

# A software tool for searching in binary text images

Nikolay Kirov Kirov  
Computer Science Department, NBU and  
Institute of Mathematics and Informatics, BAS

11 май 2008 г.

**Abstract:** In this paper we present a software tool for searching word images in scanned text documents. We consider that the document pages are represented as files in tif, jpg, gif, png, bmp and other graphic file formats. Our experiments prove the efficiency of the proposed approach and show that such type of searching can be successful. Examples of using various languages are presented. Our software is user oriented and can be applied to any collection of scanned documents.

## 1 Introduction

Optical character recognition (OCR) is the usual way of conducting text retrieval from scanned document images. It converts text images into a text file, recognizing every letter and mapping it to a number, which is called code. The most often used codes are ASCII (one byte code) or UTF-8 (two bytes code). This technique is well developed and has high accuracy.

The main problems with OCR are:

- the quality of page images
- language dependency (alphabet and coding, unknown language)
  - dictionaries
  - old grammar, obsolete words and phrases and idioms
  - old letters, outside of the coding tables
  - multi-lingual documents
- errors in automatic OCR, human intervention needed

Searching in a text file is equivalent to finding a substring in a string, which is well-known task and there are efficient algorithms for solving it. The solution can be exact – the pattern string coincides with the result, or can be approximate when the goal is to avoid some grammar changes of the searching word. Of course the last is language dependant.

---

This research has been partially supported by a Marie Curie Fellowship of the European Community program “Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria” under contract number MTKD-CT-2004-509754.

## 2 The software

Our software system for the retrieval of word images consists of a number of components. The following is a brief overview of the most important parts of the system and the necessary processing steps.

As an input data our system uses a collection of files representing a text document, each file is an image of one page of the document. Many graphic formats are acceptable as TIF, JPG, PGN, etc. The current directory containing document image files and the current file name are given on the top of main window. It is possible to do through the pages.

### 2.1 Segmentation

Lines determination is a relatively easy step in processing documents. We use horizontal projection for line extraction. If the lines are horizontal (straight lines), the histogram has near zero values between lines. The same is when the lines have small slopes.

To segment words or characters of a line, we use vertical projection. – a histogram obtained by counting the number of black pixels in each vertical scan at a given horizontal position. If the characters are well separated, the histogram should have zero values between characters. While the distances between words are larger than between characters, it is easier to separate words than characters.

For segmentation step we use a number of parameters, which are important for successful word separation (see Fig. 2):

- Minimal row height: The height of every row must be at least the value of this parameter. This help us to avoid creating (due to noise) rows with small height;
- Margin: The system reduces the page dimension from all sides by the value of this parameter and allow us to process only a part of the page. Also often page images have black lines or fields on near the ends.
- Row white: When the value at a point in row histogram is less than the value of this parameter, we suppose that this point belongs to a white space between the words.
- Row space: The white space between words must be greater than the value of this parameter. This help us to separate word images from some special symbols as dots, commas, etc.
- Minimum word length: The system does not segment words with length less than the value of this parameter. Usually it equal to the length of two or three letter words.
- Shrink white: This parameter concerns additional step when we have already separated words, and word images are framed. At this step we try to shrink the frame rectangles from top and bottom. We use horizontal and vertical histograms only for the points in a given word image. Starting from the top, bottom, left and right we decrease the rectangle size if the points of histograms have values less than this parameter. This step is very useful when the rows have small slopes (see Fig. ??).

