Introduction

Historical Archives are in varying stages of digitising collections for the purposes of preserving originals and increasing availability. Documents may be hand-written, machine-printed or a combination of both (for example, registry information), as well as being in varying states of degradation. The challenge facing custodians is not just taking high-quality images of the originals, but in extracting as much information as possible from the digital versions.

The nature of historical documents means simply extracting the text may not tell the whole story of the document. There are many other observations which could be made about a single or group of documents such as layout, style of writing, penmanship of the author or graphical elements which are not picked up by standard handwriting recognition programs. There are also many issues faced with the image-processing aspects of document analysis, before they are even in a state to attempt handwriting recognition. This study focuses on the ***difficulties faced during extraction of information from digital images of handwritten historical documents***.

Paper 1: "Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen"

Aims: The authors present their techniques for analysing the layout of handwritten index pages. They propose the use of this system for creating a searchable index of these documents. Primarily, they wish to formalise the previously mooted, but un-refined 'droplet technique' for handling ascenders and descender characters.

Reference: The paper utilises the generally accepted framework for image-processing of handwritten documents, along with existing methods for recognised printed numerical characters. For their specific purpose and document layout, they combine techniques for extracting both machine-printed and hand-written information.

Methods: A top-down approach was able to be used, due to the structured, tabular and regularly-spaced nature of the documents, which would not be present in all types of handwritten document. Existing proven techniques are used for text line recognition, offering no new advances in this area. Moore's widely recognised 'shortest-path algorithm' is used efficiently in a new setting to decide what action to take with ascending and descending characters, formalising its technique for use in line segmentation.

Outcome: The system for layout detection is quantitively proved to work with an error rate of <1%, although the authors concede that the structured nature of their documents means they have not had to face issues which less-structured documents would have presented (e.g. text skew, poor preservation) and so these results are not necessarily comparable with similar systems.

Reflection: The research shows that both existing and new techniques can be successfully used in series and that – given a specific format – a highly accurate layout analyser can be provided. The paper offers little-to-no complicating factors in the layout/standard of the documents being analysed, making comparison between this and other techniques difficult. However, the droplet technique is based on efficient accuracy, which will be useful in any proposed system as part of my own research.

Paper 2: "Access by Content to Handwritten Archive Documents: Generic Document Recognition Method and Platform for Annotations"

Aims: The authors investigate a system for 'generic' document use, rather than the domain-specific applications which have come before. Their biggest feature is intended to be the ability to store both information automatically detected in the document and user-entered annotations – both of which will be searchable.

Reference: This paper builds on existing applications of document and annotation management (such as the machine-learning WISDOM++ software) by using it with tabular and general historical documents. The project works on a top-down proposal, rather than piecing together domain-specific systems from other modular sub-systems. This approach is highly uncommon as most new systems are designed for a defined and specific purpose.

Methods: The authors have looked at the historical documents field and designed a list of ideals for the new system, including elements of both machine-print and hand-writing recognition software, combined with searchable annotations and the ability to handle tabular, structured and unstructured documents. In addition, an existing method of character recognition by analysing graphemes is utilised on titles and large surnames on forms.

Outcome: At the time of writing the paper, over 170,000 images had been scanned of different types, but still mostly structured data despite the claim that the system can also handle unstructured data. They admit that analysis takes around four-seconds per image, which is remarkably slow in computing terms and would need to be reduced as part of any future work.

Reflection: The developed system for generic documents – and idea of combining automatic and collaborative annotations – has a refreshing feel when compared to other constructed document handling systems. The authors state that their system can be more easily customised to new types of document than existing systems, although no evidence of this in practice is provided. The scalability and modular nature of this system is significant in that it creates an adaptable, but ultimately generic system. The high reliance on manual annotations to display information is a negative and further automatic-read information would drastically improve on this project.

Paper 3: "Machine Learning Methods for Automatically Processing Historical Documents: from Paper Acquisition to XML Transformation"

Aims: This paper is part of a larger COLLATE project, which aims to design and create a web-based collaborative repository for archive and cultural documents. In this paper, the authors propose the use of an existing document processing system, WISDOM++, which uses machine learning to transform documents into XML format.

Reference: The only new element of this research is the algorithms designed to be applied to document classification via WISDOM++, which are only useful for the specific types of documents they are detecting. It is implied that the larger project investigates using machine learning to design further algorithms, but this is not addressed.

Methods: The existing WISDOM++ system is used as-standard to return a wireframe-style document layout analysis. The authors then apply algorithms to classify documents based on their layouts. The document sections are then automatically labelled with section titles, ready for feeding to an OCR (text recognition) system.

Outcome: Only one set of quantitative results are given – for a single document type which report mixed accuracy results of detecting the location of known elements of that document. The results are mixed, showing near 100% recognition of machine-printed elements, 92-96% for handwritten names but only 74% for signatures. Lack of further quantitative results prevents full analysis from being made, although the authors state the results are 'promising' but in need of 'further investigation'.

Reflection: This report seems 'lightweight' in nature as very little new ground is broken or useful evidence is produced that algorithms can be designed for use in WISDOM++, a promising document-management system which is worth investigating as part of my own research. Further investigation to see if other researchers have developed efficient algorithms for WISDOM++ is a priority.

