# Preprocessing and Segmentation of Bad Quality Machine Typed Documents

Mariusz Szwoch and Wioleta Szwoch

Knowledge Engineering Department, Technical University of Gdansk,
80-952 Gdansk, Poland
{szwoch,wszwoch}@eti.pg.gda.pl

**Abstract.** The goal of the international project *Memorial* is automatic retrieval from machine typed paper documents belonging to several classes. In this paper the problem of pre-processing and segmentation of scanned archival documents is considered. The goal of these processes is to exactly determine text regions in the document for further OCR processing. Text regions are initially located for each document's class as XML templates. Then, region-matching algorithm is used to precisely locate regions in current document.

## 1 Introduction

A strategic goal of the international consortium undertaking the project *Memorial* is to enable creation of virtual archives based on documents existing in libraries, archives, museums, and public record offices [1]. To achieve this goal, the consortium focuses on computer-aided information retrieval from machine typed paper documents. Further steps, one including development of data models for the storage of retrieved information, and another, involving development of distributed virtual memorial services for navigation and information search, are planned by the consortium upon successful completion of the *Memorial* project. In the first phase of the project the archive documents from former Nazi concentration camp museums (e.g. State Museum Stutthof in Sztutowo near Gdansk) are considered.

Creation of digital libraries of archive documents requires solving many different problems connected not only with storing and management of original documents but also with automatic information retrieval from them. This information stored together with images of original documents and possibly some additional information creates complex structures called *digital documents*. Digital documents allow for advanced searching possibilities on different security levels for historians, families of victims and others.

Though, there exists several advanced OCR (*Optical Character Recognition*) systems and the structure of the machine typed documents is rather simple, the automatization of the recognition process is a difficult task bearing numerous problems of the science, technical, organizational and law nature [2].

Achieving high quality in the information retrieval (recognition) process hardly depends on applied methods and tools of image processing and recognition. In the *Me-*

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 276–285, 2004.

*morial* the professional OCR *DOKuStar* from Document Technologies GmbH is used. Numerous experiments carried out with the OCR programs revealed that the quality of recognition of documents' content strictly depends on proper image preprocessing that covers:

- image acquisition (scanning);
- document localization (layout) in whole image;
- optional skew correction and elimination of geometric distortions (mostly in digital camera acquired images);
- background and noise removing;
- localization of text regions in the document.

In this paper the research and experiments results are presented that concern archival documents' acquisition and pre-processing and also regions' localization basing on predefined documents schemes.

## 2 Acquisition of Archival Documents

The acquisition process of archival documents bears numerous problems of the science, technical, organizational and law nature. In the Memorial project a comprehensive study was carried out in order to specify correct procedures for high quality information retrieval of such documents.

### 2.1 Protection of Personal Data

Protection of personal data contained in archive documents is one of the most important aspects of the information retrieval process. In the *Memorial* project a set of false documents was prepared that could be used in some experiments and for the *Internet* presentation of the project goals and results. These test documents have been made using original typewriter and paper sheets what results in their great similarity to original patterns (Fig.1a).

However, experiments proved that this similarity is not sufficient at the image preprocessing stage. The differences between the false and original documents result from several reasons covering e.g. difficulty in simulation of such effects as mechanical damages of the paper, influence of atmospheric factors (humidity, floods, light), chemical reaction of the ink and paper, copies from other pages and many others. Example damages of original documents are presented in Fig. 1b[1].

### 2.2 Scanning

The main goal of documents scanning for the archive purposes is their faithful analog to digital conversion. However, in some cases, a traditional scanning may lose some information significant from the farther recognition process's point of view. In order

[1] Considering the demand of personal data protection all such information are removed from images presented in this paper.

to avoid such situation in the *Memorial* project, additional experiments have been carried out with alternative scanning of documents using the infrared light. Images scanned in that way are characterized by greater intensity of some image artifacts (e.g. copies from other pages) but generally their quality were distinctly worse than images scanned in traditional scanners.

Theoretically, there exist a possibility to use information from both kinds of scanned images to improve the recognition results but it requires additional storage and processing time bearing additional problem of images matching. Considering this, further research in the Memorial project has been done with different types of traditional scanners.

a)

b)

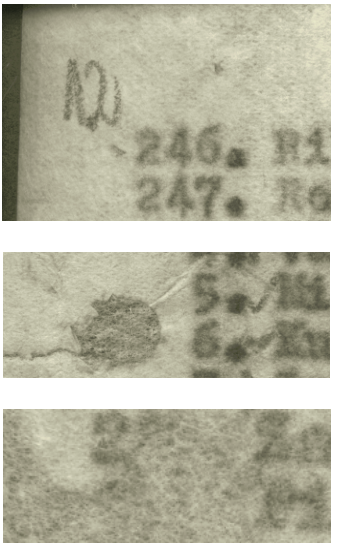

**Fig. 1.** a) false document from former Nazi concentration camp, b) fragments of original documents with example artifacts.

Optimal scanning resolution for image recognition is usually set in the range of 75-400 DPI [3]. The lower boundary of this range results from minimal width of line-shape objects occurring in the documents. These lines are expected to have at least 1 or 2 pixels in width after scanning [4]. Though, counted in this way resolution of archive documents (from concentration camp) lies between 50 and 75 DPI, in the *Memorial* project, the higher resolution of 300 DPI has been set. Using this resolution, results from the fact that scanned images are used not only for recognition but are also concurrently used for archiving purposes. These images are presented to authorized users of digital library instead of original (paper) ones.

Unfortunately, using high-resolution images increases significantly system requests for memory resources and time needed for document processing. For example, single A4 image stored as a *TrueColor* uncompressed bitmap (BMP) has a size of 26MB,

and single filter operation (window 3x3) lasts[2] about 1s. Though, all farther described algorithms had been carried out using images of 300 DPI resolution it would obviously be possible to reformulate them for documents with lower resolution which could increase their speed but decrease further localization accuracy.

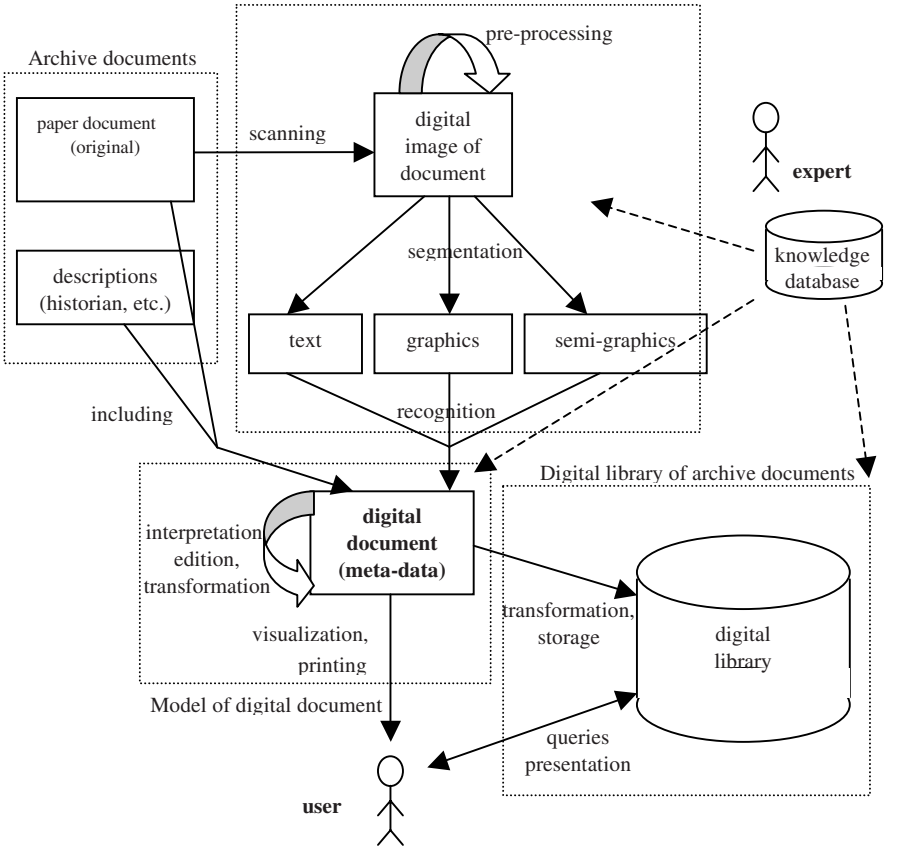


**Fig. 2.** The life cycle of digital documents.

## 3 Preprocessing of Archival Documents

The automatic document's recognition process consists of several, characteristic stages: image preprocessing, segmentation, recognition and interpretation [2]. These stages are parts of the digital document's life cycle that is presented on Fig. 2.

The efficiency and quality of the automatic recognition process is a primary condition for the final success of the whole project of digital library creation. When the only

[2] All described experiments have been carried out on PC with Athlon 2.5GHz processor with 512 MB RAM.

low recognition efficiency is achieved the only alternative is manual information retrieval from original documents.

As it was mentioned earlier, the efficiency of the OCR systems depends in high extend on proper carrying the initial stages, especially background elimination and segmentation. The experiments carried with the *DOKuStar* system, revealed its sensibility on correct localization of regions containing the text under recognition. This fact generates additional problem of exact text region localization.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE page SYSTEM "DTD\page.dtd"[]>
<page name="singlepagetransportlist" size_width="2552"
size_height="3508">
<content origin_x="0" origin_y="0">
 <region foreground_color="#000000" background_color="#ffffff"
       anchor_top="185" anchor_left="260" stretch_width="1185"
       stretch_height="265" stretch_x_tolerance="0"
       stretch_y_tolerance="0" anchor_x_tolerance="0"
       anchor_y_tolerance="0" skew_tolerance="0" skew="0" op-
   tional="False">
   <text font_name="Courier New" charset="ISO-8859-1">
     <composed_text type="">
       <line skip="265" length="0" align="unknown"
valign="unknown"></line>
     </composed_text>
   </text>
 </region>
```

**Fig. 3.** Listing of the XML schema fragment describing template regions' location.

Most recognition algorithms use some expert's knowledge that could be found in assumed preconditions, models, thresholds parameters and others. In the Memorial project the expert's knowledge is put into the schema mechanism that describes general structure of the documents of certain type with a standard text regions' localization in it. The schemas are created in XML meta-language and stay for the foundation of digital document description. The contents of each document schema fields are filled by the OCR algorithms and also by automatic and manual corrections. The example fragment of the document schema is presented on Fig.3, and location of template regions are presented on Fig. 4 a) (white rectangles).

### 3.1 Document Localization

Some amount of archive documents processed in the Memorial project has been scanned with the background of gray-green broadcloth (Fig. 4a). The first stage of processing of such documents is their location inside the image. Considering constant background characteristic (color) and unchanged light conditions (scanner) the color RGB characteristic of background has been estimated. A simple algorithm, samples the image from its edges towards the center, every M-th raw and column (M=50). Encountering N subsequent points (N=3..10) with color characteristic different from the background indicates the edge of document. Though, taking lower M values allows

for very precise determination of the document edge (also with cut corners, waved edges etc.), it is not important from the further processing point of view. The document layout of the example image (Fig. 4a) is marked as the largest light-gray rectangle. The presented algorithm is fully efficient and very fast.

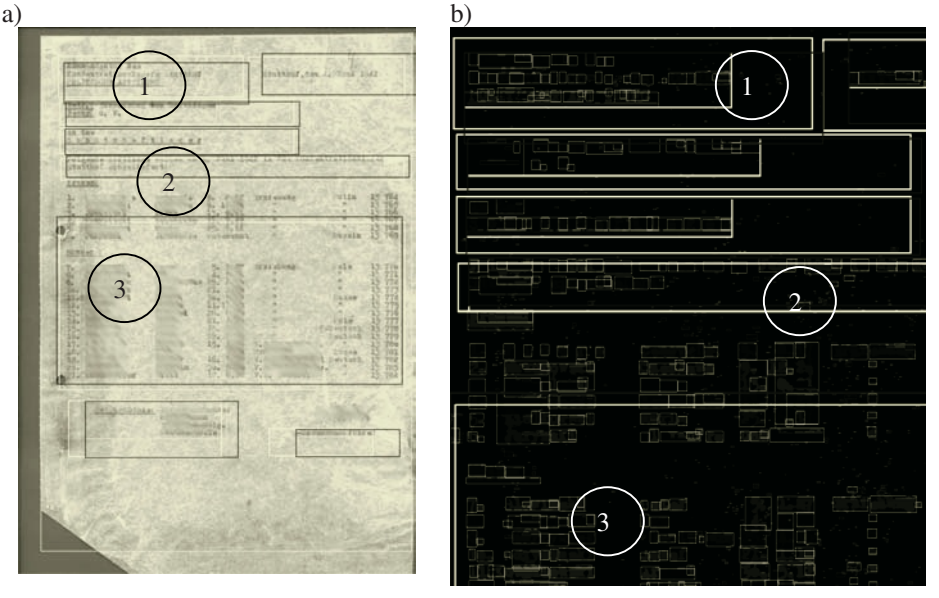


**Fig. 4.** Original, Nazi camp document with a) marked regions: template (white), moved (dark grey) and suited (light grey), b) character segments and regions.

### 3.2 Background Separation

The next step of typescript image's pre-processing is the background separation. It means removing these parts of the image, which do not contain any information that could be essential in further recognition process. The background separation leads to separation of machine-type characters whereas, graphical elements (such as signatures and handwritten notes) are treated as unwanted disturbance and should be eliminated. Unfortunately, the low quality of archive documents does not allow for using global, and even local, thresholding algorithms [5]. Moreover, interactive, *on-line* document processing does not also allow for using any advanced thresholding method with high counting complexity.

**High-Pass Filters**

High-pass filters are commonly used for edges detection particularly enabling localization of character segments. In the very first stages of research in the Memorial project, modified Sobel filters [3] (mask 3x3) were used with very good results (Fig. 5a). Unfortunately, this method failed when applied to original documents, which was probably caused by their significant blurring (Fig. 5b). Increasing mask size (5x5) had improved results of segment detection (Fig. 5c) but it had also increased filtering

time. This prevents high-pass filters from being applied for whole image but they still may be used locally to improve thresholding results in doubtful cases.

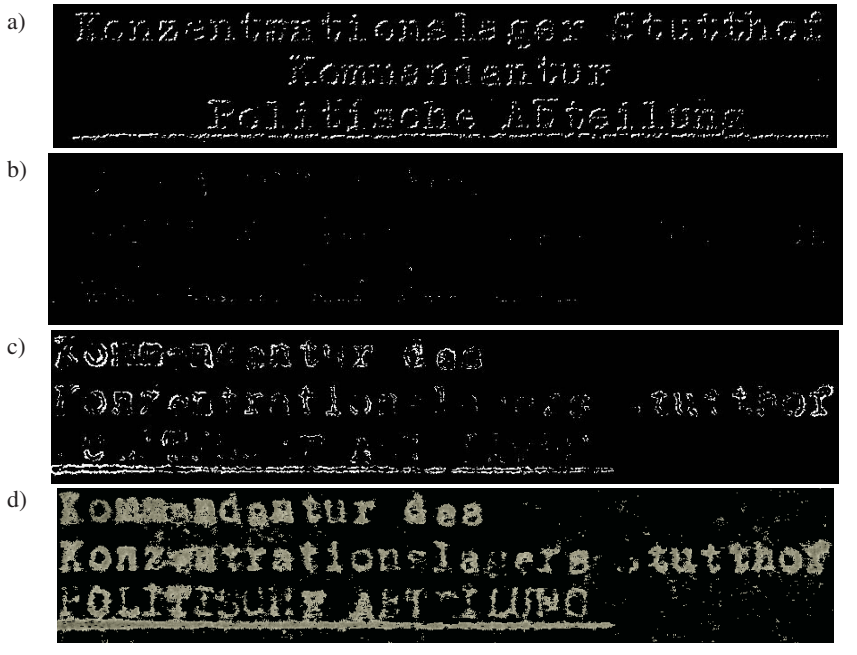


**Fig. 5.** Background separation using: Sobel filters (a – false document, b - original document (mask 3x3), c - original document (mask 5x5) and d) – algorithm for colour images.

**Thresholding of Colour Images**

When applying OCR systems for images of good quality it enough to consider monochromatic or even black-and-white images. However, in archival documents of bad quality very important information exists also in chromatic part of image. Resigning this information may cause wrong background separation and further, may lead to bad information retrieval.

Further research on this problem, results in new algorithm that bases on full colour information (RGB) contained in document image and converts it to monochromatic (and B/w as well) image. This algorithm eliminates most of background pixels in the image (Fig.5d) basing on 4 criteria that uses several thresholds $\iota_i$ calculated on the base of statistical properties of documents' images. Important fact is that choice of these thresholds is not crucial because they have safe error margin for different documents. The algorithm uses the following criteria to eliminate background pixels:

- rejecting of all pixels belonging to the grey-green background of documents (3.1). Such points may also occur inside determined document layout, mostly in the case of different document's physical damages, e.g. punch-holes, cut-corners, and torn edges, very light paper sheet etc.

- intensity $J$ thresholding using 2 safe thresholds of $\iota_{Jmin}$ (=50) and $\iota_{Jmax}$ (=180) that determines points, that certainly do not belong to characters: too dark points ($J<\iota_{Jmin}$) occurs very rarely because of ink paling and points too bright ($J>\iota_{Jmax}$) belongs to the characters' background.
- filtering all points that have at least one neighbour pixel to its right that has colour document background (green-grey). The goal of this operation is to detect large regions of document background. Information about these regions may be further used for estimating scale of document damage [2] or for locating punch-holes. Using one-dimensional filter significantly increased the operation speed though better results could be undoubtedly received using 3x3 mask.
- rejecting 'colour' points, it means all points that have a dominant red, green or blue colour component. These criterion results from the assumption that the black ink was used in typed documents and that fading was nearly equal for its all, three colour components RGB. In the criterion all points are rejected that have any colour component greater by $\iota_{chrom}$ (=30) from any other. Taking such high $\iota_{chrom}$ value, common for all colour components gives very good result providing high error margin. Additional experiments have proved that decreasing $\iota_{chrom}$ individually for each colour components could lead for even better background separation results except some specifically coloured regions.

### 3.3 Detection of Segments

Determining location of characters that belongs to text regions is the first step of document segmentation. Because the main goal of this stage is segment location (not recognition) it is possible to use methods that emphasise segments but not always preserve their shape. The algorithm worked out for the *Memorial* project consists of two steps:

- noise removal using logical filter with mask size 7x7. This filter removes all pixels that have less than 5 neighbours inside the mask window centred on them. Though filter „square" radius is rather high it filtering doesn't consume too much time because it is made only for pixels separated from the background.
- Morphological dilation of the image with rectangular (9x3) structural element that connect broken parts of the characters' segments.

Finally, the algorithm returns a list of segments that represent single or connected machine-typed characters (Fig. 4b).

### 3.4 Matching of Regions

The aim of the regions matching algorithm is to establish the real location of template regions for the processed document. This goal may be achieved by two means:

1. Finding the real location of the left upper corner of the template region preserving its original size. Such regions for example image (Fig. 4) are marked with dark-grey (blue) rectangles.

2. Finding the smallest rectangle that contains all characters of the region. Such regions for example image (Fig. 4) are marked with light-grey (yellow) rectangles.

The algorithm used to match text region founds all segments with proper size that fully lay inside the template region or intersects it having at least $\iota_{intersect}$ =40% of their size inside it. The minimal segments' dimensions depends on font size and scanning resolution and for concentration camp documents have been established as $\iota_{minsegx}$ = $\iota_{minsegy}$ = 15 pixels.

The algorithm's assumption demands that the template regions cover or intersect all characters belonging to it. The examples of such regions are marked with 1 and 2 on Fig. 4. If the above requirement is not fulfilled, the algorithm extends the region's size that it could cover all characters laying nearby, but it does not have enough information to continue this process to cover farther rows or columns. The example of such situation is marked by 3 on Fig. 4. Regions extending by subsequent rows and columns demands knowledge about the whole document structure and semantic analysis of neighbouring segments (e.g. whether they construct row or column in a table). Such semantic analysis goes beyond the algorithm's assumptions, however, it will be developed in further research under the framework of the *Memorial* project.

## 4 Summary

In this paper, the problems of pre-processing and segmentation of bad quality documents are described. Several algorithms for document localization, background separation, segmentation and regions matching are presented. All algorithms have been verified for over 50 archival documents belonging in to 4 different classes (e.g. transport lists). All documents (even within the same class) have differed in regions size and structure. The results achieved confirm very good efficiency of background separation algorithm. This confirms obvious thesis that for bad quality documents all available information should be used including chrominance. Good results were also achieved for regions' matching algorithm under the assumption that template regions cover all characters belonging to it. In other case the matching is fragmentary or missed.

High speed of presented algorithms allows for interactive on-line work with processed documents. In such case the user is able to manually correct regions boundaries. Special automatic measures could allow this task by pointing problematic areas [2].

In spite of good results achieved for presented algorithms some other high level methods for regions matching should be search. These advanced methods, using knowledge about the document structure, should deal with difficult cases of regions matching. The research in that direction, including linguistic approach [4], will be further developed in the Memorial project.

## References

1. Wiszniewski B.: The Virtual Memorial Project. http://docmaster.eti.pg.gda.pl
2. J.Lebiedź, A.Podgórski, M.Szwoch: Quality Evaluation of Computer Aided Information Retrieval from Machine Typed Paper Documents. In proceedings of Third Confrence on Recogntion Systems KOSYR'2003. Technical Uniwersity of Wroclaw, Poland, Wroclaw (2003)
3. Malina W., Ablameyko S. Pawlak W.: The foundations of digital image processing. Acad. Press. EXIT 2002 (in polish)
4. Szwoch M: Musical notation recognition using context free attribute grammars, Ph.D. thesis, ETIF Technical Uniwersity of Gdansk, Gdansk 2002 (in polish)
5. Sahoo P.K. et al: A Survey of Thresholding Techniques, CVGIP 41, (1988).