

Optical Character Recognition for Digital
Collections

GLIS 642 Final Paper

Sara Janes

For Eun Park

April 4, 2006

Introduction:

Optical Character Recognition (OCR) is a technique used “to translate pictures of characters into a standard encoding scheme, representing them in ASCII or Unicode.”¹ OCR software will take the image of a printed page and recognise each individual character, thus converting the image into digital text.

The current environment of digital libraries and archives, and electronic communications, cannot afford to neglect the vast quantities of material still existing only in print format.² Digital preservation projects provide access to these print documents in electronic format. However, for effective access, it is not only necessary that digital images of materials exist, but also that they can be found by users. OCR enables access by making it possible to index and search the textual content of images.

OCR technology is now being used in many major applications, including JSTOR (which has made over twenty million pages of journal text available since 1995)³ and the Making of America project (which has made available 3.5 million pages of nineteenth- century print material.)⁴ Using OCR to create digital text versions of these materials allows users to search their entire text, rather than relying on metadata for access to the information they contain.

1 “Optical character recognition,” Wikipedia

2 Droettboom et al, 2002.

3 JSTOR: Facts and Figures, <http://www.jstor.org/about/facts.html>

4 Making of America

Improvements in character recognition, and in the searching of OCR text, will help to make both archival and published materials useful and accessible to users of digital collections.

Literature Review:

OCR Basics

The steps of OCR, as described by Garris et al (1998) are as such:

1. Preprocessing (including line isolation)
2. Character segmentation
3. Segment reconstruction
4. Character recognition
5. Word and phrase reconstruction⁵

Most discussion of advances and challenges in OCR technology deal with steps 2 and 4: segmentation and character recognition.

For the most part, segmentation of characters is determined topologically. Preprocessing converts the document image to a black and white image; sets of black pixels which are connected form one character. Allowances are made for characters such as “i” and “j” which are composed of two disconnected segments.

Character recognition can be performed by two different classes of algorithms. Template matching is the technique of matching a character image pixel-by-pixel against a template image. The template which the character best matches is then assumed to represent the same symbol as

⁵ Garris, Wilson, and Blue, 1998.

the character image. Structural analysis decomposes characters into components (vertical lines, horizontal lines, end points) and then matches the combination of features against the features of standard character shapes. Modern OCR systems generally use a combination of both of these methods.

The problem of character recognition was originally considered to be an easy problem to solve. The first patent for OCR-related technology was granted in 1929, and significant experimentation with OCR systems began in the 1950s.⁶ However, it was found that the translation of images into text, which comes so easily to people, is a difficult process to express mathematically.

Progress has been driven by the need for OCR and by developments in computer technology. Many commercial OCR software packages are available and are being used in business applications. These are able to translate text with accuracies reaching 99% or better. However, many of these systems are optimised to only deal with text of the “best quality” -- clear, recent printed material with no degradation or distortion. The application of OCR to digital preservation projects still faces many challenges, including dealing with poor-quality printed text and manuscript materials.

Many systems have been developed to deal with difficult material using advanced algorithms or neural networks. While many of these have been successful, they consume a lot of memory and time, and are thus

⁶ Mori, Suen, and Yamamoto, 1992.

expensive and impractical for use by institutions attempting to mount digital preservation projects.⁷

OCR Text for Access and Retrieval

In the context of a digital preservation project, optical character recognition is used to provide access. OCR creates a record of the full text of a document, which can be fully searched by users. Without this full text, the user depends on metadata for access; thus, effective OCR may reduce the quantity of metadata needed for recovery of material.⁸

The effectiveness of retrieval of OCR text depends on its accuracy. Even the most accurate OCR software will occasionally misrecognise characters. In long documents, the effect of OCR error may be insignificant; Taghva et al (1994) found that OCR text corrected automatically by dictionary methods was retrieved as effectively as OCR text corrected by hand.⁹ In shorter documents, errors may cause more significant retrieval problems.¹⁰ In a digital library, most of the documents will be full-length books, and thus this is not a worry. However, archives which plan to digitise large numbers of short documents must be concerned.

In longer documents, the search term appears many times, so misrecognition of characters and words should not interfere with Boolean

7 Garris, Wilson, and Blue, 1998.

8 Tseng, 2001.

9 Taghva et al, 1994.

10 Tseng, 2001.

searching of documents. What becomes problematic is the ranking of documents. Term weighting techniques are based on the number of times a target word appears in the text; when misspellings occur frequently, this weighting becomes unstable. Tseng (2001) had success using an n-gram technique: the index is formed of sets of consecutive characters of length n, so that words which differ by only one character receive “partial match” values.¹¹

Because of the difficulty in performing OCR, indexing and searching handwritten documents is even more problematic. Manmatha et al (1996) describe a technique of “wordspotting,” in which the software compares and matches individual words within the document without attempting to recognise them. Then, a set of the most frequently occurring words would be identified manually.¹² This method, which essentially predates the digital archive movement, does not work for full-text searching, only indexing. However, if full text cannot be produced, a thorough index will still help to provide access to the documents.

