Keynote presentation given at the Swedish Symposium on Image Analysis 2010 (SSBA2010), Uppsala, Sweden, March 11-12, 2010.

Large-Scale Digitisation and Recognition of Historical Documents: Challenges and Opportunities for Image Processing and Analysis

Apostolos Antonacopoulos
Pattern Recognition and Image Analysis (PRImA) Lab, University of Salford, Greater Manchester, UK
Email: A.Antonacopoulos@primaresearch.org

***Abstract*—This paper outlines the challenges and opportunities in large-scale analysis and recognition of scanned historical documents. After a brief overview of the background of large-scale digitisation and its overall challenges, the characteristics of the documents and the artefacts encountered are presented according to how they occur throughout the lifecycle of a document. The stages of full-text digitization are presented next, with an emphasis on the document image analysis pipeline. The paper ends with some pointers to past and current notable digitization research initiatives.**

I. Introduction

Major digitization efforts are taking place around the world in order to preserve, study, showcase and – most of all – make available to the wider public the information contained in books, newspapers and other documents.

National Libraries are leading the way but even so, large-scale digitization is a relatively new activity. While special large-budget projects have taken place for a relatively long time to preserve and showcase treasures (e.g. the virtual books at the British Library [1]), it is only within the last about five years that large-scale digitization efforts have really started around the world. Such efforts pose significant challenges in terms of both technical and business (budget, legal and organisational) decisions.

The major challenges related to Image Processing and Analysis range from:

- Scanning methodology and parameters.
- Image compression decisions
- Image enhancement (noise and geometric artefacts detection and removal)
- Layout analysis (page segmentation and region classification)
- Character recognition.

The above challenges are present in practically any digitisation project for historical documents. What is crucial in the case of *large-scale* digitisation is the addition of the requirements that:

- All processes within a chosen pipeline scale up to cope with the required volume (typically millions of pages per year).
- Each process is capable of running essentially without human input (apart from initial setup).

The following sections will briefly touch on the characteristics of documents and on the nature of the artefacts present in the images (Section II), the image processing and analysis stages within the digitisation workflow (Section III) before offering some pointers to past and current significant initiatives in the area (Section IV).

II. Documents and Artefacts

Documents range widely from manuscripts to books to newspapers. In terms of printed material – the focus of large-scale digitisation – there are older (before the middle of the 19th century) very well made books as well as newer mass-produced ones of lower quality. In addition, there are newspapers (from the 18th to 20th centuries) which on the whole were not produced to last more than a day or a week, as evidenced by the relatively low quality of production.

In terms of print characteristics, books in a large library will be in Latin, English, German, French and other major language groups. The presence of more than one language in the text of a given book is not uncommon, even at the word level, presenting a significant recognition challenge. Typefaces also present a challenge either due to their individual nature (e.g. Gothic, or *Fraktur*) or in their combination within a book. Finally, layouts can vary from single to multicolumn, with decorative borders and various types drop capitals, among other illustrations.

To better understand the different artefacts that are present on a scanned page (and consequently, better target any possible solutions) they are classified below according to how (at which stage in the document lifecycle) they were introduced.

A. Printing issues

These are issues that were inherent in the document as soon as it was produced:

- Uneven inking: Non-uniform distribution of ink across the page resulting in very dark characters in some areas and very faint ones in other areas.
- Bleed-through: Ink from the text of the page behind has seeped through the substrate (paper, velum etc.) and is visible on the current page.

- Smear-over: Ink from the text of the opposite page, before it dried completely, has been smeared over the current page when the book was closed.
- Typesetting issues: Font peculiarities (including custom abbreviations) and layout issues.
- Paper texture: The background of the document is not smooth and uniform.

B. Use issues

These are artefacts introduced after the document was produced, by using it over a number of years. Depending on the type of document, varying levels of care are evident (usually archive documents have suffered the most).

- Folds and creases, especially in newspapers.
- Tears.
- Annotations.
- Stains.
- Damage and dirt from repeated handling, e.g. from turning pages.
- Repair attempts: clear tape or replacement paper to strengthen damaged parts of a page.
- Punch holes and marks from staples and clips.
- Scratches on microfilm (evident in cases where microfilm was scanned instead of the original paper document).

C. Preservation and storage issues

Such artefacts are the result of the ageing process and the conditions under which the documents were stored over long periods of time.

- Arbitrary warping of the substrate due to humidity in storage.
- Substrate discolouration due to acid in the paper or organic matter (e.g. mould).
- Fading ink.
- Discoloured paper (often unevenly).

D. Scanning issues

The following artefacts are the only ones introduced during the digitisation process and, as such, there is a possibility of eliminating them (by adopting a different strategy, where possible).

