TEXT RECOGNITION

Kazylina Y.I.
Scientific supervisor: lepustin a.v., senior teacher
Tomsk Polytechnic University, 634050, Russia, Tomsk, Lenin Avenue, 30
E-mail: kyai@tpu.ru

Optical character recognition, or OCR, is a mechanical or electronic process of a handwritten, typewritten or printed text conversion into text data, i.e. code sequence representing computer symbols, for example, in a word processor. OCR is widely used for converting books and documents into an electronic format, business accounting system computerization and text publication in the Web.

Text recognition on the PC

The process of digitization and OCR includes five steps.

- Page data input. On this stage scanned or photographed document is transferred into a computer image.
- Layout analysis. OCR application detects the arrangement of text, pictures, tables etc. on the page and divide it into blocks. The software sequentially splits the page into smaller blocks: the text into paragraphs, then into sentences, separate words and symbols. At the end layout analysis the page is represented by a set of separate symbols. The application remembers the location of every one of them on the page.
- Character recognition is the most important step of the OCR process, because the software must correctly identify all found symbols. Is it a letter 'B' used in the text (and which one of them – Cyrillic or Latin) or is it a figure '8'? If the application fails, the result will turn to nonsense. For more accurate text recognition the software combines different methods, that can conventionally be divided into two categories: pattern-comparison methods and feature-comparison methods.

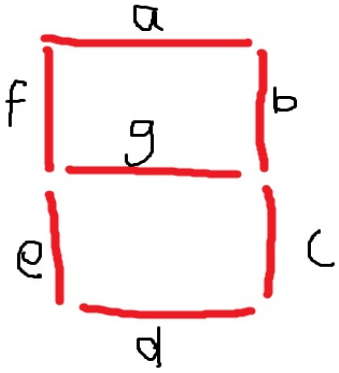

Fig1. Method of recognition

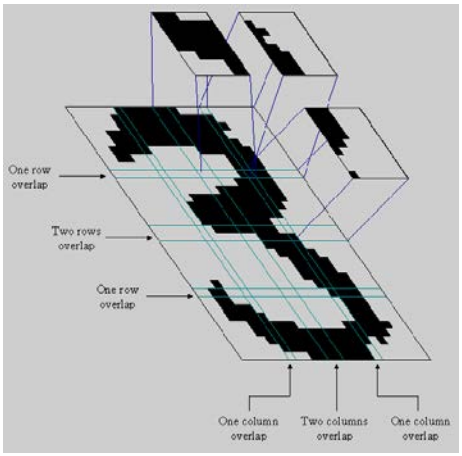

Fig 2. Method of recognition

- Document reconstruction. After the recognition process is complete the software starts to recreate pages, combining separate symbols into words, words into sentences, sentences into paragraphs etc., using the embedded vocabulary. To quicken the process the results of layout analysis (step 2) are used. Moreover, using special methods applications try to take into account grammatical features of the text so the result would represent grammatically correct sentences.

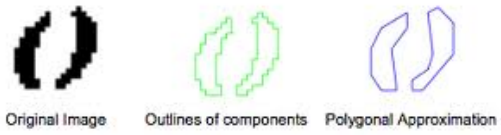

Fig 3. Method of recognition

- Document saving. OCR application saves the recognized text in the user defined format (text only - TXT; page layout – Microsoft Word files of PDF

Text recognition software

The most widely known text recognition software are ABBYY FineReader, CuneiForm, OmniPage и Readiris. They are available in different versions, for home and professional use.

ABBYY Finereader 10.0.102.95 is a popular text recognition application by the Russian company ABBYY. Finereader provides high-quality recognition and document format keeping. There are three versions of this package: Home Edition, Professional Edition and Corporate Edition. The software recognizes text in more than 180 languages, with the built-in spell checking for 38 of them.

