Proceedings of the International Workshop on Document Image Analysis for Libraries (DIAL2004), Palo Alto, January 2004, IEEE Computer Society Press, pp. 336-341.

Document Image Analysis for World War II Personal Records[†]

A. Antonacopoulos and D. Karatzas
Pattern Recognition and Image Analysis (PRImA) Group, Department of Computer Science
University of Liverpool, Liverpool L69 3BX, United Kingdom
http://www.csc.liv.ac.uk/~prima

Abstract

Complete collections of invaluable documents of unique historical and political significance are decaying and at the same time they are virtually inaccessible, necessitating the invention of robust and efficient methods for their conversion into a searchable electronic form. This paper presents the issues encountered and problems addressed in the MEMORIAL project, whose goal is the establishment of a digital document workbench enabling the creation of distributed virtual archives based on documents existing in libraries, archives, museums, memorials, and public record offices. Successful approaches are described in the context of the chosen data class: a variety of typewritten documents containing personal information relating to the presence of individuals in World War II Nazi concentration camps.

1. Introduction

There is a significant need to analyse and index old and historical documents into digital libraries. Such documents are dispersed across different institutions of varying financial and technical means and pose a number of problems in their conversion to searchable electronic form. It is not surprising, therefore, that complete collections of invaluable documents are not accessible on-line.

Contrary perhaps to popular belief, vast numbers of documents of historical importance are not necessarily very old (cf. medieval documents). In fact, relatively "modern" ink and paper deteriorates very fast. In addition, historical documents are not only interesting to historians and other scholars, sometimes they play vital political and administrative roles.

The document classes involved in the project described in this paper are typically characterised by the above aspects. The documents concerned are decaying, unique sources (no copies exist elsewhere) of personal information that constitute the only record and proof of existence for many thousands of people during the dark years of Nazi occupation in Europe. Thousands of documents containing this type of information are closely guarded in various individual museums and archives making access to the whole body of knowledge contained in them virtually inaccessible. This is a significant problem as, apart from researchers, there are government agencies that initiate queries regarding the existence and trail of individuals during World War II. All such queries are executed manually at present by an archivist/historian searching through the documents available on each individual site.

The significance of the project and of the document image analysis methods involved, in particular, expands far beyond the chosen document class into practically any typewritten document (which includes enormous numbers of documents of the 20th century).

The most common approach taken by archives and libraries is to at least digitise (scan) the paper documents (to at least preserve them) even if no further document analyses processes take place. 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 to convert file cards from the archives of the Natural History Museum in London, UK [1]. This project 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. The results are indexed hierarchically in a database.

The work described in this paper is of a different nature and necessitates a relatively different approach. First, many of the documents (as in most archives) are fragile, and mass scanning is heavily resisted by curators. 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). Finally, there many not be as ordered a logical structure in the text and position of documents as in a card index, for instance (although there usually is some logical information that historians / archivists are able to specify).

The implication of the above issues is that there is a requirement for more involved and, at the same time, more generic document analysis. The volume of text and

[†] This work is supported by the European Union grant IST-2001-33441.

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 is used, for which the characters are segmented and individually enhanced in advance by the methods developed for the project.

The project is in mid-way at this point in time and the methods described here have only yielded preliminary (but largely representative) results.

The remainder of the paper presents an overview of the project (Section 2) and of each individual stage involved. More precisely, Section 3 outlines the digitisation and document structure definition processes. Section 4 describes each of the following document image analysis steps: pre-processing, background removal, character location and character improvement. The character recognition and post-processing stage is summarised in Section 5, while a brief description of the final web-enabled system is given in Section 6 before the paper concludes with Section 7.

2. The MEMORIAL project

The "MEMORIAL" (http://www.memorialweb.net) project is funded (€1.5M) by the European Union (Fifth Framework Programme – Information Society Technologies priority) and undertaken by a multi-national consortium. The overall goal of the project is to enable the creation of distributed virtual archives based on documents existing in libraries, archives, museums, memorials, and public record offices. The full title of the project is: "A Digital Document Workbench for Preservation of Personal Records in Virtual Memorials", which also hints to the nature of the document dataset selected for study: documents that contain information about people.

