Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005),
Seoul, Korea, August 29 - September 1, 2005, IEEE, pp. 48-53

Semantics-Based Content Extraction in Typewritten Historical Documents[†]

A. Antonacopoulos and D. Karatzas

PRImA Lab, School of Computing, Science and Engineering, University of Salford,
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract

This paper presents a flexible approach to extracting content from scanned historical documents using semantic information. The final electronic document is the result of a "digital historical document lifecycle" process, where the expert knowledge of the historian/archivist user is incorporated at different stages. Results show that such a conversion strategy aided by (expert) user-specified semantic information and which enables the processing of individual parts of the document in a specialised way, produces superior (in a variety of significant ways) results than document analysis and understanding techniques devised for contemporary documents.

1 Introduction

The conversion of historical documents from their ageing (paper or other substrate) form to a suitable electronic representation is an endeavour that gives rise to significant issues that are not usually encountered in the processing of contemporary documents.

One can observe that historical documents suffer from artefacts due to ageing, extended use and previous preservation attempts all of which variably affect the visual quality of the information (e.g., text) and which are not straightforward to always specify or predict (the nature and effects of these artefacts are examined in more detail later on in this paper). Such conditions seriously challenge traditional document image analysis methods that assume (effectively) a smooth background and uniform quality of writing, for instance.

Equally importantly, there are considerable (non-image related) issues that govern and characterise such conversion attempts. These can be broadly referred to as issues related to semantics of the content and to the representation/use of the extracted information.

First, there are issues stemming from the semantic richness of information contained in historical documents. The original content of the document is important not only for the *prima facie* message it contains but also for its historical context, and different levels of interpretation that are possible given knowledge about people, places and events that are referred to directly or indirectly. For instance, some text may actually be a person's name and, depending on which section of a given document type it appears, that name may belong to a man or a woman, a prisoner or military officer etc.

A further distinguishing characteristic of historical documents is that, in addition to the original content, most frequently they contain evidence of the *use* of the document as well (which must be preserved and interested as well). The semantic evidence of this use (as opposed to the physical evidence in terms of use-specific degradations) is present in the form of annotations and other foreground elements added later by people other than the author of the document.

Second, the exploitation of the information extracted from historical documents is most often considerably different to that of the information extracted from contemporary documents. This fact impacts on the conversion process as the resulting document representation has to be flexible and comprehensive enough to cater for the required use scenarios.

Finally, it must be noted that historians need to see the information as contained in the document, without any alteration of the text, for instance (i.e., mis-spelt words in the original must not be corrected during document analysis).

It can be concluded from the above that a conversion strategy aided by (expert) user-specified semantic information that processes individual parts of the document in a specialised way, will inevitably produce better (in a variety of significant ways) results than document analysis techniques devised for contemporary documents.

This paper presents a flexible approach to extracting content from scanned historical documents using semantic information. The final electronic document is the result of a "digital historical document lifecycle" process, where the expert knowledge of the historian/archivist user is incorporated at different stages with the aid of a number of interactive tools. These tools together with others that

[†] This work has been supported through the EU grant IST-2001-33441.

are automated (e.g., image analysis) form part of the "Digital Document Workbench" developed as a the result of the MEMORIAL project (www.memorial-project.info).

The flexibility of the approach further manifests itself in its ability to combine the segmentation of entities at different levels (regions, lines, words or individual characters) with the application of thresholding methods that are more suited to each level, depending on the input document. Such a locally adaptive process with access to semantic information is best suited (and perhaps the only realistic approach) to the segmentation and enhancement of text and (ultimately) to achieving higher recognition results by OCR.

The digital historical document lifecycle is described in section 3. In Section 4 the focus of this paper, the content extraction phase, is introduced. The main emphasis is on the segmentation stage, which is described in detail. Results showing the effectiveness of the whole approach are presented and discussed in Section 5.

2 Comparison with previous work

There is scarcely any report in the literature of conversion of this type of typewritten documents into a logically indexed, searchable form. A notable exception is a project that involved the digitisation of (mostly typewritten) index cards with a bank-cheque scanner and the subsequent curator-assisted extraction and recognition of taxonomic terms and annotations [1].