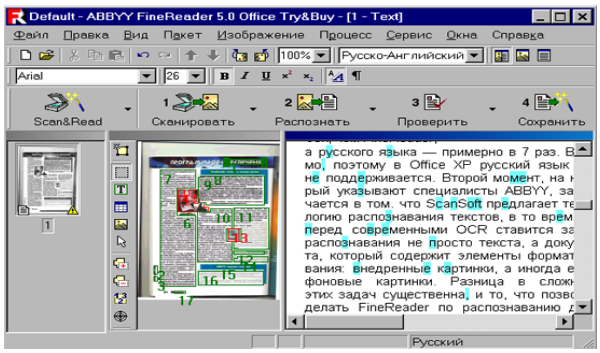

Fig 4. Main window programm

OCR CuneiForm is a free software for scanning and text recognition by the Russian developer Cognitive Technologies. Initially OCRCuneiForm was being developed as a commercial product, but in the December 2007 the developer company started to distribute the software for free, and in the April 2008 opened the source code. The future development of the text recognition system is planned – OpenOcr.Org project, supported by the CognitiveTechnologies company and OpenSource developers community. OCRCuneiForm provides fast, easy and high-quality text recognition with document format keeping. The number of supported for recognition languages is more than 20, among them are Russian, Ukrainian, English, German, French, Spanish, Italian, Portuguese, Swedish, Finnish, Serbian, Croatian, Polish; and also there is a recognition of mixed English-Russian text.

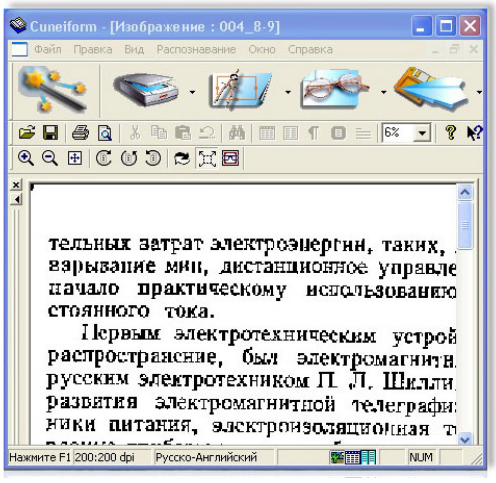

Fig 5. Main window programm

OmniPage software features a high speed and accuracy of the recognition. There are over 120 languages for recognition with different alphabets: Latin, Greek, Cyrillic, Chinese, Japanese and Korean languages. As well as FineReader, OmniPage recognizes documents, made with digital cameras, with the help of the image correction technology 3DCorrection. As well as other OCR applications, Readiris convert scanned document images in an editable format. Readiris recognizes documents with a complex layout, tables and pictures. There are Pro and Corporate versions and additional modules for Middle Eastern and Eastern languages recognition. The number of supported for recognition languages is more than 120, including Russian and Middle Eastern languages.

SimpleOCR is a free OCR system for recognition of texts from scanners and pictures. The result quality is comparable with many commercial analogs. The recognition accuracy can reach 99% - a very high performance for such systems. This version of SimpleOCR works with documents in English and French only, but vocabularies for other languages will appear in the future.

Text recognition libraries

Today there are a lot of text recognition libraries, but most of them are commercial, and we will need free ones, so the list of libraries is significantly reduced. The list of libraries we've examined: Tesseract, Pumanet, AForge, OpenCV.

- **Tesseract-Ocr in Visual Studio** is a free text recognition library. To include it, all necessary components have to be downloaded.
- **AForge.NET Framework** is a base of C# and is oriented for researchers and developers in computer vision and artificial intelligence – image processing, neural networks, genetic algorithms, machine learning, robotics etc.
- **Puma.NET** is a shell for Cognitive Technologies CuneiFrom recognition library which allows to add recognition functions in any .NET Framework 2.0 (or higher) application. API is provided in a set of simple classes. High performance and recognition results accuracy can be reached with just a few lines of code.
- **OpenCV** (Open Computer Vision) is an open-source computer vision library, providing a set of data types and numerical algorithms for image processing with computer vision algorithms. Implemented is C/C++.

References:

1. **http://www.mathworks.com/matlabcentral/answers/47668-optical-character-recognition-for-seven-segment-display-digit-matlab**
2. **http://pumanet.codeplex.com/**
3. **http://www.abbyy.ru/finereader**