The document classes selected for analysis are taken from World War II Nazi-run concentration camps. The Stutthof camp (http://www.stutthof.pl) provides the following selection of representative documents:

- **Transport lists** – 3 kinds of documents, where names of Stutthof prisoners are present – 1) lists of prisoners that arrived at Stutthof, 2) lists of prisoners that were moved from Stutthof to other camps, and 3) lists of prisoners freed from the camp. A sample transport list can be seen in Fig. 1. Similar transport lists (as compiled by the SS) exist in a number of museums and archives throughout Europe.
- **Catalogue cards** – these are cards created by historians shortly after the end of the war and contain information about people from various sources (as typewritten text and stamps). There are about 200,000 of these cards in Stutthof alone and, in terms of the project, provide a different class of documents but with typewritten text fields to test the applicability of the framework developed. An example catalogue card can be seen in Fig. 2.

Fig 1. A transport list of prisoners moved to camp on 12 February 1944.

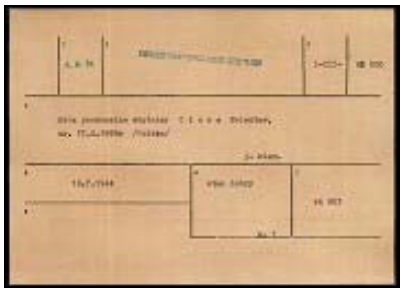

Fig 2. A catalogue card.

The remainder of this paper will concentrate on the broad class of transport lists.

3. Document input

This section describes the steps preceding the document image analysis stage, which is the main concern of this paper. The two steps of scanning and document structure definition are only briefly described.

3.1. Scanning

The transport lists exist in a variety of physical conditions (in terms of damage and decay) and most often they are the actual duplicate pages (carbon copies) produced when the original lists were typed by the SS (and later destroyed by the same people upon fleeing at the end of the war). As such, the surviving documents are printed on rice (a.k.a. Japanese) paper which is very thin, rendering the use of an automatic document feeder impossible.

There are other important issues raised by the type of the paper (see Fig. 1). First, the presence of a background texture. There is a texture which is very prominent as multi-colour noise in the colour scans. Scanning can potentially improve the quality of the resulting image (studies were carried out with a variety of scanners and set-ups), although historians prefer to see a facsimile of the original (with the background texture) when they study the documents. The decision was made to retain the fidelity of the scanned documents to the originals and to place the burden on the image analysis stage.

Second, the text in the transport lists is produced after each key of the typewriter has hit the original paper, then the carbon sheet and finally the rice paper. As such, characters are not sharply defined but blurred and often faint (when the force was not enough to carry through to the rice paper or the quality of the carbon was not so good. The decision was made to scan the pages without any attempt to improve them during scanning and, therefore, defer processing to the image analysis stage.

One requirement imposed for consistency of further processing, is that documents are scanned against a dark colour background (covering the scanning area surrounding the paper document).

Documents are scanned in 300dpi, in 24-bit colour TIFF format (with lossless compression) and, using an indexing tool (created as part of the project) entered into a working repository with simple metadata attached.

3.2. Document structure definition

For many classes of archive documents, it is possible to define a correspondence between the physical and the logical structure of a document. This is especially the case for documents having a fixed form-type physical structure. In the case of the transport lists, there is definitely a logical structure (in an oversimplified way: certain blocks of information followed by lists of personal information, followed by certain closing blocks of information) but there is not a fixed layout correspondence. It is the responsibility of the historian/archivist user of the final system to group together very similar documents and, using a tool developed by the consortium, create a *template* where physical (generic) entities on a page are associated with logical information.

The subsequent extraction of the textual content of the document is guided by the document template. The template is an initially "empty" container XML structure. This structure specifies generic regions (as rectangles) of a predefined type of content (e.g. table block, salutation block etc.). Region specifications are interpreted by the document image analysis methods to extract precise text regions (e.g. textlines, table cells etc.) from the scanned document pages. This (geometrical parameter) information is then inserted into respective regions of the XML template to produce *content* XML files ("filled" document structure) - one for each respective page of input.

The OCR process subsequently looks up each of the content XML regions in correspondence with the enhanced image in order to resolve problems with recognising individual words and groups of characters (using corresponding dictionaries – peoples' first names, geographical place names etc.). The recognised text (possibly after additional human editing) is then inserted in the appropriate positions in the content XML, completing the document conversion process

4. Document image analysis

The main goal of the image analysis stage is to prepare the ground for optimal OCR performance (compensating for off-the-shelf OCR inadequacy to deal with the document class in hand). This goal is in reality twofold. First, the quality of the image data has to be improved to the largest possible extent afforded by the application. Starting with a colour scanned document with a large number of artefacts (noisy background, paper discolouration, creases, and blurred, merged and faint text, to name but a few) the result must be a bi-level improved image where characters are enhanced (segmented, restored and faint ones retrieved from the background) as much as possible.