The work described in this paper is of a different nature and necessitates a wider-ranging approach. First, many of the documents (as in most archives) are fragile, and curators heavily resist mass scanning. Second, the paper is frequently damaged by use and decay and, sometimes, heavily stained. Third, the characters typed on the paper may not be the result of direct impression but of impression through the original paper and a carbon sheet as well (characters in carbon copies are frequently blurred and joined together). Fourth, there may not be as ordered a logical structure in the text and position of documents as in a taxonomy card index, for instance (although there usually is some logical information that historians / archivists are able to specify). Finally, the volume of text and the relatively unrestricted dictionary possibilities evident in many of the documents does not permit the use of experimental (purpose-built) OCR. An off-the-shelf OCR package must be used and optimised (as far as possible) instead.

3 Digital historical document life-cycle

The MEMORIAL project has introduced a *Digital Document Life-Cycle (DDLC)* development model supported by a specially developed *Digital Document Workbench (DDW)* toolset.

The left arm of the DDLC model (see Figure 1) represents analysis of information aided by the user, whose domain knowledge is gradually being transformed into a control structure of processes for engineering the final product, represented by the right arm.

The first phase of the DDLC model is *digitization*, which yields a raw digital image of a paper original. This process may be performed as an entirely manual activity, e.g., photographing extremely rare documents with a digital camera, as well as batch scanning of more sturdy documents (e.g. inventory cards of a museum) with an automatic document feeder scanner. The most important task of this phase is to assure proper scanning parameters, for otherwise if the quality of scanned images turns out to be unsatisfactory later on during the cycle, taking paper documents out from the archive to scan them again may be impossible or costly. A Repository Management Tool (RMT) has been developed to aid the archivist in the management of scanned images.

The next phase of DDLC is *qualification*, when documents similar in structure, purpose and meaning are grouped into semantic classes. Documents within the same semantic class can be processed throughout the rest of the cycle in a specific way, "tuned" individually for each class defined. This has been made possible by introducing two concepts: a *document template*, and *phase tuning* (the latter is achieved through rigorous evaluation of quality at different levels using a Quality Evaluation of electronic Documents tool (QED), described elsewhere [2]).

A template is an XML file, specifying formally the document layout and content in a form that is both machine readable and can be directly manipulated by an expert user knowing document semantics. It shall be noted that the distinction between classes is at the level of a single page, since a class template combines in a specific way both the page layout and the content of a page, as explained later. This distinction, however, is introduced at the archivist's discretion, who interprets a document and defines a template. A range of DDW component tools have been developed to operate on document templates, including a template Electronic Document eDitor (EDD) tool for generating template XML files. An intuitive graphical interface provides a fair separation of a document expert from XML intricacy.

The document template is consulted by the *segmentation* and *extraction* phases, where textual regions in the raw document image are segmented and improved by the Image Processing Tool (IPT), resulting in "clean" bilevel image regions that are processed by OCR and a *document content* XML file is produced.

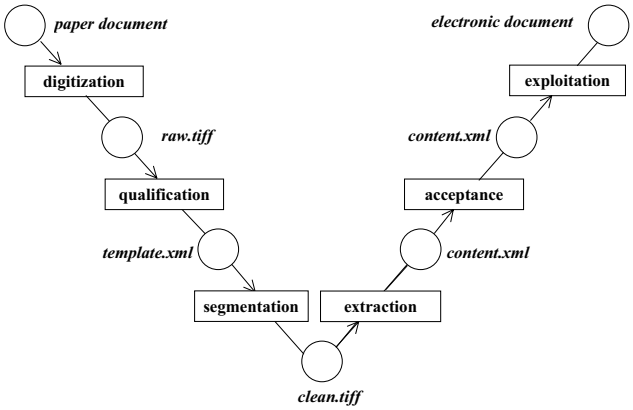

Fig. 1. Digital historical document life-cycle.