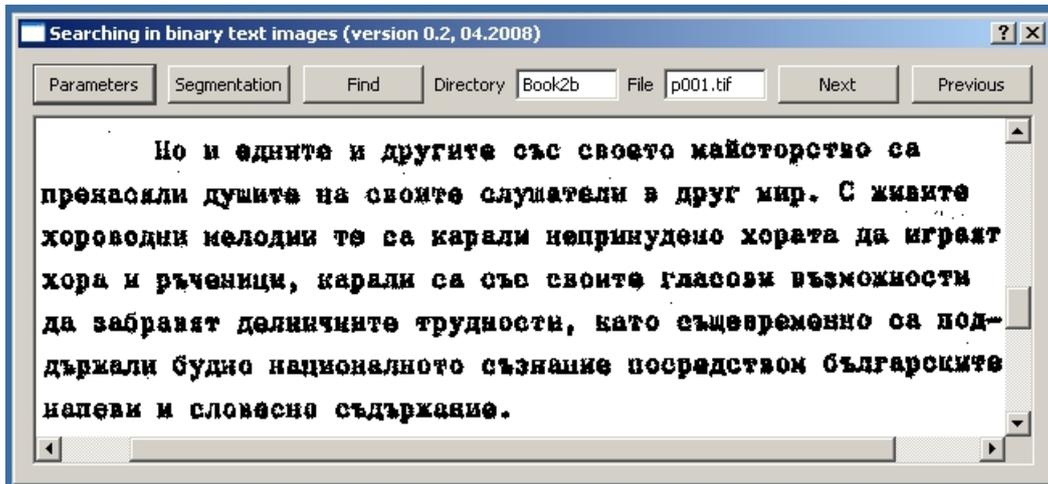


Figure 1: Main window

## 2.2 Retrieval

Before calculating the corresponding Hausdorff distance between the pattern word and a word under investigation, we must place ... we must dispose the word images ... such that the sets of black pixels coincide ... There are three options for defining a translation vector ...

Dubuisson and Jain [7] introduced so-called Modified Hausdorff Distance (MHD), one of the best measures for words similarities (see [4] for parallel with MHD and other Hausdorff type measures). A bit better results were obtained in our examples changing lightly their definition and called this modification Sum Hausdorff Distance (SHD), [2].

We can see a part of the retrieval data in Find window. Pushing GoTo button the page with

General views of user screens are presented of Figs 1, 2 and 3.

The code is written in C++ with help of Qt – a cross-platform application development framework [8].

## 3 Experiments

We present a number of experiments using the following text documents:

- Bulgarian (Fig. 4), Bulgarian typewritten document (about 1940), 335 pages, tif (2400×3200), 1 BitsPerPixel
- Bulgarian book (5), Christomatia (1884), 1000 pages, tif (2300×3800), 8 BitsPerPixel
- Old Greek, (Fig. 6), Old Greek text (approximately in the third century BC), 50 pages, jpg (1077×1416), 8 BitsPerPixel
- Hindi (Fig. 7), Hindi book (1858), 178 pages, tiff (2800×5000), 1 BitsPerPixel
- Old Spanish (Fig. 8), Text in Spain (1901) [10], 30 + 57 pages, gif (1400×2500), 4 BitsPerPixel
- Handwritten Russian (Fig. 9), Handwritten document in Russian (1840) [12], 44 pages, jpg (700×900), 24 BitsPerPixel

