Преглед НЦД 13 (2008), 9–16

Nikolay Kirov
(Institute of Mathematics and Informatics
Bulgarian Academy of Sciences)

A SOFTWARE TOOL FOR SEARCHING IN BINARY TEXT IMAGES

Abstract: 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 search could be successful. Examples of using various languages are presented. Our software is user oriented and can be applied to any collection of scanned text 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 symbol 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. Searching words in a text file is a relatively easy task.

But sometimes OCR is a very difficult process requiring dictionaries in the corresponding language. Often human efforts are needed to correct OCR errors which is quite tedious work. Here are some obstacles to successful OCR:

- The quality of page images.
- Language dependence (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.

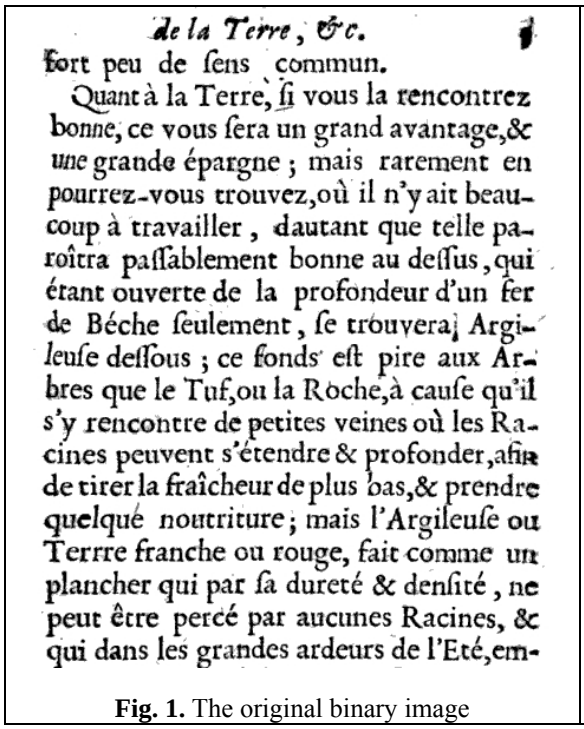
An example of unsuccessful automatic OCR can be seen on Figs 1 and 2 (see [11]).

One of the base reasons for converting binary text images to text file is the search. Searching in a text file is a well-known task - finding a substring in a string, 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 searched word. Of course the latter is language dependant.

We suggest a different approach: instead of applying two steps - OCR and searching in text documents, we will directly search words in scanned text documents. We can organize retrieval of words, similar to a given pattern word, searching in the binary text images. Similar ideas can be found in [5] and [6].

Here we present a software tool for searching word images in scanned text documents. The document pages can be represented as binary images in any graphic file format. 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.



de la Terre, &c.

fort peu de ſens commun.

Quant à la Terre, ſi vous la rencontrez bonne, ce vous ſera un grand avantage,& une grande épargne ; mais rarement en pourrez-vous trouvez,où il n'y ait beaucoup à travailler , dautant que telle paroîtra paſſablement bonne au deſſus, qui étant ouverte de la profondeur d'un fer de Béche ſeulement, ſe trouvera Argileuſe deſſous ; ce fonds eſt pire aux Arbres que le Tuf,ou la Roche,à cauſe qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder,afin de tirer la fraîcheur de plus bas,& prendre quelqué nourriture ; mais l'Argileuſe ou Terre franche ou rouge, fait comme un plancher qui par ſa dureté & denſité , ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté,em-

Fig. 1. The original binary image

```
m»il)uanrΓ¤, la l'erre,^! vouz la rencomrex
t,onne,ce vouz lera un ^ranΓ¤ avanra^e,sc
u/B»e ^ranΓ¤B« epar^ne ; maiz raremenr en
pourrex-vouz trouvex,on il n'v air tieau-
couri a travailler , claur^nc ^ue teile ria-
rorrra rrallΓ¤lllernenr cionne au clelluz,c^
am ouverre cle la rirokoncleur Γ¤'un iec
Γ¤e Lecrre ieulemenr , re rrouveral ^r?l-
leule cleclouz ; ce fonclz eli riire aux ^r-
Γ¶rez c^ue le l"uf,on la l^t,cke,a caule c^u'il
z'^ reuconrre cle r/erirez veiuez on lez ^.a-
cinez peuvenr z'ccenΓ¤resc rirofoncler,arlB«
cle rirer la ir,uclleurcle^luz baz,sc rirenΓ¤re ^
uelc^ue nourricure; malz l^r^ilcule ou '
I'errre francrre ou rou^e, fair comme un
plancrrer c^ui riar ia Γ¤urere sc clenlrre , ne
peur ecre rierce riar aucunez li.acinez, sc
c^ui cl^nz lez ^r,inclez arcleurz cle l'^,te,emB»
```

Fig. 2. The text obtained by auto OCR

2. The software

The following is a brief overview of the most important parts of our software system and the necessary steps in the searching process.

Input data of the software are collection of files representing text document. Each file is an image of one page of the document. Many graphic formats are acceptable as TIF, JPG, PGN, GIF, etc.

