properties through an internal classification of text fonts to find a match. It repeats the process for all characters within a word, and then runs that information through a dictionary of complete words to find a match. It extends this process through sentences, lines and text blocks until – ideally – all text in the image has been identified.

Character recognition technology goes back at least to the 1920s and has had many different uses: from text-to-speech conversion to enable partially-sighted people to access printed documents, to automatically reading barcodes and cheque numbers. The following chapters treat OCR from the standpoint of an institution that is building a repository of text-based digital images. The text identified by OCR provides an important means of making those images discoverable, searchable and reusable.

Main uses of OCR

OCR has four main uses in the discovery and presentation of digital content. Broadly complementary, they are listed here in order of complexity:

- **Document retrieval -** feeding the OCR'd text to a search engine for indexing allows users to search for documents which contain the terms they are interested in.

Presenting search results in this fashion allows for OCR results to be quite poor while still making the document searchable to users. However, document retrieval of this sort can only establish if a particular

term appears in a document, not where in the document the term is found. Its usefulness is therefore limited for very long or complex documents.

- **Full text retrieval -** also relies on indexing the OCR results, but in this case the search results are presented in context to the end user. A search for a certain term will either produce a series of "highlighted" results on the digital image, or a set of context snippets from the OCR text. The OCR text is not displayed directly to the viewer, but because this method allows for a contextual view underlying OCR mistakes will be more visible to the user.

One of the most widely used standards in full text retrieval is the Portable Document Format (PDF)[1]. In a PDF, the digitised images and the corresponding text are bundled into a single file, with additional information about the position of individual words in the image. This means that users can search within a PDF and see highlighted results for their search terms. It also means that the OCR accuracy has to be higher than in basic text retrieval.

PDF documents are easy for a user to download and print, but if the image quality is high and the document contains a lot of pages, they can have very high storage implications for both the user and the institution making the PDF available.

- **Full text retrieval using enrichment -** search and retrieval may go beyond a literal match of search terms with words occurring in the documents, for instance by attempting to retrieve spelling variants, inflected forms, synonyms, or approximate matches of the search terms. The need for enrichment is especially urgent for historical language.

- **Presenting the full OCR text to the user -** here the OCR result is shown to the end user as a representation of the original document. The major consideration is OCR accuracy. If words are not represented accurately by OCR, users will not be able to find what they need and will lose confidence in the resource. Accuracy has to be very high and cannot normally be assured without human intervention, which has obvious time and cost implications.

A relatively recent development in this regard has been the growth of cooperative correction systems, where raw OCR is presented to the user via a web browser and users correct the errors – helping to make the item or the resource more discoverable for other users. The most prominent experiment in this field so far has been the Australian National Libraries Newspaper digitisation project[2], but IMPACT is contributing to the development of this idea with its CONCERT Tool[3].

- **Full text representation with xml mark-up -** full text representation is rarely done without some xml mark-up to preserve details of layout, structure or provenance, but xml mark-up can also be used to allow end users to annotate or tag items of interest, or otherwise link them to relevant items within an electronic corpus.

A common standard in full-text representation is METS/ALTO, which is a combination of metadata standards originally designed for displaying the OCR results of newspapers (a relatively complex layout type). The METS[4] file contains data about the physical or structural layout of the source document, while the ALTO[5] file contains the OCR result and structural information about individual pages. The combination of METS and ALTO makes it possible to not only search for and highlight individual words on the page, but also whole sections or articles, even parts of articles, images, tables, or charts.

[1] *Abobe PDF Technology Center*; 2009; Adobe Systems Ltd: http://www.adobe.com/devnet/pdf/pdf_reference.html Retrieved 12.03.2010
[2] *Many Hands Make Light Work;* Holley, R; 2009: http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf Retrieved 12.03.2011.
[3] *IBM and EU Partner to Enable the Digitization of Historic European Texts on a Massive Scale*; IBM Press Release; 2010; http://www-03.ibm.com/press/us/en/pressrelease/32380.wss Retrieved 12.03.2011
[4] Metadata Encoding and Transmission Standard Homepage; 2009; Library of Congress: http://www.loc.gov/standards/mets/ Retrieved 12.03.2010
[5] Analysed Layout and Text Object description; 2009; Content Conversion Specialists (CCS) GMbH: http://www.content-conversion.com/alto/ Retrieved 12.03.2010

Unlike PDFs, however, METS/ALTO files usually depend on server-side software and stylesheets for presentation in a browser, and are less suitable for distribution to the general public.