The subsequent *acceptance* phase introduces again expert user interaction for two reasons. First, the content XML file generated by OCR may contain incorrectly recognized characters; therefore loading such an electronic document into the target database (digital archive) will reduce the quality of information available during the subsequent *exploitation* phase. Second, it is imperative that the historian inspects any automated correction of the original characters by OCR, for instance, or some geographical names misspelled by international authors. The correction of the document content in the first case requires modification of the content, and is supported by the content editor Generator of Electronic Documents (GED). In the second case only annotation is allowed, and is supported by the multivalent browser Viewer of Electronic Documents (VED).

4 Content extraction

At this phase (incorporating the "segmentation" and the "extraction" stages of DDLC) the textual content of the document is extracted. The segmentation stage is responsible for the optimal location and preparation (enhancement) of textual entities in the image in order for the subsequent OCR stage to produce the best possible results. For completeness, it should be noted that generic commercial OCR alone performs poorly on this type of documents, and results shown later in this paper support this.

As mentioned earlier, information contained within the document template file, which provides generic information about a class of similarly structured documents, is consulted during the segmentation stage, and a document content template is produced to describe more accurately the specific document being analysed. Information regarding the existence and actual positioning of the document regions is added to the content template. The content template is then used to control the subsequent stage of OCR, during which it is further refined to include the recognised text. By the end of the content extraction phase, the content template provides a detailed description of the document.

4.1 Document characteristics

Historical documents share certain common characteristics in terms of artefacts present in the source image. Most commonly these will be artefacts related to ageing (e.g., discolouration, disintegration of parts etc.) and repeated use (e.g., dirt, punch holes, tears, rust from metallic clips etc.).

In the document classes encountered within the MEMORIAL project, additional characteristics are observed, typical of typewritten documents. First, each character in the document is *individually* produced by pressing the corresponding typewriter key. In contrast to printed documents, each individual character in the document may appear stronger of more faint than its neighbours. The difference can be considerable in some cases.

Second, a typewritten document may not be produced in its entirety at once. Instead the paper may be removed at some point and reinserted to make corrections and further additions. This can result in a non-uniform skew angle throughout the document and in non-uniform spacing between text lines.

Finally, it was typical in the case of official documents to produce a carbon copy at the same time. Usually, the carbon copy was produced on a very thin paper (a.k.a. Japanese paper) which has prominent texture. Due to the mechanical nature of this process (the force from the typewriter key has to be transferred through the original paper and through the carbon sheet before a character is produced on the carbon copy) the characters of the carbon copy are usually blurred and faint.

4.2 Segmentation of non-content regions

The purpose of this step is to remove any areas (e.g., scanner borders and reconstructed paper regions) in the image that cannot possibly contain any useful information and at the same time have inconsistent characteristics with the remainder of the document area [3].

4.3 Text segmentation and enhancement

Central to the phase of content extraction is the derivation of a suitably enhanced bi-level version of the textual regions in the document image. As mentioned earlier this is a necessity when using commercial OCR whose optimal results naturally depend on the quality of its input. Therefore, the processes involved in this stage and their measure of success are determined by the quality of

OCR results. This translates to an attempt by the segmentation stage to minimise instances of problems that give rise to OCR errors. More specifically, the result of this stage should minimise the existence of filled-in, merged and broken characters.

Given the fact that historical typewritten documents suffer from the presence of various problems, as explained in Section 4.1, there is a significant need for a flexible approach.

Flexibility in the approach described here is achieved in three ways. First, different thresholding techniques can be used in different situations. Second, segmentation can be performed at a number of different levels (region, line, word and character (very important especially for typewritten documents). Finally, depending on the situation, an iterative process can take place where progressively finer segmentation-thresholding steps ensure better results.

The combination of segmentation level and thresholding method (afforded by the flexibility of this approach) has been studied by the authors and it can be observed that each class of documents can benefit from the selection of a different processing path (a particular thresholding method applied individually at a specific segmentation level).

Global thresholding, as expected, failed to produce reasonable results, even at the level of individual characters. This is due to numerous reasons, for instance the considerable amount of background noise present in the document, the ageing conditions (which often produced an uneven background), the different strength by which each character was typed etc.