The software system supports three user windows - **Main**, **Parameters** and **Find**. **Main** window displays one page of the document. The current directory containing document image files and the current file name are given on the top of **Main** window. It is possible to go forward and backward through the document pages (see Fig. 3).

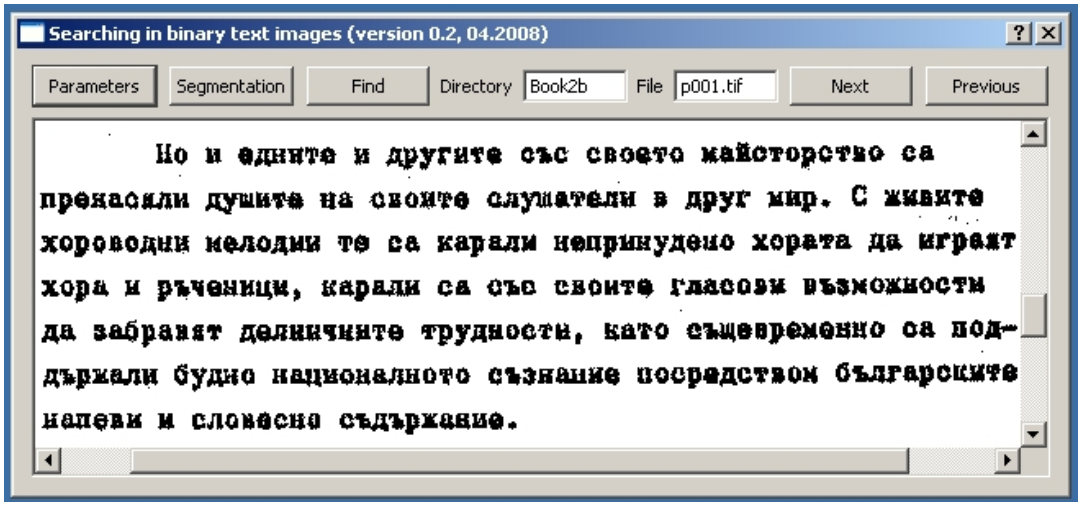


Fig. 3. Main window

Three main steps are essential for successful word searching: segmentation, searching and result representation.

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

To segment the words in a line, we use the vertical projection. - a histogram obtained by counting the number of black pixels in each vertical scan at a given horizontal position. If the words are well separated, the histogram should have areas of zero values between words. Because the distances between words are larger than between characters, it is easier to separate words than characters. Segmentation of words and characters is also an important step in every OCR process.

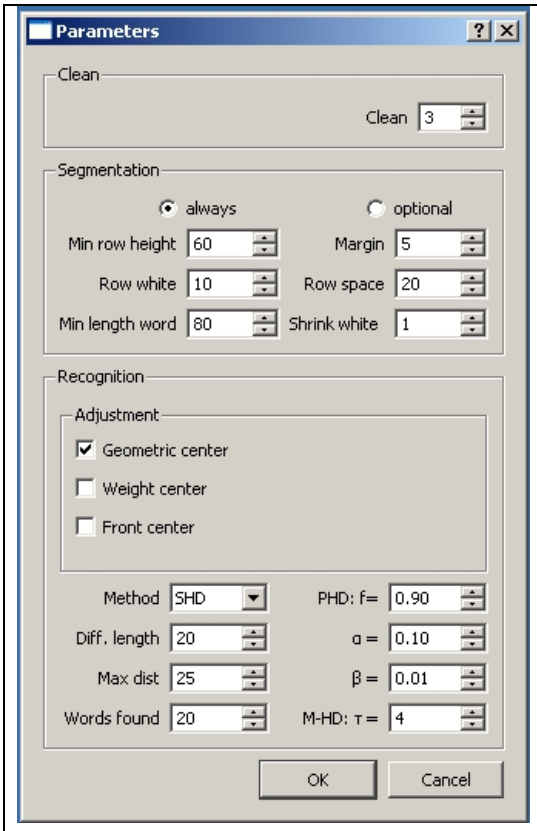


Fig. 4. Parameters window

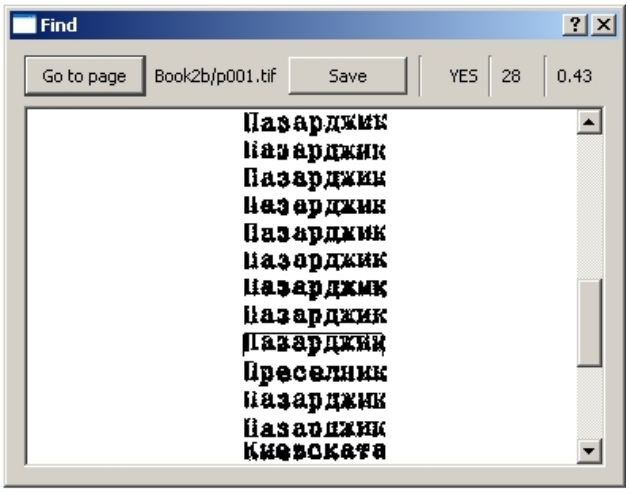


Fig. 5. **Find** window

As a result of the segmentation, every word is associated with a word image -minimal rectangular frame that contains the corresponding word. So we consider any word image as a rectangle, which consists of white and black pixels. The black pixels form a set, which is used in calculating word similarities.