Main challenges in OCR

The accuracy of an image's OCR result is greatly dependent on the properties of the image itself, the means by which the image was created, and the properties of the underlying source material and text. As a general rule OCR will produce its most accurate results if:

- the layout of the text is simple, with no tables or illustrations;
- the text itself is in a modern, computer-generated typeface;
- the digital image preserves a high contrast between the text block and non-text detail (including blank space);
- the image has been created from a perfectly flat and straight scan (if a digital copy from an analogue source);
- the text of the analogue source is clear, well aligned and consistently presented;
- the basic material of the analogue source is undamaged;
- the text is in a single language;
- the image has been taken from the original physical source and not a degraded surrogate (such as a microfilm)

If any of these qualities is lacking OCR accuracy will suffer to some extent, and combinations will decrease accuracy further. OCR is also a significant cost within a digitisation project and institutions engaged in digitisation will need to decide whether OCR is an appropriate or feasible way of making their collections available to their stakeholders.

Whether to use OCR

As the above demonstrates, even the most up-to-date OCR software run on the ideal material cannot guarantee 100% accurate text recognition in all cases.

For smaller digitisation projects, therefore, it may be cheaper and more efficient to have typists key in the text to be electronically displayed. In the context of large-scale and mass digitisation, however, manual rekeying of material is most often too costly and time-consuming.

Any digitisation project investigating OCR should have three main considerations:

- **Suitability of material for OCR -** some types of material are better suited to OCR than others. For instance, handwritten or manuscript material will rarely give the uniformity of character needed for OCR software to be effective; badly damaged paper can lead to machine-unreadable text, etc. But if a project decides to OCR, the second consideration comes into play: namely, the level of accuracy needed in order to deliver a satisfying experience to end users.
- **Accuracy threshold needed -**OCR software packages promise a certain level of accuracy within defined parameters. They will tend to define accuracy as a percentage figure of characters recognised correctly, per the total volume of characters converted. Moreover, the accuracy will tend to have been measured on an 'ideal' document and will not therefore give a true indication of how the software will work on historical material.

In addition, character accuracy may not be the most the best indicator of the overall usefulness of OCR results. Simon Tanner of King's College, London explains:

> For example: a page of 500 words with 2,500 characters. If the OCR engine gives a result of 98% accuracy this equals 50 characters incorrect. However, looked at in word terms this could convert to 50 words incorrect (one character per word) and thus in word accuracy would equal 90% accuracy. If 25 words are inaccurate (2 characters on average per word) then this gives 95% in word accuracy terms. If 10 words were inaccurate (average of 5 characters per word) then the word accuracy is 98%. In terms of effort and usefulness the word accuracy matters more than the character accuracy – we can see the possibility of 5 times the effort to correct to 100% across the word accuracy range shown in this simple example.[6]

Many institutions engaged in mass digitisation of historic text materials using OCR will set an acceptable threshold of OCR accuracy in advance of scanning, and visually check the OCR accuracy against that target on random batches of material. Most institutions will factor this checking into the workflow of their digitisation projects and programmes (i.e. it is an ongoing task through the digitisation lifecycle). For a one-off project, however, it is usually less costly and time consuming to digitise a smaller selection of relevant material first and test the practical OCR accuracy on this sample. This has the distinct advantage of pointing up the potential problems presented by the material and/or image capturing process, and allows an institution to change its workflow or capture standards to take these problems into account before ramping up to full production.

- **Potential further use of OCR results -** if OCR results are being used primarily for indexing and retrieving online or through a catalogue (i.e. where the user will never see the actual OCR results), the absolute accuracy of the results is sometimes less significant. Search engines such as Google employ fuzzy searching, where a misspelled or misrecognised word will be matched against the actual word it most resembles, so a strict 90% character accuracy level can nevertheless result in a 98% retrieval rate. Using OCR as an indexing tool is increasingly common practice in large-scale and mass digitisation.

[6] *Deciding Whether Optical Character Recognition is Feasible*, Tanner, S., 2004 http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf Retrieved 11.03.2011