Locally adaptive methods proved to be more effective when applied to individual words or characters. This is expected since, due to the nature of these methods, in areas where no actual foreground pixels (e.g., characters) are present, background noise is usually labelled as part of the foreground. This effect can be minimized by accurately segmenting text regions, thus making sure that most of the background area is excluded before thresholding. Therefore, locally adaptive thresholding methods tend to produce better results at the segmentation level of individual words or characters.

It has further been observed that certain local thresholding methods in some documents perform better at the level of individual words than that of individual characters. This might appear at first inconsistent, as (in typewritten documents) a word might comprise characters typed with different strengths. Nevertheless, given the area of a single character, there are cases when there is not enough background available for the thresholding method to make a sufficiently good threshold judgement.

Indicative segmentation results are shown in Figure 2 for two document classes with different characteristics (for brevity, only the levels of region and character are shown). Results from applying different thresholding methods at different segmentation levels are shown in Figure 3 (again, for brevity, results on lines and words are omitted). Note that in the "transport list" class, the best results are obtained by applying the GPP [2] method at the region level, whereas in the "catalogue card" class, the application of Niblack [5] thresholding at the individual character level yields the best results.

4.3.1 Region-level. Initially, each of the text regions indicated (loosely described) in the document template is mapped to the actual content of the document being processed and the region boundaries are adjusted accordingly. To this end a thresholded image of the loosely described region is used. If the segmentation level defined in the processing path is that of individual regions, thresholding (using the thresholding algorithm specified) takes place over the whole region (now precisely described), and the process finishes here. In any other case, segmentation continues by identifying lines of text in the region.

4.3.2 Line-level. To identify lines of text in a given text region, the vertical projection profile of the text region is calculated in the greyscale image. The text line segmentation process is then based on analysing the patterns of local minima (ideally representing white space between lines) and maxima (indicating a high count of foreground pixels) in the histogram.

The analysis takes into account information about the expected size of a character box (e.g., dimensions of typewritten character) but allows for some degree of flexibility. In addition, each text line identified is verified by examining previously identified lines, since adjacent text lines are expected to be separated by similar distances.

The projection profile analysis method avoids the use of absolute thresholds and, therefore, is largely uninfluenced by uneven illumination artefacts and moderate amounts of skew.

If the segmentation level defined in the processing path is that of individual text lines, each text line is thresholded using the method specified, and the process ends at this point. In any other case the segmentation proceeds further, to identify individual words.

4.3.3 Word-level. Given a text line box, the horizontal projection profile is calculated and analysed to identify suitable spaces between words. The left and right delimiters of each word are thus located.

The top and bottom boundaries of a word are initially identified as equal to the top and bottom of the text line rectangle, but are subsequently adjusted to fit the given

word as tightly as possible. The reason for doing that is twofold. First, it is beneficial to the subsequent processes (thresholding or further segmentation) to exclude as much background space as possible from the word box. Second, by adjusting the top and bottom of each word individually, the skew angle of the text line can be followed closely, word by word.

The process of the adjustment of the top and bottom of the word rectangle starts by identifying tighter top and bottom word boundaries within the bilevel image of the word (thresholded using the specified method). The expected height of a character is taken into account during this process, which is capable of identifying the top and bottom boundaries even if they extend above or below the textline boundaries. The identified boundaries are then verified in the greyscale image and adjusted further if necessary.

Given the adjusted current word rectangle position, the text line rectangle (i.e., the vertical boundaries within which the next word is expected) is amended before locating the next word, so that is vertically centred to the rectangle of the current word, and extends one quarter of the height of the character box above and below it. This ensures that the skew angle of the line is followed word by word throughout the text line.

If the segmentation level specified is that of individual words, the segmentation process stops here and a bi-level image of each word is produced by thresholding using the algorithm specified. Otherwise, the last segmentation step is to locate individual characters inside each of the identified words.

4.3.4 Character-level. For this step, the horizontal projection profile of each word is calculated. As the information is rather limited (only one character height), care must be taken in the ensuing analysis to ignore the effects of noise while taking into account valid (but possibly faint) characters, possibly diacritical marks.

