Computerized Recognition System for Historical Manuscripts

Artem Skabin

Petrozavodsk State University, Petrozavodsk, Russia
artb00g@gmail.com

Abstract. The article describes the process of creating a universal computerized recognition system of historical manuscripts, including historical shorthand records dating back to the 19th and early 20th centuries. We discuss the problem of getting the original graphical representation of symbols from historical manuscripts using a threshold binarization method. We search for a similar graphical representation of symbols in the database. Moreover we present a prototype of a computerized recognition system of historical manuscripts.

Keywords: character recognition, historical manuscript, binarization, threshold method.

1 Introduction

Nowadays Russian archives have a large amount of encoded shorthand records. The reason is the shorthand writers' inability to decipher historical records. During the 19th and early 20th centuries Russian shorthand writing was in the making. So the existing documents were deciphered in different systems. Moreover, modern shorthand writing differs significantly from the historical stenography systems of the 19th century. The main difficulties in decoding shorthand records are:

1. the lack of specialists in the area of 19th and early 20th century shorthand writing systems. There are only old books;
2. the shorthand writer's inability to use standard characters, because usually stenographists deciphered the texts themselves and used their own symbols;
3. there was a widespread custom of skipping vowels or replacing repeated combinations of characters and words with one symbol;
4. the fact that some characters from shorthand records could have a similar spelling, but depending on certain physical parameters such as height, can have different meanings.

The aim of this work is to create a universal computerized recognition system of historical manuscripts, including from historical shorthand records of the 19th and early 20th centuries. It is to solve the problem of description and decoding historical transcripts, as well as to introduce new documents to the scientific world.

2 Description of the developed system

Special features of the developed system are a historical account of the 19th and 20th centuries spelling characteristics, an account of the individual characters of different shorthand writers, the ability of critical analysis, the usage of dictionaries for help in deciphering the texts, etc. [1] The information system will be publicly available and offered to be used by archive professionals and librarian scientists. The fine-tuning of the system was done using the transcripts by Snitkina partially decoded C. Poshemyanskoy and P. Olkhin’s book [2]. Recognition of any text includes the following steps:

1. image preprocessing, usually an image binarization;
2. segmentation, i.e. selection of the text in the preprocessed image, such as characters, combinations of characters, words, lines;
3. analysis of the derived segments by establishing the values, characteristics, comparing with reference standards that could be found in the knowledge base;
4. decoding by choosing the most appropriate word forms from dictionaries of equivalents with a specific language model.

Additional difficulties in the text recognition are caused by curving rows, brightness drops, transparency of the text on the reverse side and other defects of the original text and image. Manuscript recognition is more difficult in contrast to the recognition of printed texts [3].

The goal of the research is to create fairly universal software for computerized recognition of historical manuscripts for which it has been impossible in the past. The suggested computerized recognition system of historical manuscripts with the possibility of intelligent decision support will significantly accelerate the process of conversion of manuscripts into text files and increase the accuracy of their decipherment. The software will have the following characteristics [1]:

1. the system automatically monitors the state of keying in and interactively displays information to the user;
2. the system returns the user the word forms variants sorted by frequency of occurrence in the database and information on the absence of the words typed in the database.

3 Binarization of historical manuscripts

While decyphering historical manuscripts a problem with image binarization occurs. Due to the aged condition of the image and the fact that the shorthand records were made in pencil on yellow paper the threshold method in color components (RGB) was not suitable for this task. This was caused by the characters and background pixels which have similar values as color components. As it can be seen on the histograms (Fig. 1) the absence of two clearly defined peaks, does not allow choosing the threshold value for binarization. Similar results are observed (Fig. 1), if we use HSB colour scheme decomposition (hue, saturation, brightness). If using threshold method binarization of brightness you can get a clear character, with a small amount of noise.

Threshold intensity was experimentally determined by choosing those values for which the most precise symbols came out with the least amount of noise. Binarization was performed on around 1500 fragments from these 30 historical documents. The best results are achieved when the percentage of black pixels after binarization is approaching 13% of the total number of pixels.

