

From Book to Text File

Realisation of a Book Scanner, and Programming of an OCR Software

by Christian Graber and Elias Mügler

coached by Guido Lang and Tobias Witschi

Kantonsschule Kreuzlingen, 2006

1 Introduction

The basic idea of our project is the transfer of a complete book to a computer. The entire transfer process shall be carried out without human intervention; i.e., the turning of the individual pages of a book, and the scanning of the pages, shall be performed automatically. Another part of the project is the programming of an OCR software, which is able to filter the text from the scanned images.

The idea of this project originates from an optional course at our school. In that course, we were shortly informed about the principles of OCR. We wanted to program such a software as a school-leaving final work. In our German classes, we were also made aware of the Gutenberg project: Texts without copyright protection are offered on that web site. Investigations showed that the volunteers in that project put the books onto their scanner by hand, page per page. This gave us the idea to automatise just this process.

We were given ten months' time for our project. Experts doubted that it could be realised in such a short time. Still, we did not allow ourselves to be discouraged, and tackled this task.

2 The design

2.1 *The setting of tasks*

The objective of the design work was the development and realisation of an equipment which is able to scan the individual pages of a book, and to turn them over automatically. Due to the short time available for the realisation of this project, we had to introduce certain limitations: A ring binder in the A5 format was used.

2.2 *The basic idea of the page-turning system*

The first question which has to be answered is: How can a page be turned over? Two basic systems can be considered here. In the first system, a page is sucked up, then lifted. A stick slides under the lifted page and turns it over. In the second system, a small roller pushes the page towards the middle of the book. A gripper seizes the page, lifts it, and turns it over.

The suction system is more efficient, because its design is simpler, and there is no risk of damaging the paper in the process. The book page might get damaged if a roller were used. Furthermore, it is difficult to grip the lifted page; it might be more than one page. Therefore, we decided to continue our work with the suction system.

The second question is: How is the page suctioned? One possibility is the use of a vacuum generator and suction caps. Another possibility is the use of flat discs that are lying on the page. Air streams through a small opening in the centre of the discs. According to Venturi and Bernoulli, a vacuum is created by the high velocity of the air stream. We may assume that a single page only is lifted at a time. In the case of suction caps, however, there is the risk that more than one page is lifted at a time, or none at all. Furthermore, the page does not adhere directly to the discs during the blowing process: There is always a thin layer of air in-between. The paper may get damaged in places if suction caps are used. In a comparison, and in experiments, the second solution rates considerably better.

The next decision refers to the drive system of the equipment. Here, too, a decision between two possibilities has to be taken: Electric motors can be used in a multitude of ways, and they can be controlled very precisely. The second possibility is a pneumatic system. Since almost all movements in the system go from one stop to another, the application of pneumatics is more adequate.

2.3 *The design of a model*

Once the movements had been defined, we could start the design work. We used a CAD software and realised the design in three dimensions, and to scale.

2.3.1 The first model

The page-turning system, which is presented in chapter 2.2, can be seen in the first model (fig. 1). A scanner is positioned at the top of the frame. This frame is lowered along the two lateral guides onto the book. The book is lying on the base plate and is positioned against an abutment. When the scanner is in the upper position again, the rectangular arm sucks the page and lifts it. The stick on both horizontal rails turns the lifted page over. The arm can turn outward by means of a second joint: Now there is room again for the scanner.

When we analysed and evaluated this system, we soon found defects. The page-turning system which we had chosen cannot be implemented in this model. We see this in fig. 2: The arm cannot turn outward because the guides for the scanner are in the way. Furthermore, the scanner cannot go all the way to the book, because the stick is in the way. Therefore, we had to abandon this first model.

2.3.2 The second model

The second model (fig. 3) shows the cylinders which we used for the machine. Now the arm no longer turns outward, but makes a linear motion, driven by a band cylinder. Generally, more attention was paid to the space requirements. Because of the cylinders, this model has two plates, one above the other. The valves and the control unit could be accommodated in-between.

2.3.3 The third and final model

The third model (fig. 4 and 5) was created on the idea to turn both cylinders around so that the whole structure is reduced to one single plane. This makes it more stable. Furthermore, this model has two suspended plates on which the book is lying. The book can give way when the scanner is pressing on it; therefore, both pages are at the same height during the scanning process. The third model eliminated all shortcomings of the previous models, and was ready for production.

2.4 Structure and production

In a first step, we arranged all elements on a board. This was the best method to find the optimal routing of the cables and the best placing of the control unit. Subsequently, we bolted the definitive installation onto an acrylic glass plate (definitive installed machine in fig. 6). We went to the workshop ourselves and endeavoured to do as much as possible ourselves. The suspension system of the plates caused us the most trouble during the manufacturing process, because of jamming. Therefore, we had to enlarge the diameter of the guide bores. Furthermore, we screwed another plate with exact bores to the bottom of the suspension system, so that the springs are running straight.

2.5 The results

The machine is working as planned (e.g., fig. 7). Our equipment fulfils the goal. It is able to leaf through a ring binder correctly and without damaging, and to scan the individual pages (procedure in fig. 8a-k). The turning of a page takes 15 seconds. Scanning takes between 20 and 60 seconds, according to the resolution which is required for the OCR process.

2.6 Discussion

We are very pleased with the results. The turning-over procedure works impeccably and quickly. When a book is being digitised, scanning is the procedure that takes the most time. Alternatively, a digital camera with a high resolution could be used instead of a scanner. Thus, the mechanical system could be simplified, and the cycle could be accel-

erated. However, this is a question of costs. The system could now be developed further so that all kinds of books could be digitised. To this end, a system would have to be developed which retains those book pages that have a tendency to fold back.