Since the word boundary is accurate enough to exclude any inter-word space, the first character separator is predicted at a distance equal to the expected character box width from the left edge of the word rectangle, and is adjusted according to the location and strength of histogram minima. Based on the adjusted position of the first character separator, the next one is predicted at a distance equal to the given character box width and adjusted accordingly, and the same procedure is repeated for the length of the word rectangle.

Finally, each individual character is thresholded using the algorithm specified in the processing path.

4.4 Character recognition and further stages

The OCR process (OCE DOKuStar V3.6) is provided with the enhanced image file and the location of each logical entity (from the intermediate content template). At the end of this step, the recognised characters are inserted in the content XML structure.

Individual dictionaries have been created with valid text for different semantic entities. For instance, a dictionary of first (proper) names and a dictionary of place-names (with a variety of spellings and transliterations) is being used to improve the recognition rate for the corresponding semantic entities in the relevant document classes. It must be noted that the ability to apply different recognition parameters to different semantic entities, is only possible due to the semantics-rich document architecture devised.

This stage completes the automated content extraction phase. As mentioned earlier, the resulting content is validated and augmented in the subsequent acceptance stage of the digital historical document life-cycle before being exploited in the variety of use scenarios outlined in the beginning of this paper.

5 Results and discussion

This paper has presented the case for the conversion of historical documents into an electronic representation, based on semantic information provided by the expert (e.g., historian) user. The approach devised to recover the content from the scanned images was described, focusing especially on its flexibility of specifying combinations of segmentation level and locally adaptive thresholding method for optimal results.

The effectiveness of the whole approach is assessed by evaluating acceptance-testing scenarios. The scenario related to the uniform model only is discussed here, for reasons of brevity. The model takes into account four metrics: the average OCR confidence level (as output by the package), the percentage of correctly recognised characters, the percentage of correctly recognised words and, finally, the document preparation time ratio (indicating time/cost savings as opposed to human transcription). Quality (effectiveness of the system) is expressed in the range of 0–1. Three different cases are compared in terms of quality value: the direct application of the off-the-shelf package to the document, the application of the OCR package following thresholding by Otsu's method [6], and finally, the comprehensive approach of the MEMORIAL project.

The uniform model (graph shown in Fig. 4) applies equal weights to each of the metrics. It is evident that the whole approach constitutes an overall improvement to both the manual transcription and to the semi-automated

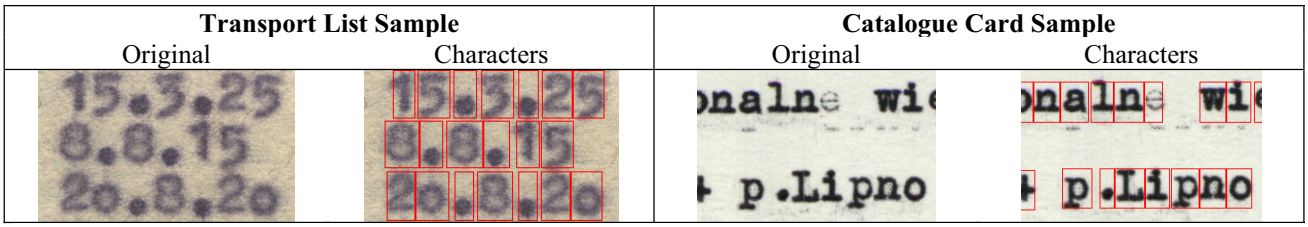

Fig. 2. Segmentation at the level of characters for two document classes.