Paper 4: "Fast Handwriting Recognition for Indexing Historical Documents"

Aims: Propose a computer architecture system for decreasing the time taken to search a large lexicon for the best match to a handwritten word image, with no decrease in accuracy.

Reference: Building on recognised methods of speeding up processing, then combining ideas. The authors intend for their faster architecture to be used by other researchers to improve detection accuracy, which is recognised to drop significantly as the lexicon size increases (as low as 58% for a 20,000-word lexicon).

Methods: Existing techniques in processor speed optimisation are analysed for their individual gains on a given task, then combined to form a new architecture which is also tested on the same task. Final architecture is based on parallel processing, each processor finding the most probable match from its section of the lexicon. A final process compares the identified word from each processor and makes final decision.

Outcome: The headline claim is that they were able to reduce the time to recognise a single word (from a 20,000-word lexicon) from 6.755 seconds to 0.376 seconds, although the architecture also increased from 1 to 4 processors. The research does not address any other parts of the document analysis process, such as extracting and generating an image of the input word, despite their system requiring high quality input images to maintain accuracy.

Reflection: The authors state that 'handwriting recognition is the most important part of document analysis' and this is certainly true for the process of extracting data from the text of a document. The research performed in this paper to propose a faster architecture is very useful, but only if combined with a similarly high-quality document layout and text-line segmentation system. Any time savings are only a positive, provided they do are not couple with loss of accuracy.

Paper 5: "Text Line Segmentation of Historical Documents: A Survey"

Aims: Study available text-line segmentation methods and analyse their benefits and appropriate uses. Includes studies of multiple languages, including non-Latin characters for an all-encompassing overview.

Reference: The paper does not necessarily build on or contribute new knowledge to the field but compares uses and applications of other research in the area, proposing the areas in which they believe the best gains are to be made from current and future research.#

Methods: Splitting existing methods into categories by general technique (including the most widely-used 'smearing' technique), the authors address the pros and cons of each technique in general, before mentioning particular research which has improved knowledge in that area, or is notable for reasons of increased compatibility, intended use or increased functionality – such as the ability to process skewed lines, overlapping characters or unstructured documents.

Outcome: The general conclusion is that there is no one system – or technique – which will suit all needs. The authors admit that machine learning is the next step in this field, as well as a type of 'feedback' loop where detected document characteristics control elements of the text-line segmentation process. They also conclude that text-line segmentation is only one piece of the overall document-analysis picture and no studies have been made (at the time of writing) into the overall process.

Reflection: As an overview, this is an interesting and useful study which confirms that the 'Holy Grail', one-system-fits-all approach is currently unobtainable. However, it confirms that machine learning is the logical next step in research, leading me to lean my own research into this area.

Paper 6: "Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map"

Aims: The authors propose a method to extract text from a gray-scale document as an alternative to the standardly-accepted method of Binarization. They seek to show that not only is it possible, but that less data will be lost from the document due to filtering in the binarization process.

Reference: The authors have previously published other work on document analysis of gray-scale documents, which is a relatively untouched area of historical document processing. In this paper they propose an analogue-equivalent of the binary 'smearing' method.

Methods: A mathematical formula is applied to each pixel within a grayscale document to create an 'ALCM Transform' image – essentially smearing nearby pixels to locate areas of dense text. After this point, identical techniques to those used with binary images are implemented to complete the document analysis.

Outcome: No quantitative results are produced and with a sample size of only 30 documents, this data would not be accurate in any case. Qualitative reports show that the authors identify some shortcomings themselves, the most notable being errors in identifying non-full lines of text.

Reflection: Despite only being tested on a handful of documents, the idea of leaving documents as gray-scale has obvious data-retention benefits and negates the need to create high-functioning binarization systems as a pre-processor to document analysis. A generic system would surely need to have this functionality, however the simplistic methods adopted here could be improved if combined with elements of machine learning. This is worth considering my own research.

References

Bulacu, M., van Koert, R., Schomaker, L., & van der Zant, T. (2007). Layout Analysis of Handwritten Historical Documents for searching the archive of the Cabinet of the Dutch Queen. *Proceedings of 9th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 357-361). Curitiba, Brazil: IEEE.

Couasnon, B., Camillerapp, J., & Leplumey, I. (2007). Access by Content to Handwritten Archive Documents: Generic Document Recognition Method and Platform for Annotations. *International Journal of Document Analaysis and Recognition (IJDAR)*, 223-242.

Esposito, F., Malerba, D., Semeraro, G., Ferilli, S., Altamura,)., Basile, T., . . . Di Mauro, N. (2004). Machine Learning Methods for automatically processing Historical Documents: from paper acquisition to XML Transformation. *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL)* (pp. 328-335). Palo Alto, USA: IEEE.

Govindaraju, V., & Xue, H. (2004). Fast Handwriting Recognition for Indexing Historical Documents. *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL)* (pp. 314-320). Palo Alto, USA: IEEE.

Likforman-Sulem, L., Zahour, A., & Taconet, B. (2007). Text Line Segmentation of Historical Documents: A Survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 123-138.

Shi, Z., Sethur, S., & Govindaraju, V. (2005). Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map. *Eighth International Conference on Document Analysis and Recognition (ICDAR)* (pp. 794-798). Seoul, South Korea: IEEE.