Second, individual semantic entities must be precisely located and described in the content XML structure. The required level of abstraction for the semantic entities is defined in advance. The document template XML structure coarsely outlines the location of regions in the image (e.g., there is a table region contained within a given notional rectangle). The document image analysis methods must locate the required instances of logical entities (e.g., individual table cells) and enter this information in the content XML. Using the resulting content XML as a map, the OCR process can be directed to that specific location in the image and fill-in the recognised text into the content XML structure.

Fig. 3. A sample scanned document.

The main steps of the document image analysis stage are *pre-processing*, *background removal*, *character location* and *character improvement*. Each of these stages is described in the remainder of this section.

4.1. Pre-processing

The first step in the processing chain is to locate the paper document within the image resulting from the scanned area. As it can be seen in Fig. 3, there is a dark outer region surrounding the paper document. As mentioned earlier, this is the result of the requirement to scan each document against a dark background (to facilitate processing). In fact, the chosen colour was dark green as this is a colour that is very unlikely to occur as that of the paper document itself.

The surrounding dark region is identified by examining the Lightness component of the image (which is converted from RGB to HLS). This special consideration is necessitated by the lack of colour calibration of the different scanners used. The edges of the paper document are thus located.

Assuming that the edges of the original paper are mostly straight and pair-wise perpendicular, a first attempt is made to calculate and correct for skew. This correction step appears to be sufficient in the majority of cases.

Next, areas of reconstructed paper are identified. The presence of this type of area is an artefact resulting from earlier document restoration attempts, where missing paper (due to tears, holes etc.) is "grown" back using liquid paper (see areas along the left edge of the document in Fig. 3). Identification of such areas (of typically different colour/texture than the background) facilitates the subsequent identification and removal of the original paper background.

Reconstructed paper areas are identified in the image based on information contained in the Lightness and Saturation components of the HLS data and by performing connected component analysis, careful filtering (certain regions of printed information share similar characteristics) and morphological operations. The result of the pre-processing step can be seen in Fig. 4, where the identified surrounding area is shown in green and the reconstructed paper areas in orange.

Fig. 4. The identified surrounding area (green) and the reconstructed paper areas (orange) for the document in Fig. 3.

4.2. Background removal

A process of pivotal importance is the separation of the foreground from the background. The removal of the background enables subsequent processes to focus more detailed processing on the remaining foreground regions only. As mentioned earlier, however, there are various artefacts whose presence hinders this process.

Having removed the surrounding region (scanner background) and the regions of reconstructed paper, the remaining pixels can be analysed as part of the document itself.

Following experiments, the Lightness component of the image is chosen for this step. So far, the use of a dynamically derived threshold (based on the Lightness histogram) and its application to the image has been used (followed by morphological operations to repair resulting gaps in the background). Further experimentation is taking place, although it is not very important at this stage to precisely remove all the background pixels. In fact, the process errs on the side of caution as there are instances of faint characters surrounded by darker ones that will be removed as they exhibit many characteristics of the background (similarly, there are relatively dark artefacts that subsequent processes have to analyse and identify by

more refined processing). The result of the background removal process can be seen in Fig. 5.

Fig. 5. The document in Fig. 3 with all non-foreground elements removed (white).

4.3. Character location

In order to improve the text regions to the effect that merged characters are separated, and faint ones are "lifted" from the background, the approach described here performs an individual character location and enhancement process. This approach is novel in this type of application and is afforded by the regularity of the typewriter font.

In order to locate individual characters in the image, a top-down approach is followed. First, the regions of interest are looked up in the XML document template. This minimizes the overall processing effort required, since character location only takes place within given areas instead of the whole image (although the methods can be extended to work with the whole image). For each text region of the template, a two-step process takes place to locate the characters: first, the identification of textlines in the region is performed, and then for each textline extracted, the characters within it are segmented.

Both steps involve careful analysis of projection profiles using statistical analysis and a novel metric that takes into account the presence of noise and in some cases, due to the cautiousness of the previous processes, parts of background.

The main properties of the typewriter font exploited here are the constant font size (character height and maximum width) and the fact that there are no vertical overlaps between parts of adjacent characters. Thus, the distance between textlines is practically predictable and adjacent characters can be separated based on character width analysis.

In the current version of the system, in order to identify textlines, the horizontal projection is calculated for each text region, taking into account only the pixels labelled as foreground. A statistical analysis follows, during which a number of potential locations for text line separators (corresponding to minima in the projection profile that have a high likelihood to correspond to white space between text lines) are identified. A histogram of the distances between subsequent separators is then constructed, and the most frequently encountered distance is identified as a benchmark. Each potential separator is subsequently scored, based on whether the distance between itself and its neighbouring separators is consistent with the benchmark distance. The separator with the highest score is then selected as the first true textline separator and the rest of the potential separators are reassessed and labelled based on the benchmark distance.