For segmentation step we use a number of parameters, which are important for successful word separation (see Fig. 4):

• **Minimal row height**: The height of every row must be at least the value of this parameter. This helps us avoid creating (due to noise) rows with small height;

• **Margin**: The system reduces the page dimension by the value of this parameter and allows us to process only a part of the page. Also often page images have black lines or fields 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 the white space between the words.

• **Row space**: The white space between words must be greater than the value of this parameter. This help us 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 equals to the length of two or three letter words.

• **Shrink white**: This parameter concerns an additional step conducted when we have already separated words, and words are framed. At this step we try to shrink the height of rectangles, using horizontal histograms only for the points in a given word image. We decrease the rectangle height if the points of histograms have values less than this parameter. This step is very useful when the rows have small slopes (see Fig. 6).

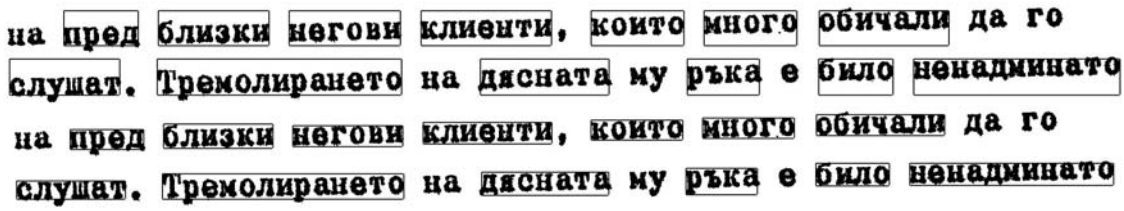


Fig. 6. Small slope of rows: word segmentation before and after the step "shrink"

2.2 Searching. After segmentation of a page, we must choose a pattern word image - this is a word, which we want to find in the document pages.

Searching starts when the user pushes the button **Find** in the **Main** window. It activates the process of inspection all pages for segmentation and measuring similarities of segmented words and the pattern word.

Before calculating the corresponding Hausdorff distance between the pattern word and the word under investigation, we must dispose the word image at a suitable position with respect to the pattern image. We simply calculate a translation vector which adjusts the images. There are three options for defining the translation vector - by connecting geometric centers, mass centers or the left sides of word images (see the frame **Adjustment** in **Parameters** window - Fig. 4).

Although the **Parameters** window gives us the possibility for choosing a number of methods for word similarities, we apply only Modified Hausdorff Distance (MHD) in our experiments. Dubuisson and Jain [7] introduced this method, one of the best measures for words similarities (see [4] for parallel with other Hausdorff type measures).

We can see a part of the retrieval data in **Find** window. The number of words in this window is set in **Parameter** window - the field **Words found**. Pushing **GoTo** button, the page containing the marked word is displayed in **Main** window.

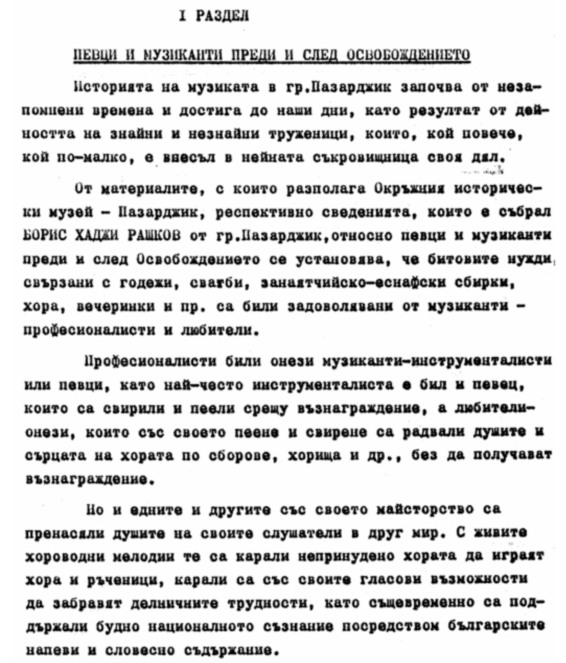
General views of user screens are presented of Figs 3, 4 and 5. The program code is written in C++ with help of Qt - a cross-platform application development framework [8].

The results of searching words are presented on Figs 15–22. The pattern word is indicated by a frame in the text.

4. Conclusion

We presented a software tool for searching in binary text images. We described briefly the possibilities of our preliminary version of the program. Experiments with 7 different languages from various epochs give us the certainty in practical benefit of our approach. It is quite universal and does not require any specific features of the concrete language. Word

searching can be applied to any collection of scanned documents, immediately after the graphic files have been created by the scanner device.