Digital preservation projects dealing with heritage materials face particular challenges which are often not encountered by other OCR applications. These projects must obtain accurate digital text from materials which are not well suited to scanning or to recognition of characters or words. Older text may be degraded in quality, with “broken

¹¹ Tseng, 2001.

¹² Manmatha et al, 1996.

characters” (which a segmentation program will read as two separate characters) or suffer from noise or low contrast. Also, one must consider the large amount of textual heritage materials in manuscript form, which is difficult to OCR accurately.

Challenge: Degraded Text

Older or poor- quality printed text frequently presents difficult challenge to commercial OCR systems. Characters which are normally printed continuously instead appear to be composed of several unconnected parts, due to fading below binarization thresholds. Various context- specific techniques have been used to solve this problem, including checking dictionaries. However, these techniques do not work on mixed- language documents and are inherently restrictive.

Droettboom (2003) uses a method which tests all possible combinations of the components of these “broken characters,” and then determines which possible set of characters has the best mean confidence across the entire word. He was able to reach 93% accuracy on the broken characters independent of type of document; while this is below the standard accuracy for commercially viable OCR systems, it is a significant improvement.¹³

Xu and Nagy (1999) also discuss the problems of degraded page quality. They conclude that this does not create a fundamental barrier to accurate OCR, as software can be trained sufficiently on any character

¹³ Droettboom, 2003.

set. However, their techniques, and Droettboom's, do depend on having an accurate (human-generated) transcript for a part of the document.¹⁴

Challenge: Manuscript

Reading handprinted or handwritten materials is a challenge for OCR systems. The primary difference between printing and handprinting, in terms of character recognition, is that in mechanically printed text the characters are much more uniform. Characters written by hand will vary in shape, size, and skew. This variation among the same character in one sample makes it difficult to train software to recognise text.

It is much more difficult to properly segment handwritten text than printed. In certain cases, it can even be challenging to isolate individual lines and words.¹⁵ Segmentation of letters within words is a particular problem, as it is impossible to segment by contiguous characters as one would for print or handprinting. Han and Sethi (1995) suggest an iterative approach: their program uses knowledge of average width of characters to segment the word, then attempt to recognise characters, and depending on those results either re-segment or move on.

Recognition of characters in handwriting is also difficult, because of the extreme variability within and between handwriting styles. Structural analysis methods work best, as characters can be classified according to regular features (such as bars and loops) or singularity

¹⁴ Xu and Nagy, 1999.

¹⁵ Garris, Wilson, and Blue, 1998

features (such as end points and crossing points.)¹⁶

Recognition of handwriting has had some success in experimental applications, but it is still error-prone. The biggest challenge to attempting to create a universal handwriting recognition system is the great diversity in penmanship. It can be suggested that a system such as Gamera (see below) which can be intensively trained to recognise a particular hand might be more effective.

A Comparison of Two OCR Programs

To complement the literature research, I evaluated the performance of two OCR systems on somewhat degraded print and on manuscript.

The commercial OCR program I tested was ReadIris 11, for which a demo version is available for download.¹⁷ This program is comparable in price and advertised accuracy to other major OCR engines such as OmniPage or PrimeOCR. It is expected to have high accuracy for print documents, and should also be able to recognise handprinting. ReadIris features an “interactive learning” feature, which allows the user to correct questionable character classifications in real time.

I also tested the open source OCR engine Gamera (“Generalised Algorithms and Methods for Enhancement and Restoration of Archives”), which is freely available to any interested user.¹⁸ It is a cross-platform program based on Python scripting. The idea behind Gamera is that

16 Han and Sethi, 1995

17 IRIS OCR Software

18 The Gamera Project Homepage

rather than create one universal OCR program which can handle any type of material, it is more effective to design a basic structure which then can be optimised by document experts to process a particular type of material. “The goal is to leverage the user’s knowledge of the target documents to create custom applications rather than attempting to meet the needs of diverse users with a monolithic application.”¹⁹ In effect, the user creates the character dictionary for comparison and has control over all algorithms used for segmentation and recognition. I note that having had very little experience with this software I was not able to use all of the features to best deal with the challenging textual samples.

For a sample of print, I used several scanned pages from the 1888 Chambers' Encyclopaedia of Universal Knowledge.²⁰ These scans are of good quality (400 ppi, grayscale) but they show some slight text degradation in the original. Many of the characters are segmented and some distortion appears close to the spine of the book. This material is a reasonable example of older (though not archaic) print material which might be scanned for a digital preservation project and for which accurate OCR would be important.