Within each text line identified, a similar process takes place during which individual characters are separated. However, due to the cautiousness of the background removal process, the vertical projection for each text line is calculated taking into account all the pixels of the text line (including those initially labelled as background).

An example of individual characters precisely located within the image can be seen in Fig. 6.

4.4. Character improvement

Having identified the position of all characters in the image, localised processing can take place for each character. This processing, aims at enhancing the characters and producing a bi-level image of each character, which will be used by OCR in the next stage.

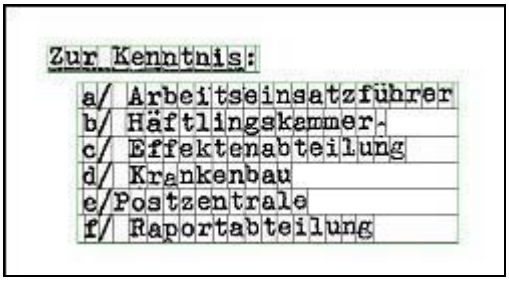

Fig. 6. Example of individual characters located and enhanced within text regions.

During this enhancement step, the full information from the original image is taken into account. This effectively means, that all pixels in the box of the character are assessed, not only the ones that were classified as foreground during the background removal step. This approach provides the opportunity to recover from any background/foreground misclassification that may have occurred during previous steps.

A number of contrast enhancement and adaptive thresholding approaches can be performed at this point. Experimentation still takes place but as an initial approach, the method proposed by Niblack [2] has been adopted. This initial decision was made after consideration of a number of alternatives (including variants of histogram equalisation techniques [3] and Weszka and Rosenfeld's [4] approach). A key characteristic of Niblack's method seems to be the accurate preservation of character edges while, however, it does not perform very well on areas of degraded background not containing any pixels of a character. A sample result can be seen in Fig. 6.

It should be noted that very encouraging results have been obtained, with merged characters correctly separated and faint characters (previously classified as background) recovered. The ability to locate individual characters constitutes a very significant benefit for any enhancement process and this is one of the characteristic advantages of this project.

5. Character recognition and post-processing

An off-the-shelf OCR package is given the enhanced image and the location of each logical entity (from the intermediate content XML structure). At the end of this step, the recognised characters are inserted in the content XML structure.

The OCR package cannot be trained directly on the document class in hand. However, the results of the OCR are post-processed taking into account the type of the logical entity to which they correspond. For instance, if the logical entity is a date, only digits and separators (e.g., hyphens) are considered and the result can be further validated. Similarly for surnames, names and placenames, although the frequent practice of using German spelling (and re-naming) of places and peoples' names make the process more complicated. Experiments are still being carried out to establish initial figures of measurable improvement in recognition rate as opposed to applying OCR to the original image (before enhancement). It is hoped that these results will be reported at the final version of this paper.

6. Web database and user access

The project will create a prototype portal where historians, government officials and the public can initiate a query through the web. The final approved content XML document structures will form the basis for extracting selected information to include in a web database application. Each type of user will be authenticated first and then receive the result of their query applied to each of the participating archives (appropriately censored according to their user status).

7. Concluding remarks

This paper has given a brief overview of the issues and solutions identified so far in the MEMORIAL project, which is still under way. A very significant class of documents, namely typewritten documents of the 20th century necessitates the invention of robust and efficient methods for their conversion into a searchable electronic form.

The document analysis methods outlined in this paper are a step towards fulfilling this need. These methods attempt to cope with a large number of artefacts present in the scanned images and, taking into consideration key characteristics of typewritten text, have yielded very promising initial results. In fact, it is possible to examine the document mage at each individual character location and apply a different enhancement process as required (as a response to quality issues arising from the fact that typed characters are produced independently of each other on a given page). Furthermore, merged characters can in the majority of cases be successfully separated, thus removing another major source of OCR error. Work continues in refining the stages of the system presented and in expanding the repertoire of enhancement methods.

References

[1] A. Downton, S. Lucas, G. Patoulas, G. Beccaloni, M. Scoble, and 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.

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

[3] M. Sonka, V. Hlavac and R. Boyle, *Image Processing, Analysis and Machine Vision*, 2nd ed, PWS Publishing, 1999.

[4] J.S. Weszka and A. Rosenfeld, "Threshold Evaluation Techniques", *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-8, pp. 622–629, 1978.

0-7695-2088-X/04 $20.00 © 2004 **IEEE**