I РАЗДЕЛ

ПЕВЦИ И МУЗИКАНТИ ПРЕДИ И СЛЕД ОСВОБОЖДЕНИЕТО

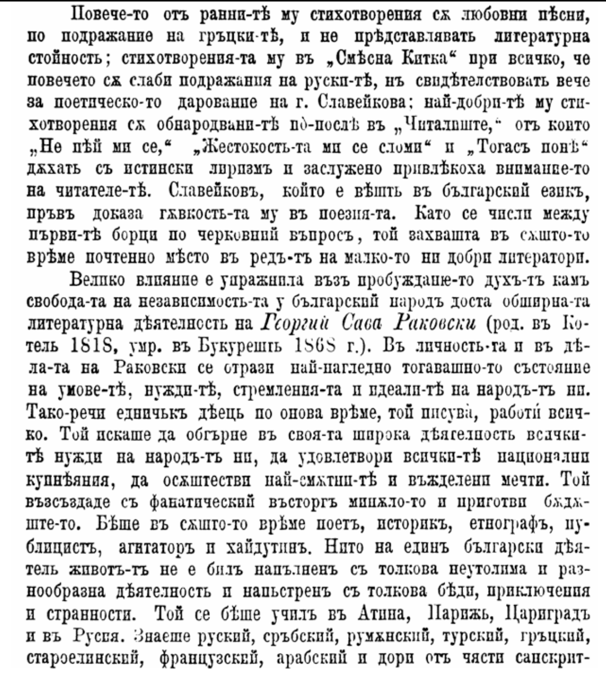
Историята на музиката в гр.Пазарджик започва от незапомнени времена и достига до наши дни, като резултат от дейността на знайни и незнайни труженици, които, кой повече, кой по-малко, е внесъл в нейната съкровищница своя дял.

От материалите, с които разполага Окръжния исторически музей – Пазарджик, респективно сведенията, които е събрал БОРИС ХАДЖИ РАШКОВ от гр.Пазарджик,относно певци и музиканти преди и след Освобождението се установява, че битовите нужди, свързани с годежи, сватби, занаятчийско-еснафски сбирки, хора, вечеринки и пр. са били задоволявани от музиканти – професионалисти и любители.

Професионалисти били онези музиканти-инструменталисти или певци, като най-често инструменталиста е бил и певец, които са свирили и пеели срещу възнаграждение, а любители-онези, които със своето пеене и свирене са радвали душите и сърцата на хората по сборове, хорища и др., без да получават възнаграждение.

Но и едните и другите със своето майсторство са пренасяли душите на своите слушатели в друг мир. С живите хороводни мелодии те са карали непринудено хората да играят хора и ръченици, карали са със своите гласови възможности да забравят делничните трудности, като същевременно са поддържали будно националното съзнание посредством българските напеви и словесно съдържание.

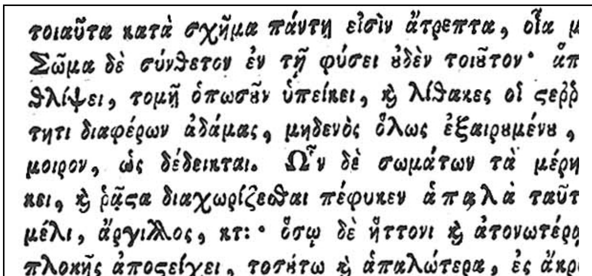
Fig. 7. Bulgarian typewritten document



Повече-то отъ ранни-тѣ му стихотворения сѫ любовни пѣсни, по подражание на грѣцки-тѣ, и не прѣдставлявать литературна стойность; стихотворения-та му въ „Смѣсна Китка“ при всичко, че повечето сѫ слаби подражания на руски-тѣ, нъ свидѣтелствовать вече за поетическо-то дарование на г. Славейкова: най-добри-тѣ му стихотворения сѫ обнародвани-тѣ по-послѣ въ „Читалище,“ отъ които „Не пѣй ми се,“ „Жестокость-та ми се сломи“ и „Тогасъ понѣ“ дъхать съ истински лиризмъ и заслужено привлѣкоха внимание-то на читателе-тѣ. Славейковъ, който е вѣщь въ българский езикъ, прьвъ доказа гъвкость-та му въ поезия-та. Като се числи между първи-тѣ борци по черковний въпросъ, той захваща въ сѫщо-то врѣме почтенно мѣсто въ редъ-тъ на малко-то ни добри литератори.