The manuscript sample consists of five pages of my own handprinting, produced as neatly and regularly as possible. By using only capital letters, I made sure it fit the guidelines of what is supported by ReadIris 11 and limited the character set for classification. The text

19 Droettboom & al, 2002

20 Personal correspondence with Erik Fitzpatrick of Vikipedia (March 22, 2006)

consists of the first six paragraphs of H.G. Wells' The War of the Worlds,²¹ a sample which contains all English letters except “Z” and several numeric characters.

Camera on Print

The first page of text was divided into approximately 12000 characters by a simple continuity algorithm (all contiguous pixels form a character; those separated by white space are separate characters.) Using the interactive classifier, I was able to identify 6093 characters to use as classifier glyphs. This process took approximately five hours, including processing time for the “group and guess” function. The low percentage of glyphs identified is due to the high number of partial-character glyphs (such as dots for “i”s and “j”s) and punctuation marks, which I chose not to identify at this stage in experimentation.

Having finished creating a set of classifier glyphs from the first page, I then loaded a second page of the Encyclopaedia. This page again had approximately 12000 characters. Running the “group and guess” function on this page using the previous classifier set took several minutes.

The level of accuracy of identification was mixed. For common characters, particularly the standard small letters, accuracy was very high. Characters with smaller classifier sample sets, such as the capitals

²¹ The full text of this novel is available digitally from several sources, including Project Gutenberg

and italic characters, had much lower accuracy. Due to the large number of characters and my lack of experience with the program, I did not obtain a numerical value representing accuracy.

Running a second page through the interactive classifier in order to create a stronger set of classification glyphs would take less time than the first page, and likely increase the accuracy of classification further. My impression is that with proper training the Gamera system should be able to match commercial OCR engines in terms of accuracy on the printed page.

ReadIris 11 on Print

I used the “interactive learning” feature on ReadIris to train the program on the same image as I used to train Gamera, and then ran the engine on the second page. This process took much less time than with Gamera, even when using interactive learning.

ReadIris did not seem to be able to deal as well with words distorted by page folds, nor could it recognise ligatures such as “ae” or “fi”. The other major difficulty was that it did not allow the user to redefine how the segmentation of characters had been chosen.

Approximately seventy words in total had been corrupted, implying a somewhat higher rate of character misclassification. Considering the total number of characters on the test page, this result falls within the expected accuracy range (120 misclassified characters out of 12000

implies a 99% accuracy.)

It should be noted that the interactive learning had very little effect on the recognition of characters in this sample.

Gamera on Manuscript

In training Gamera on the manuscript sample, from three pages the total number of characters examined was approximately 3000 and I was able to extract 2410 glyphs to the classifier. The test page was found to have about 800 characters – excluding punctuation and disconnected characters, approximately 600 characters total. After classifying the final page, I found only sixty misclassifications: thus, after only a few hours and three pages of training, Gamera achieved an accuracy rate of 90% on handprinting. This rate is comparable to that achieved on the print sample, especially considering the lower number of classifier glyphs used.

I postulate that an even higher accuracy rate would be possible with the same amount of training if all letters used had been consistently connected in my own writing ('E' and 'H' frequently were recognised as multiple characters) and had this sample included no numeric characters. However, most manuscript texts to be recognised will have peculiar challenges which interfere with OCR accuracy, and each of these may have to be overcome individually. It is impossible to alter existing documents to make them more amenable to OCR; thus, it is necessary to

solve these problems technologically instead.

ReadIris 11 on Manuscript

I ran ReadIris 11 on the same sample handprinting pages, enabling the interactive learning feature. This is one of the few commercial programs which advertises handprinting- recognition capability, and recognition of handprinted characters was for the most part good.

The accuracy for this program was similar to that achieved by Gamera, with forty- two words total corrupted: considering that many of the problems stemmed from errors in segmentation, it became unhelpful to count the total number of character errors.

The primary difference in performances between these two systems seems not to be the quality of recognition of individual characters, but the importance of the learning mechanism. Gamera cannot classify any characters without bootstrapping from characters the user has classified by hand. ReadIris performs nearly as well without the “interactive learning” tool as with.

Suggestions for Future Work:

My experimentation with the Gamera program was educational, but I am aware that I was not able to use all the features it offers. The software is still in development, so it can be expected that future

versions will be more powerful and, eventually, more accessible.

In its current form, the software is not encouraging to the new user. That alone may be incentive for an institution to decide to instead purchase more well-packaged and intuitive software, even though it may be fundamentally capable of less.

I plan to continue working with the Gamera software, to better determine its abilities and limitations. I do agree with its developers that an application which can be trained by specialists to best deal with a particular difficult type of document is more promising than the somewhat nebulous idea of a universal OCR program.

To date, most large collections of digital text have included primarily published material. A few digitally available collections of private papers have relied on manual transcription of the text.²² With improvements in OCR abilities to render manuscript into digital text, archives will be able to open many more of their collections to the worldwide community. This will provide access to materials which otherwise may have remained hidden.