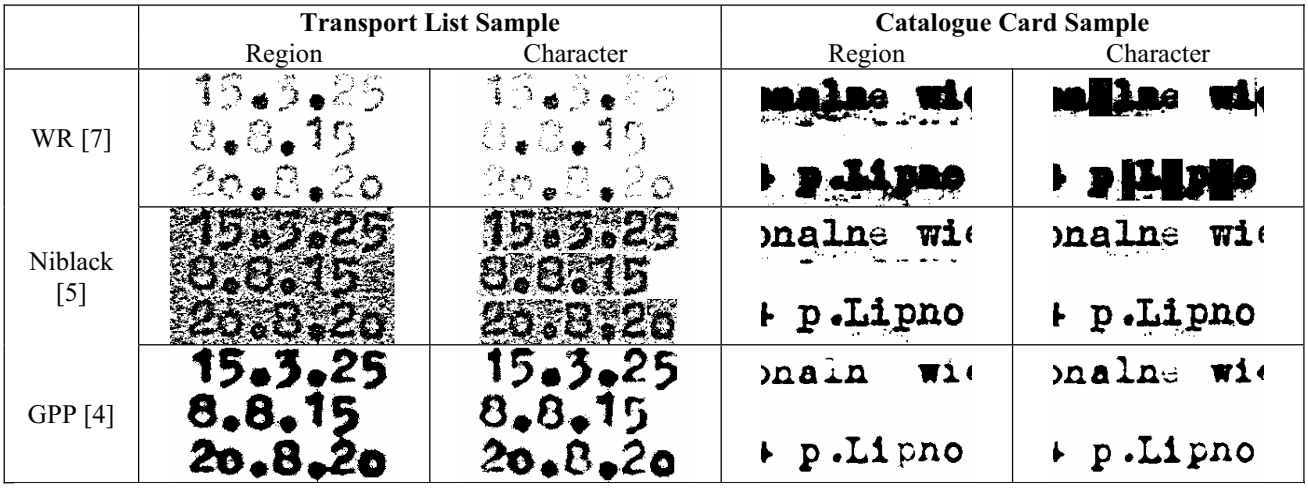

Fig. 3. Results of applying different processing paths to two document classes.

application of off-the-shelf packages. Moreover, the richness of information (semantically tagged) obtained by the approach described here is far superior to the output of generic OCR.

Acknowledgment

The authors wish to acknowledge the work of H. Krawczyk and B. Wiszniewski (partners in the MEMORIAL project) who were responsible for the non-document image analysis aspects.

References

[1] A. Downton, S. Lucas, G. Patoulas, G. Beccaloni, M. Scoble, G. Robinson, "Computerising Natural History Card Archives", *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Edinburgh, UK, August 3–6, 2003, pp. 354–358.

[2] A. Antonacopoulos, D. Karatzas, H. Krawczyk and B. Wiszniewski, "The Lifecycle of a Digital Historical Document: Structure and Content", *Proceedings of the ACM Symposium on Document Engineering (DocEng2004)*, Milwaukee, USA, October 28–30, 2004.

[3] A. Antonacopoulos and D. Karatzas, "A Complete Approach to the Conversion of Typewritten Historical Documents for Digital Archives", *Proceedings of the 6th IAPR Workshop on Document Analysis System (DAS 2004)*, Florence, Italy, September 8-10, 2004, Springer LNCS, pp. 90–101.

[4] B. Gatos, I. Pratikakis, and S.J. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents", Proceedings of the 6th IAPR Workshop on Document Analysis System (DAS 2004), Florence, Italy, September 8-10, 2004, Springer LNCS (3163), pp 102-113.

[5] W. Niblack, *An Introduction to Digital Image Processing*. Prentice-Hall, London, 1986.

[6] N. Otsu, "A threshold selection method from gray-level histograms" *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-9 (1979), pp. 62–66.

[7] J.S. Weszka and A. Rosenfeld, "Threshold Evaluation Techniques", *IEEE Transactions On Systems, Man and Cybernetics*, Vol. SMC-8, IEEE, 1978, pp. 622-629.

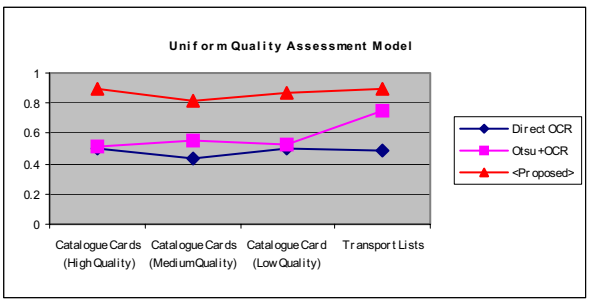

Fig. 4. Uniform quality assessment model graph.