- Page curl: a different issue from arbitrary warping (a storage issue, as mentioned earlier), this is the curvature evident towards the spine of a bound volume. To avoid this, volumes could be unbound (not a viable solution for most libraries) before scanning, or a different scanning methodology can be used. In practice, the presence of page curl is frequent.
- Skew: this can be uniform across the page (e.g. a simple rotation of the whole image can eliminate it) or varying, caused usually by scanner feeders whose rollers on either side of the document operate at slightly different speeds.
- Show-through: unlike bleed-through (ink seepage from the page behind), this is caused by shining light through the page behind. In principle, show-through can be eliminated by placing a dark card behind the page to be scanned. However, this is not possible when robotic book scanners, for instance, are used.
- Other illumination artefacts: shadows, for instance, can be caused by the ripples of arbitrary warping on the page.

III. Full-Text Digitisation Wokflow

Libraries go to great lengths to plan and execute their digitisation projects. The following are steps followed (in principle by the libraries:

- **Document preparation.** Physical examination, recording of artefacts and entering metadata.
- **Scanning (and perhaps recognition – OCR)**. This step, the most interesting to our research community, is usually subcontracted. Quality specifications and expectations vary a lot depending on budget and contractor capabilities.
- **Examination of documents**. After scanning this is to establish whether documents were damaged in any way.
- **Examination of scans**. This quality check can only be done for relatively small samples.
- **Examination of recognised text**. Again this is only performed for small samples.
- **Hosting / presentation.** Images are usually shown to readers and text is used for searching (resource discovery).

It is interesting to note that the scanning and recognition step tends to cost almost as much as all the other steps combined. That step, can be broken down to the following stages, relevant to image processing and analysis:

- **Scanning.** Resolution, colour depth and compression decisions play a significant role.
- **Image enhancement.** To facilitate further steps and to create a better viewing experience for the readers. Typical operations include, *page splitting* (if more than one page is present in the scan), image *border removal* (cropping to the page), *geometrical artefact correction* (deskewing and dewarping), *noise removal* and *binarisation* (if required by further stages).
- **Layout analysis.** This stage includes *page segmentation* (identification of coherent regions of interest: blocks, textlines, words and characters) and *region classification* (identification of the type of the contents of each region e.g. text, graphics, line art etc.).
- **Character recognition.**

- **Post-processing.** Improvement of OCR results using dictionaries (in addition to those used by OCR itself), named-entity gazetteers and other background knowledge. Crowdsourcing is also an option for large-scale distributed correction of text (usually by members of the public).

There is significant room for improvement in all the above steps. Methods must be robust to artefacts of historical documents and, as mentioned earlier, be fast and require almost no human intervention. Business decisions may dictate that quality expectations be lowered slightly in favour of the above properties.

IV. Past and Current Research Initiatives

Past projects have focussed either on proof-of-concept experimental methods or have concentrated on specific types of documents only. Some notable such projects:

- METAe - Metadata Engine [2]: Research on extracting relatively simple layout information and metadata. OCR is trained on Fraktur.
- DEBORA (Digital AccEss to BOoks of the ReNAissance) [3]: Research prototypes of document image analysis methods developed to demonstrate feasibility of concepts.
- MEMORIAL [4]: Prototype of a complete system for full-text digitisation of archive documents.
- VIADOCS [5], COLLATE [6] and others: Research in digitization of specific types of semi-structured archive documents.

Commercial initiatives such as Google Book Search [7] and the now abandoned Microsoft Live Book Search [8] also include some historical books. It is the author's observation, though, that Google Book Search does not (on the whole) place significant emphasis on the quality standards required of national libraries and it is rather secretive of the methodology used.

A major 4-year EU-funded project, IMPACT [9], is under way to improve the large-scale extraction and enrichment of text from historical documents (books and newspapers). It is the first such project where most national libraries in Europe collaborate with document analysis researchers to improve every step of the digitisation process.

Acknowledgment

The work described here has been supported in part through the European Union 7th Framework Programme grant IMPACT (Ref: 215064).

References

[1] http://www.bl.uk/onlinegallery/virtualbooks/index.html
[2] http://meta-e.aib.uni-linz.ac.at/
[3] http://www.comp.lancs.ac.uk/computing/research/cseg/projects/debora/
[4] http://www.memorialweb.de/
[5] http://www.nhm.ac.uk/research-curation/research/projects/lepindex/aboutproject.html
[6] http://www.cultivate-int.org/issue6/collate/
[7] http://books.google.com/books
[8] http://en.wikipedia.org/wiki/Live_Search_Books
[9] http://www.impact-project.eu/