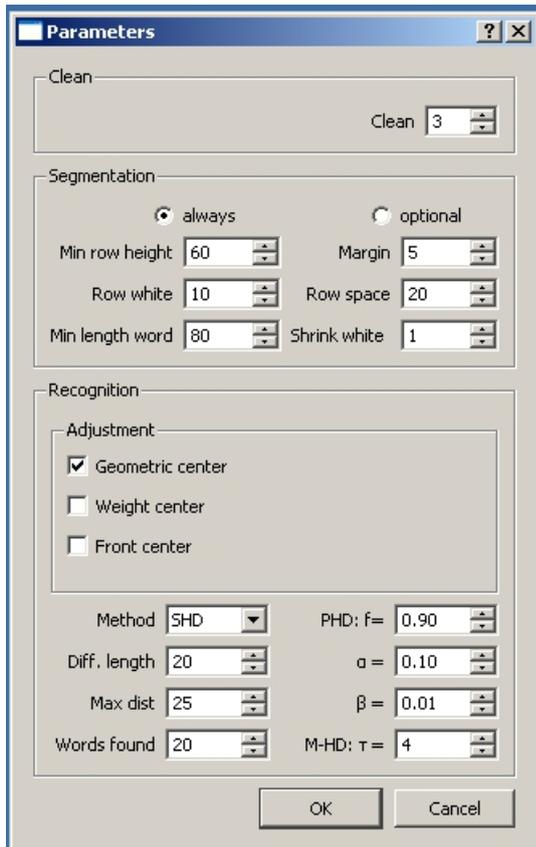


Figure 2: Parameters window

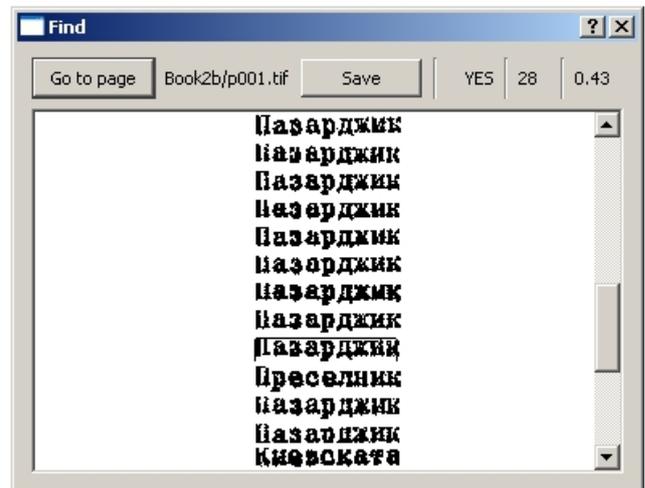


Figure 3: Found window

- French (Fig. 10), Text in French (1692), [11], 388 pages, jpg (2048×3550), 8 BitsPerPixel
- Slavonic manuscript (Fig. 11), Zbornik “Zlatoust” (1574) [9], 747 pages, jpg (1249×1890), 24 BitsPerPixel.

Накар и рядко, в кръжката си е свирил и на мандолина пред близки негови клиенти, които много обичали да го слушат. Треволирането на дясната му ръка е било ненадминато от никой мандолинист в града.

Figure 4: Bulgarian typewritten document

Повече-то от равни-тъ му стихотворения са любовни пѣсни, но подражане на грѣцки-тъ, и не прѣдставляват литературна стойност; стихотворения-та му въ „Смѣсна Книга“ при всичко, че повечето са слаби подражания на руски-тъ, нѣ свидѣлствуват вече за поетическо-то дарование на г. Славейкова; най-добри-тъ му стихотворения са обнародвани-тъ по-послѣ въ „Читалиште“, отъ които „Не пѣй ми се,“ „Жестокост-та ми се сломи“ и „Тогасъ понѣ“

Figure 5: Bulgarian book

The results of searching words are presented on Figs 12-19. The pattern word is pointed by frame in the text. The file name of

## 4 Conclusion

In this article we do not discuss the quality of image preprocessing particularly the important step of segmentation. Also we have no data of number of searching words in the text, because this is tedious work which cannot be done by computer. It follows that we cannot produce the standard recall/precision retrieval estimation (see [2]). In addition, we cannot catch the words which are incorrect segmented as well as these which are break at the end of a line







- [5] T. Konidakis, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, S. J. Perantonis, *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback*, International Journal of Document Analysis and Recognition, 9, (2007) 167Ц177.
- [6] Hwa-Jeong Son, Soo-Hyung Kim, Ji-Soo Kim, *Text image matching without language model using a Hausdorff distance*, to appear in: Information Processing and Management, (2008).
- [7] M.-P. Dubuisson, A. Jain, *A Modified Hausdorff Distance for Object Matching*, In: Proc. 12th Int. Conf. Pattern Recognition, Jerusalem, Israel, 1994, pp. 566-568.
- [8] Trolltech: <http://trolltech.com/>
- [9] Дигитална Народна библиотека Србије, Кирилски рукописи, Збирка словенских рукописа Јернеја Копитара, Зборник “Златоуст” [Digital National library of Serbia, Cyrillic manuscripts, Jernej Kopitar’s collection of slavic manuscripts, Zbornik “Zlatoust”] <http://www.digital.nbs.bg.ac.yu/>
- [10] Alonso, Rogelio M., Cartilla histyrico-descriptiva del tѳrmino municipal de Macuriges. Habana: Impr. La Propagandista, 1901, HOLLIS Catalog, Harvard University, <http://lms01.harvard.edu>
- [11] Nicolas de Bonnefons, Ch. de Sergy, (1692), University of Gent, Digitized by Google (2007) <http://books.google.com/books?id=uxg0AAAAQAAJ&hl=bg>
- [12] Дом живоначалной Троицы, Свято-Троицкая Сергиева Лавра, Собрание славянских рукописей, 43: Житие схимонаха Феодора <http://www.stsl.ru/manuscripts/book.php?col=2&manuscript=043>
- [13] Harvard Repository, Special collection,

Address: Nikolay Kirov Kirov  
 Institute of Mathematics and Informatics,  
 Bulgarian Academy of Sciences,  
 "Acad. G.Bonchv"str., Block 8  
 1113 Sofia, BULGARIA

E-mail: [nkirov@math.bas.bg](mailto:nkirov@math.bas.bg), [nkirov@nbu.bg](mailto:nkirov@nbu.bg), [nkkirov@gmail.com](mailto:nkkirov@gmail.com)  
 WEB: <http://www.math.bas.bg/~nkirov>