Велико влияние е упражнила възъ пробуждание-то духъ-тъ камъ свобода-та на независимость-та у българский народъ доста обширна-та литературна дѣятелность на *Георгий Сава Раковски* (род. въ Котель 1818, умр. въ Букурешть 1868 г.). Въ личность-та и въ дѣла-та на Раковски се отрази най-нагледно тогавашно-то състояние на умове-тѣ, нужди-тѣ, стремления-та и идеали-тѣ на народъ-тъ ни. Тако-речи единичькъ дѣецъ по онова врѣме, той писува, работи всичко. Той искаше да обгърне въ своя-та широка дѣятелность всички-тѣ нужди на народъ-тъ ни, да удовлетвори всички-тѣ национални купнѣяния, да осѫществи най-смѫтни-тѣ и въжделени мечти. Той възсъздаде съ фанатический въсторгъ минѫло-то и приготви бѫдѫще-то. Бѣше въ сѫщо-то врѣме поетъ, историкъ, етнографъ, публицистъ, агитаторъ и хайдутинъ. Нито на единъ български дѣятель животъ-тъ не е билъ напълненъ съ толкова неутолима и разнообразна дѣятелность и напьстренъ съ толкова бѣди, приключения и странности. Той се бѣше училъ въ Атина, Парижъ, Цариградъ и въ Русия. Знаеше руский, сръбский, ромѫнский, турский, грѣцкий, староелинский, французский, арабский и дори отъ части санскрит-

Fig. 8. Bulgarian book



τοιαῦτα κατὰ σχῆμα πάντῃ εἰσὶν ἄτρεπτα, οἷα μ
Σῶμα δὲ σύνθετον ἐν τῇ φύσει οὐδὲν τοιοῦτον· ἅπ
θλίψει, τομῇ ὁπωσοῦν ὑπείκει, καὶ λίθακες οἱ στερρ
τητι διαφέρων ἀδάμας, μηδενὸς ὅλως ἐξαιρουμένου,
μοιρον, ὡς δέδεικται. Ὧν δὲ σωμάτων τὰ μέρη
κει, καὶ ῥᾷστα διαχωρίζεσθαι πέφυκεν ἁπαλὰ ταῦτ
μέλι, ἄργιλλος, κτ:· ὅσῳ δὲ ἧττονι καὶ ἀτονωτέρ
πλοκῆς ἀποστείχει, τοσούτῳ καὶ ἁπαλώτερα, ἐς ἄκρ
οὐδενὶ γάρ ἐστιν ἐντυχεῖν, οὗ τὰ μέρη μὴ ὁπωσοῦν
γνύμενα.

Γʹ. Τὸ σκληρὸν σῶμα ὑπὸ κούφης τε καὶ ἀσθεν
πεφυκὸς, εὔθραυστον ἀκούει· τοιαῦτα χάλυψ
τὰ κεράμεια σκεύη· Τούτων τὰ στερεὰ μέρη οὐχ οὕτω
ἀλλήλοις, διὸ καὶ ῥᾷστα τῆς ἀμοιβαίας παφῆς ἀφίστατ

Δʹ. Τὸ ἐκ πολλῶν οἱονεὶ λεπίδων πάνυ λεπτῶν ἀ
στάμενον, εὔσχιστον σῶμα προσείρηται· τούτου τὰ
μόρια στερρότερον προσκεκόλληται ἀλλήλοις, ἢ λεπὶς
ἄρα τὰ τοιαῦτα σώματα εἰς λεπίδας ἀναλυόμενα·
καλούμενοι λίθοι οἱ ἐξ Ἰβηρίας, καὶ Καππαδοκίας,

Fig. 9. Old Greek text

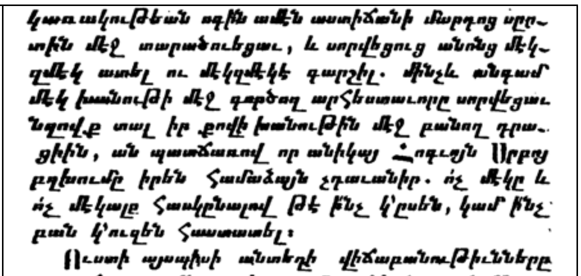
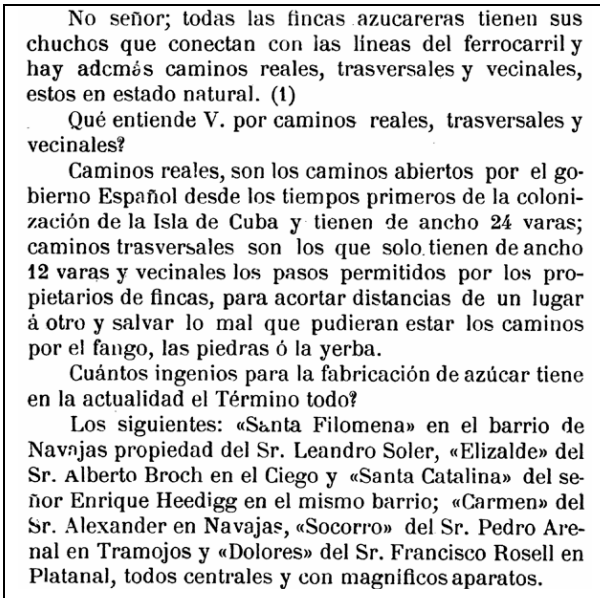


Fig. 10. Armenian book



No señor; todas las fincas azucareras tienen sus chuchos que conectan con las lineas del ferrocarril y hay además caminos reales, trasversales y vecinales, estos en estado natural. (1)

Qué entiende V. por caminos reales, trasversales y vecinales?