3 The control system

3.1 *The setting of tasks*

This part of the work deals with the controlling and monitoring of the machine. Since the machine is driven by pneumatics, valves are used to control the air streams. Hall probes at the cylinders act as sensors, the signals of which are processed and have to be synchronised by the computer.

3.2 *System selection*

During the process of realisation, two different systems were considered: A self-made control unit, and a programmable logic control (PLC) system. A self-made control unit would be very flexible and could be adapted very precisely to the machine. The effort required for its development, however, is considerable. Since a PLC system also fulfils these requirements and is safer and easier to handle, we opted for a PLC system.

3.3 *System 1: Logo and IO-Warrior*

At the beginning, we worked with a Siemens LOGO! programmable logic control system (fig. 9). This PLC system does not have an integrated facility for the communication with the computer; the connection cable only serves for the downloading of programs. Therefore, the data exchange was realised with an USB-I/O card (fig. 10): An output port of the PLC system was connected to an input port of this card, and, conversely, an output port of the card was connected to an input port of the PLC system. This enabled the transmission of handshake signals. The computer can inform the PLC system when a scanning procedure is completed and a turning-over procedure may be started. After the turning-over procedure, the computer receives a signal and may start the scanning procedure again.

Since programming the LOGO! works with "drag & drop", the programming of longer programs, and, in particular, of sequences, is very time-consuming and confusing. For this reason, we changed the PLC system.

3.4 System #2: S7-200

The Siemens S7-200 (fig. 11) can be programmed with three different programming languages, among others, with instruction sets, with which procedures (chains of steps) may be programmed easily and clearly. Moreover, the S7-200 has an RS-485 interface and, therefore, is able to communicate directly with the computer. Therefore, it is now possible to exchange intermediate results, further to handshakes.

3.5 Development

Since we had never before worked with a PLC system, we first arranged small test circuits. When these worked perfectly, we designed the overall program, implemented it, and simulated it. The PLC system sends status information to the computer during the turning-over procedure.

3.6 Results

The PLC system controls the operating sequences of the machine impeccably, and the communication with the computer is without any problem.

3.7 Discussion

We are very pleased with the control unit, since it fulfils its task correctly. Moreover, we have chosen a system that we can expand without much effort.

4 OCR

4.1 The setting of tasks

The purpose of OCR is the filtering of text from a book page which is available as a graphical metafile. The software should be able to control a scanner, to scan a double page, and then to perform OCR.

4.2 Motivation

Although several cheap OCR programs are already available in the market, it appealed to us to program this software by ourselves. Originally, the project stemmed from this idea.

4.3 The operating principle

The program is given the task of filtering the whole text from a page. The idea is to subdivide this major problem into a multitude of minor problems. To this end, the page is subdivided into individual characters, which are then recognised.

To subdivide a page into characters, it is first subdivided into lines of text. To this end, the number of black pixels is counted along every horizontal line of pixels (fig. 12). If this number is zero, an empty space between two lines of text is identified. This separates the individual lines of text. Now the same procedure is applied to every vertical line of pixels within a horizontal line of text (fig. 13). If the amplitude along such a vertical line of pixels is zero, an empty space between two characters is identified. If the amplitude remains zero over several rows of pixels, an interval between two words is identified.

Now we can look at all characters individually. Each character is compared to a reference alphabet, and the character with the least deviation is displayed every time.

4.4 Testing of words

For fonts that comprise characters that are almost identical (e.g., small L and capital I, fig. 14), the above-mentioned method fails. To counteract this, the character with the second-least deviation is also displayed if such deviation from the optimal character is minor. In the case of the word "Island", the software is uncertain about the first and third character; therefore, it finds several possibilities for these positions: "[Ii]s[Ii]and". Now all possible words are formed: lsland, lsIand, Island, IsIand. These words are now searched for in a word database, and since "Island" is the only word that is found there, this is the word that will be displayed.

4.5 Development

The software was developed with Borland Delphi. The DelphiTwain feature was used to control the scanner. Text files were used as databases since they are easy to manage. Bitmaps were used to enable an easy analysis of the images with Delphi.

4.6 Results

95% of the text is recognised correctly. With the additional testing of words, up to 99% is recognised correctly.

4.7 Discussion

The software yields good results at the recognition of simple fonts. For books that also comprise images, the software would have to be revised: The pages would have to be analysed prior to the recognition of text. Since the scanned image data are stored intermediately, they can be transferred to some other software at any time without problems.

5 Conclusion

The machine, the control unit and the software fulfil their task. The system is able to scan and recognise a complete book. Despite the fact that there is still a very high potential for further development in our project, we are very pleased with it. We are proud to have mastered this problem with success. We will have a look at any possibility of further development and implement it, if feasible.

6 Personal impressions

6.1 Christian Graber

The teamwork in this project fascinated me. We were given the opportunity to delve into new fields of knowledge, and we were always supported by specialists. I am pleased at the interest the industry showed in our project. This work gave me some insight into the world of mechanical engineering. It was a great chance for me.

6.2 *Elias Müggler*

This project gave me an insight into many interesting specialised fields of knowledge. I learned the application of pneumatics, the development of SPC systems, how to manufacture in a workshop, and how to tackle large software projects. I took great pleasure in forming new contacts with industry, and with other junior researchers.

7 Thank you

We would like to thank all persons and firms that gave us good advice and support. In particular, we would like to thank Harry Lüthi, Roger Hess, Tobias Witschi, Ronny Balmer, Rolf Laager, Hans Menzi, Guido Lang, as well as MOWAG, SMC Pneumatik, and Siemens. Many thanks indeed!