OCR technology has proven itself useful in non-textual applications, including the use of bar codes. It is entirely possible that automated recognition of images can some day be applied to good use for collections of (non-textual) images. In order to reach that point, however, much more research must be done on how humans “read” and

²² E.g. The Newton Project

understand visual materials.

Ideally, a state could be reached when “machine- readable” becomes synonymous with “human- readable.” As computers can more easily process human communication and recordings, the interactions between computers and their users will become smoother and more natural.

Conclusion

Optical Character Recognition (OCR) is a useful tool for an institution mounting a digital preservation project. It allows users to search and browse the full text of digitally- preserved document images, providing full access to the information presented.

OCR is a challenging field, as systems must be developed to properly deal with a wide variety of textual materials. The success of the technology for rendering clear, printed text as digital text is encouraging, but it must be noted that there remain many unsolved problems.

If systems can be developed to deal accurately and efficiently with challenging materials such as manuscript text, then users of digital preservation projects will be given effective access to a new set of materials. Progress has been made in this field, and experimental systems have been successful; all so far looks encouraging. OCR is a technology which institutions making textual resources available digitally should not ignore, nor take for granted.

References

- Assessing the Costs of Conversion. The Making of America IV: The American Voice 1850- 1876. A Handbook Prepared from the Andrew W. Mellon Foundation. University of Michigan Digital Library Services, July 2001.
- Droettboom, Michael. "Correcting Broken Characters in the Recognition of Historical Printed Documents." Joint Conference on Digital Libraries: Association for Computing Machinery, 2003.
- Droettboom, Michael, et al. "Using the Gamera Framework for the Recognition of Cultural Heritage Materials." Joint Conference on Digital Libraries: Association for Computing Machinery, 2002.
- Fenton, Eileen Gifford. "An OCR Case Study." Handbook for Digital Projects: A Management Tool for Preservation and Access. Ed. Maxine K. Sitts. Andover, MA: Northeast Document Conservation Center, 2000.
- The Gamera Project Homepage. 2006. Johns Hopkins University. April 3, 2006. <<http://ldp.library.jhu.edu/projects/gamera>>
- Garris, Michael D., Charles L. Wilson, and James L. Blue. "Neural Network-Based Systems for Handprint Ocr Applications." IEEE Transactions on Image Processing 7.8 (1998): 1097- 112.

- Han, Ke, and Ishwar K. Sethi. "Off- Line Cursive Handwriting Segmentation." Document Analysis and Recognition 2 (1995): 894-97.
- IRIS – OCR Software. 2006. I.R.I.S. April 3, 2006. <www.irislink.com>
- Junker, Markus, and Ranier Hoch. "An Experimental Evaluation of Ocr Text Representations for Learning Document Classifiers." International Journal on Document Analysis and Recognition 1 (1998): 116- 22.
- “Optical Character Recognition” Wikipedia, the Free Encyclopaedia. April 3, 2006.
<http://en.wikipedia.org/wiki/Optical_character_recognition>
- Manmatha, R, et al. "Indexing Handwriting Using Word Matching." First ACM International Conference on Digital Libraries. Bethesda, Maryland: Association for Computing Machinery, 1996.
- Mori, Shunji, Ching Y. Suen, and Kazuhiko Yamamoto. "Historical Review of OCR Research and Development." Proceedings of the IEEE 80.7 (1992): 1029- 58.
- Taghva, Kazem, et al. "The Effects of Noisy Data on Text Retrieval." Journal of the American Society for Information Science 45.1 (1994): 50- 58.
- Trier, Øvind Due, Anil K Jain, and Torfinn Taxt. "Feature Extraction Methods for Character Recognition - a Survey." Pattern Recognition 29.4 (1996): 641- 62.

Tseng, Yuen- Hsien. "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text." Journal of the American Society for Information Science and Technology 52.5 (2001): 378-90.

Xu, Yihong, and George Nagy. "Prototype Extraction and Adaptive OCR." IEEE Transactions on Pattern Analysis and Machine Intelligence 21.12 (1999): 1280- 96.

Digital Collections Consulted

JSTOR – The Scholarly Journal Archive. 2006. April 3, 2006.

<www.jstor.org>

Making of America. 2005. April 3, 2006.

<<http://www.hti.umich.edu/m/moagrp>>

The Newton Project. 2005. The Newton Project. April 3, 2006.

<<http://www.newtonproject.ic.ac.uk/>>

Project Gutenberg. 2006. Project Gutenberg Literary Archive Foundation.

April 3, 2006. <<http://www.gutenberg.org/>>

Vickipedia. Ed. Erik Fitzpatrick. 2006. April 3, 2006.

<<http://vickipedia.livejournal.com>>