Caminos reales, son los caminos abiertos por el gobierno Español desde los tiempos primeros de la colonización de la Isla de Cuba y tienen de ancho 24 varas; caminos trasversales son los que solo tienen de ancho 12 varas y vecinales los pasos permitidos por los propietarios de fincas, para acortar distancias de un lugar á otro y salvar lo mal que pudieran estar los caminos por el fango, las piedras ó la yerba.

Cuántos ingenios para la fabricación de azúcar tiene en la actualidad el Término todo?

Los siguientes: «Santa Filomena» en el barrio de Navajas propiedad del Sr. Leandro Soler, «Elizalde» del Sr. Alberto Broch en el Ciego y «Santa Catalina» del señor Enrique Heedigg en el mismo barrio; «Carmen» del Sr. Alexander en Navajas, «Socorro» del Sr. Pedro Arenal en Tramojos y «Dolores» del Sr. Francisco Rosell en Platanal, todos centrales y con magnificos aparatos.

Fig. 11. Text in Spain

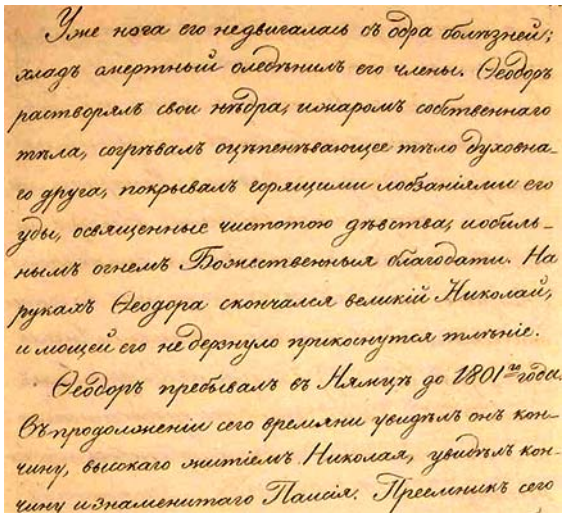
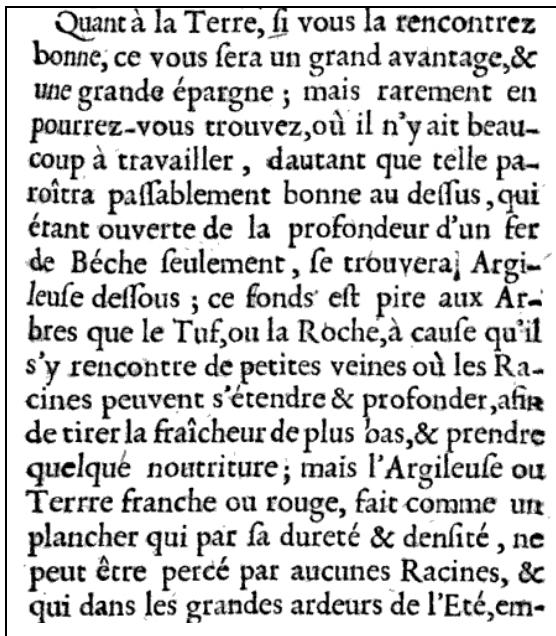


Fig. 12. Handwritten document in Russian



Quant à la Terre, ſi vous la rencontrez bonne, ce vous ſera un grand avantage,& une grande épargne ; mais rarement en pourrez-vous trouvez,où il n'y ait beaucoup à travailler , dautant que telle paroîtra paſſablement bonne au deſſus, qui étant ouverte de la profondeur d'un fer de Béche ſeulement, ſe trouvera Argileuſe deſſous ; ce fonds eſt pire aux Arbres que le Tuf,ou la Roche,à cauſe qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder,afin de tirer la fraîcheur de plus bas,& prendre quelque nourriture ; mais l'Argileuſe ou Terrre franche ou rouge, fait comme un plancher qui par ſa dureté & denſité , ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté,em-

Fig. 13. Text in French

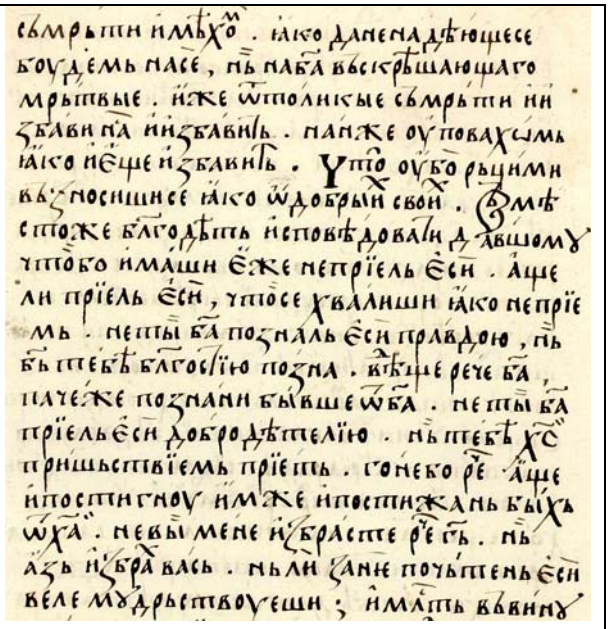


Fig.14. Slavonic manuscript

There are some possibilities for improvements in the software.

- Increasing the efficiency and speeding up the search.
- Searching with a part of word as a pattern.
- Character segmentation of a page (or pages) and composing pattern word from well separated letters.
- Feedback -- making second search for the same word with a different pattern word.
  - The user can choose this word among the correct words found in the first search (see [5]).
  - Produce a new pattern as an average of all or part of the correct words.
- Automatic or semi automatic choice of parameters based on image information.

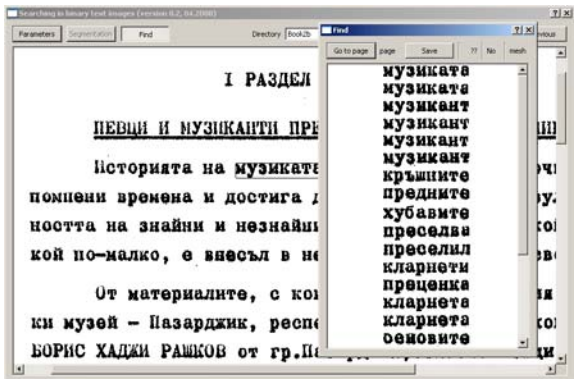


Fig.15. Bulgarian typewritten document

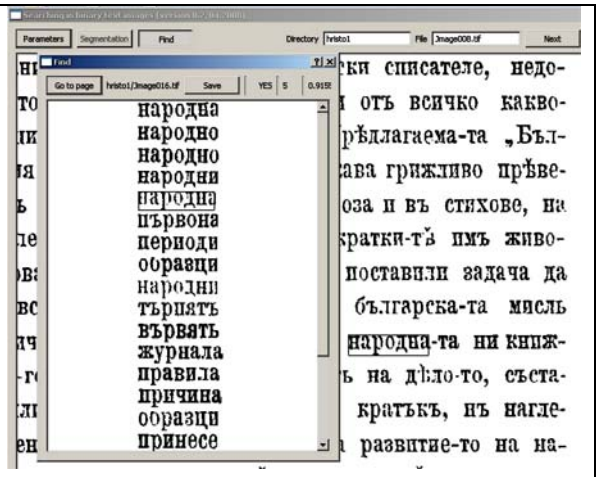


Fig.16. Bulgarian book

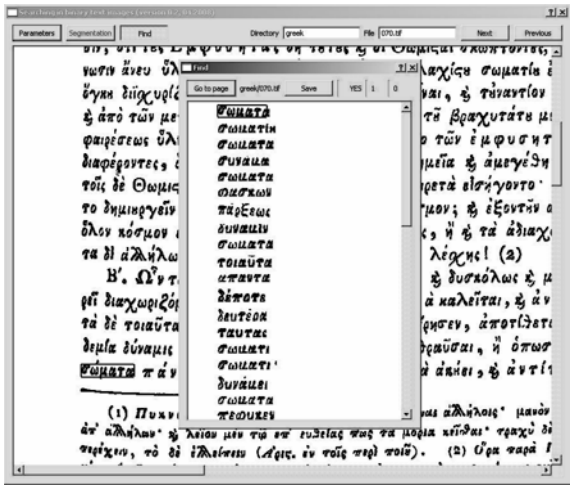


Fig.17. Old Greek text

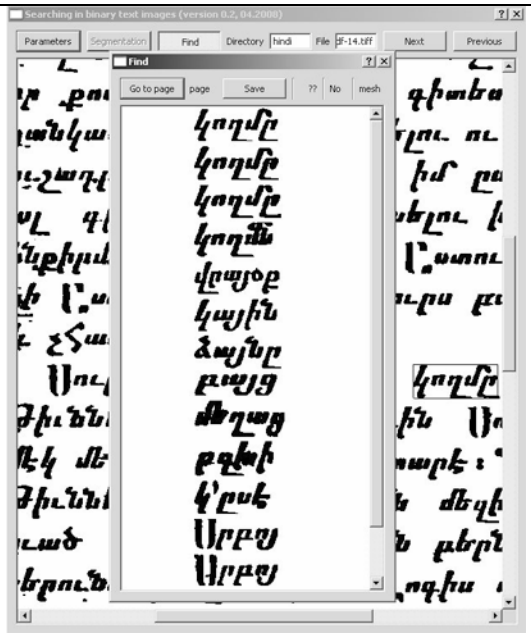


Fig. 18. Armenian book

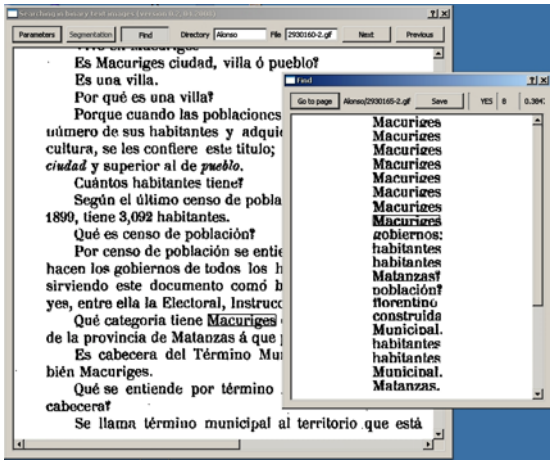


Fig.19. Text in Spain

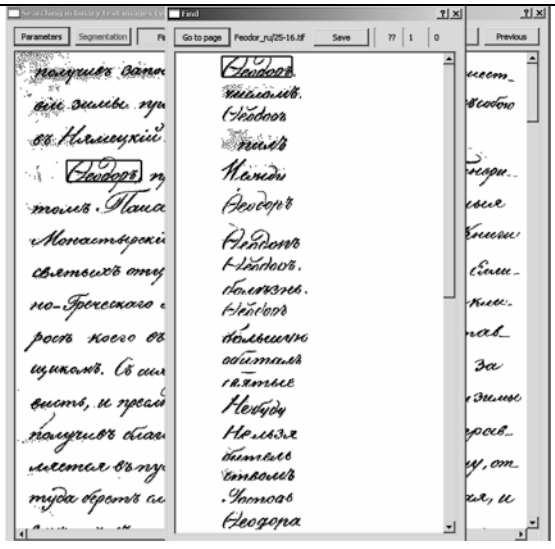


Fig.20. Handwritten document in Russian

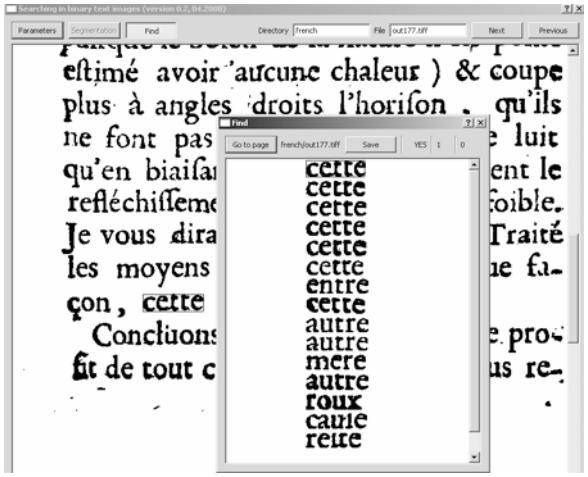


Fig. 21. Text in French

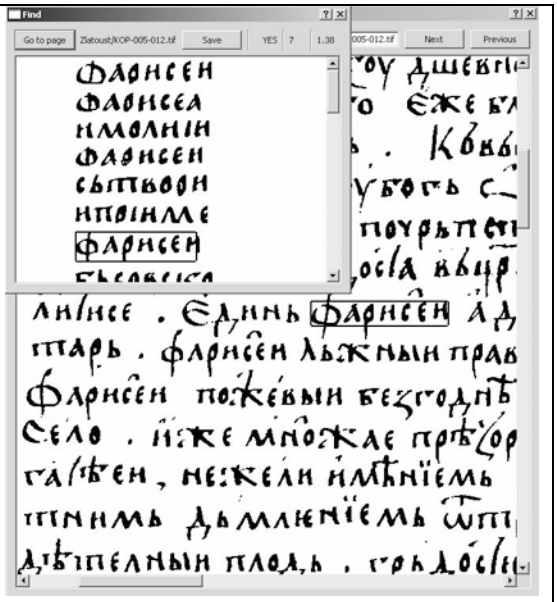


Fig. 22. Slavonic manuscript

In spite of relatively low efficiency of the Hausdorff type methods [4] (the searching process takes a lot of time) we believe that even the modern, high level personal computers will be able to solve the problem within a reasonable time. We think also that the accuracy of retrieval is sufficient for practice.

References

1. A. Andreev, N. Kirov, *Hausdorff Distance and Word Matching*, Proceedings of the International Workshop "Computer Science and Education", June 3–5, 2005, Borovetz–Sofia, Bulgaria, 19-28.
2. A. Andreev, N. Kirov, *Word image matching in Bulgarian historical documents*, NCD Review 8 (2006), 29–35.
3. A. Andreev, N. Kirov, *Some Variants of Hausdorff Distance for Word Matching*, NCD Review 12 (2008), 3–8.
4. A. Andreev, N. Kirov, *Text Search in Document Images Based on Hausdorff Distance Measures*, Proc. CompSysTech'08, 2008 (accepted).
5. T. Konidaris, 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*, Internat. J. Document Anal. Recognition 9 (2007), 167–177.
6. Hwa-Jeong Son, Soo-Hyung Kim, Ji-Soo Kim, *Text image matching without language model using a Hausdorff distance*, Information Processing and Management (2008), to appear.
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, Ћирилски 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 tйrmino 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=uxgOAAAAQAAJ&hl=bg
12. Дом живоначальной 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 (from books.google.com)

nkirov@nbu.bg