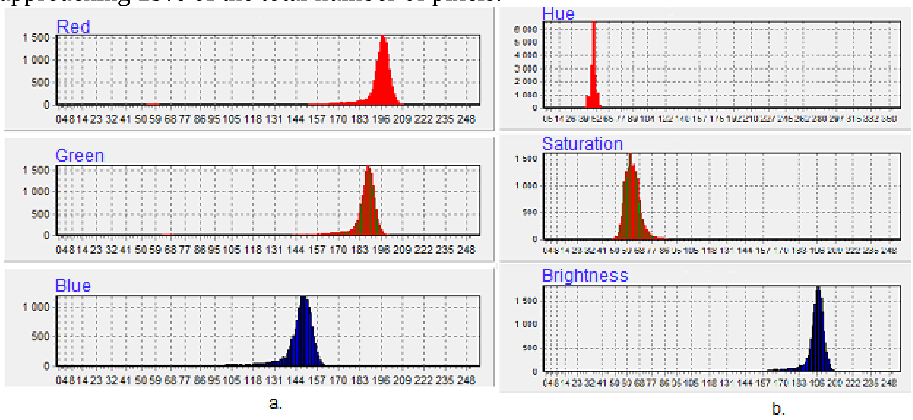


Fig. 1. Histograms of color schemes RGB (a) and HBS (b).

4 The module for creating original graphical representation of characters

The system was divided into several modules. One of them is the module for creating an original graphic representation of the characters. Fig. 2 shows an interface module for creating original graphic characters. The main window consists of two areas: the original image (original transcript) is on the left panel where the user selects the desired character (Fig. 2), the selection location is displayed on the second panel. There the processed shorthand record is located, i.e. all of the characters that could be found in the same place where the original symbols are.

After selecting a character the user should click on the "hot key" or a combination thereof. Then the system performs binarization selection and segmentation. If you receive multiple segments, the system prompts the user to choose which segment or segments correspond to the original character. If you selected several segments, the system binds [4] the broken "pieces" and provides the user with the result. When the user is satisfied with the result, the symbol is saved in the database and is located on the right panel respectively to the location (coordinates) on the original image. If the result does not meet user’s requirements, it is possible to edit the received symbol with a simplified graphics editor.

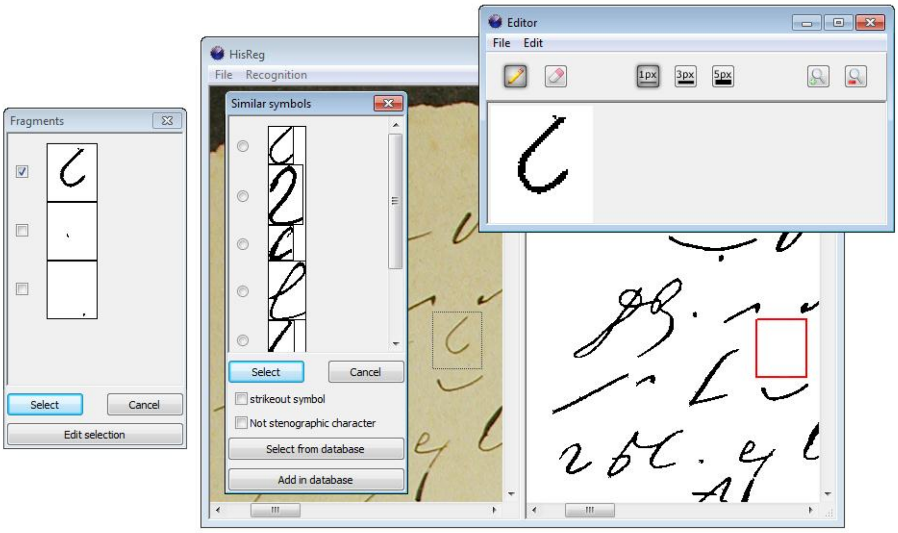


Fig. 2. Interface module for creating original graphical representation of characters

The creation of an original graphical representation of characters is a difficult task for the following reasons:

1. The original image is quite old and was written in pencil on yellowed paper which has distortions, various types of damage; moreover some shorthand records have irrelevant records with no meaning or there are lines intersecting with the symbols;
2. There were gaps of characters in binarization, as some pixels of the character had a similar color to the pixels of the paper;
3. There was a need for segmentation into individual characters of the symbols that were written together.

5 Character search in the database

During the creation of original graphic characters, the problem of forming a characters database appeared. The database is needed to avoid duplication and redundancy of graphic symbols. The database is extensible, i.e. if in the process of recognition the character has not been previously met, it is added to the database. The basis of the database was taken from a sample of 250 characters, which is equal to the number of characters in the alphabet used by Snitkina, randomly selected from the manuscripts. We used the following methods to compare the current characters with the characters from the database: pattern matching, the comparison with the skeleton of the pattern, the boundary distance method and method of projections. