

A Device for automated Scanning of Books

Stephan Ullrich
Arved C. Hübler
Klaus Kreulich
Carsten Enge

Chemnitz University of Technology/Institute for Print and Media Technology
stephan.ullrich@mbv.tu-chemnitz.de
09107 Chemnitz
www.tu-chemnitz.de/pm

Abstract

Fast access to older theses and dissertations is still difficult, because they are often available only in printed form. If in digital form, content of these books could be available more easily. However, scanning of bounded books is a time-consuming and costly process - the pages of the book must be turned manually.

At the Institute for Print and Media Technology, a device was developed, which can turn the sides of a bounded book automatically. While the book is scanned, the operator is free for other tasks, as for example the control and revision of already scanned material and the input of corresponding data records (meta data) into a catalogue SYSTEM.

The presentation gives an overview of the technical functionality, application possibilities and limits of the device.

A second main point is the presentation of a draft for the integration of the device into a software based workflow. After the optical character recognition an examination of the scanned pages of the book should be done within this workflow. Missing pages must be recognised, and the operator is informed about these failures. In another step the structure of the book is analysed. Among other things this information can be used to produce an XML-File. Thus the publication and archiving of the thesis as an ETD in online data bases can be supported easily.

analysis of extensive works easier in many scientific disciplines.

Another relevant operational area is the production of print-products based on digitized books. On the one hand the digitized works can be printed on demand as reprints using print-on-demand-technologies. On the other hand the components of digitized books can be used for the production of generic books. The media production processes necessary for that are one of the research areas at the Institute for Print and Media Technology at present.

Surely numerous further applications can be found for digitized books. However, relative few special works were digitized in the last years. Reason for this might be the time-consuming, inefficient and expensive processes of digitizing.

In the following a SYSTEM, which automates the digitization of books as far as possible, is presented. This SYSTEM consists basically of a device that is able to turn over and to scan book pages, and a software-based workflow to process digital data.

Preface

Innumerable scientific works, like theses, magazines and monographs, are often available in printed form in libraries and archives only. The supra-regional access to these works, which are partly rare and precious, is still difficult: The document must be ordered and dispatched by a delivery service and so even the risk of damage exists.

This is different with digitized books. They can be retrieved via Internet in a few minutes worldwide. The original remains in the bookshelf and is preserved, while the digital copy can be worked with.

Apart from a higher availability of the work the digitization offers additional search possibilities and operational areas. Text recognition supplements not only the data in the library catalog around a full text search, but also forms the basis for a structural analysis, which tables of contents and indices of a work can be recognized with and worked into data bases. Over and above that text recognition allows the development and evaluation of whole corpi using data mining concepts and makes the

Scanning of Books

A flat bed scanner is hardly suitable for the preparation of digitized books: The book must be turned over page by page and layed onto the glass plate of the scanner and the scan-process must be started manually. This work is monotonous, with heavy books physical loading, and often leads to heavy damage of the bound work as well.

In the first glance the destruction of the bookbinding is a timesaving alternative, because the sheets can be inserted into a sheet-fed scanner. Due to the damage of the book this method is hardly not acceptable - particularly for older, valuable works.

The so-called book or table scanner represents a solution to scan bound works carefully. Contrary to a flat bed scanner the book lies opened on a table or a so-called book-cradle. A scanner or a digital camera is attached above. The operator turns the pages over and releases the scanning-processes after each turn. This procedure is repeated until the last page of the book is scanned. The

operator has to control and to catalogue the scanned files after turning the complete book page by page.

The main disadvantage of these book scanners is that the operator is continuously busy during the digitization. This work is very monotonous, nevertheless requires highest concentration to scan the book without failures. An increase of the effectiveness can be obtained, however the efficiency can be mainly increased by automatic support for turning of the pages. Such a device - called Bookreader - is developed at the Institute for Print and Media Technology.

The Bookreader - A Device for automated Scanning of Books

The Bookreader consists of several components. The core component is a newly developed device, which turns the pages of a book automatically. A book scanner is attached above this device, which is connected to the turning equipment and on the other hand to a computer, running the software for scanning and processing the digital data. The scanner and the appropriate software can be exchanged depending on intended purpose. For example, a fast gray tone scanner is often sufficient for scientific works, whereas a slower color scanner can be used for full-color books.

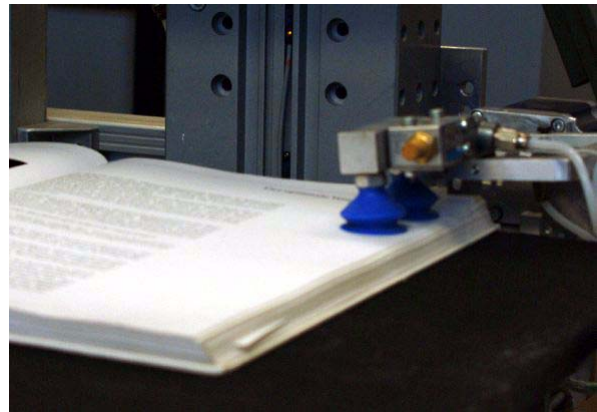
Sequence of functions

To be scanned automatically by the Bookreader, a book must be put on the book-cradle and fixed by clips. The book-cradle consists of two plates, which are both connected and installed on an own pneumatic cylinder. With this construction the different height of the two book halves of the opened book becomes balanced when a glass plate is lowered from above the book before the scanning process. So the top pages are on the same level after the divided glass plate from above is lowered. This adjustment of the two book halves on an equal level ensures a continuous quality and sharpness during the scanning. After the scanning of the actual top pages the glass plate will be raised again and the right page can be turned.

Turning a book-page over takes place in several steps:

1. Taking the current highest (right) page of the opened book.
2. Separating this page from the underlying stock.
3. Turning of the page.

Taking the page to turn and the separation from the pages underneath works similar to separation of sheets in sheet-fed presses. In practice, for sheet feeders a combination of blowing nozzles and suction cups became generally accepted.



At the Bookreader air is blown into the upper right edge of the opened book by a nozzle. This loosens the pages up and thus supports the release of the upper book page. Two suction cups touch down on the toggling page and fix it by negative pressure.

Due to the different air permeability of different papers it can occur that the suction cups take more than just one page. For this reason the bar with the suction cups is tilted around a horizontal axis. Because of the flexural rigidity of paper a possibly still adhering second page drops.



While the page taken is raised, it is bulged. Injecting additional air supports this curve of the paper. Thus sufficient space is formed to move a lever under and to turn the page.



Lowering the divided glass plate, which moves down angled into the inner margin of the book, prevents falling back of the turned page. When lowering further the two parts of the plate in one plane the possibly still curved upper pages of the book are smoothed. In interaction with the book-cradle the differences in height are levelled between both parts of the book. After that, a signal is sent to the scanner and the scan-process can start. When the scan is finished data are processed in the connected computer and the turning process starts again automatically. The entire procedure is repeated until the last page of the book is captured.

Technical data

To specify technical data of the prototype over 3000 different books from different departments were measured in the library of the Chemnitz University for Technology. Extremely large and extremely small formats were not considered in this examination.

With the present solution, pages of standard books with a size between 10x15cm (about 4"x6") and 30x45cm (about 12"x18") can be turned automatically with sufficient high reliability. That covers 96% of the examined books.

The speed of the SYSTEM depends on two components: on the one hand of the speed of the used scanner and on the other hand of the speed of the page-turning device. Without scanning turning one sheet takes about 20 seconds, so maximum 360 pages can be turned over per hour.

By further improving of the device, both the speed and the variability of the formats can be increased.

Software workflow

After scanning of one or several books the digitized pages are available as TIF-files. Depending on the type of the

original and the intended purpose different workflows can be formed for their processing. Thus e.g. a digital magazine requires more complex text recognition and processing compared to a simple reprint.

At the Institute for Print and Media Technology a simple workflow was first realized where standardized interfaces of commercial products are used. PDF-files are eventually produced allowing full text search and the access to contents of the work by navigation elements and references.

The TIF-files can be transferred into different file formats. The 100% accuracy is not yet given using state-of-the-art OCR-software. Especially when reading scientific work the clarity of formulas and technical terms and the quoting ability are very important. Therefore manual control of the OCR result is necessary for the conversion into text-based file formats, e.g. RTF, HTML or XML. One alternative, which can be provided fast and automatically, is the non text-oriented PDF-format, in which the OCR result can be linked subsequently with an image of the original in one file. Moreover, the PDF-format is a widespread standard format for the document exchange in the print and media field and fulfills the requirements of both conventional and digital printing processes. Later converting of contents after manual control of the OCR result into other, textbased file formats, e.g. RTF, HTML or also XML is possible.

The Software "Adobe Acrobat Capture" is used for text recognition. Apart from the transformation and collection of the individual TIF files into one PDF-file, this software can also recognize the logical page numbers and the structure and/or the table of contents of the scanned book and makes them available as references and navigation elements in the form of Bookmarks.

This workflow will be extended by the following components:

- Control of the digitized book: Possible failures in turning over the pages can be recognized by control of the succession of the logical page numbers. The operator will be informed about missing pages so they can be supplemented by subsequent scanning. Double pages with same page number and identical content are recognized and removed from the digital book.
- Storage of the data in a content management SYSTEM: "Acrobat Capture" can communicate with a content management SYSTEM via the ODMA-interface¹ and save the TIF-files and the PDF-documents in this. Already within "Acrobat Capture" the possibility of entering metadata is given for the description of the digital objects, which will be transferred when storing the files into the content management SYSTEM. The content management SYSTEM is based on a client-server architecture. Different software for the

¹ The Open Document Management API (ODMA) enables the integration of standard desktop applications in document management SYSTEMs. (<http://www.infonuovo.com/odma/>)

treatment of the digital objects can be integrated to the client.

- Integration of the Metadata Encoding & Transmission Standard (METS)² into the content management SYSTEM: The METS format is a kind of a container, in which apart from the content-related description of a work by descriptive metadata administrative metadata for the description of technical and legal data of an object and structural metadata for the information about relations of several objects among themselves can be stored. Thus METS supports not only the search but also the administration of digital objects. For the administration of descriptive metadata several well-known standards can be used within METS, e.g. Dublin core and MARCXML. This allows the import of the metadata of the original from a library catalog and in addition the conversion into other metadata formats.
- Integration of export-modules into the content management SYSTEM: Different clients have different requirements to a digitized book, also concerning the file format. Export modules will ensure that the requirements of such clients can be fulfilled.

Summary

In this paper, a SYSTEM is described that allows the decrease of manual work, increase of efficiency and quality

in the process of digitizing books and other bound printed matter.

The Bookreader automates the work of the page turning and allows the operator to run other devices or to handle more sophisticated tasks. Thus more books can be digitized efficiently with the same personnel effort. Further improvements of the SYSTEM include the increase of the operating speed and the integration of a magazine so that a book can be replaced by another book automatically after scanning.

Due to the modular concept of the SYSTEM it can be adapted and optimized to meet different requirements. This applies also to the software-workflow, that process graphic data and is open to further development and demands.

Acknowledgements

This work is supported by DFG (Deutsche Forschungsgemeinschaft) within the project "Decentral Digital Research Library" ("Verteilte Digitale Forschungsbibliothek").

2 For the description of objects in digital libraries METS is being developed by the Digital Library Federation. (<http://www.loc.gov/standards/mets/>)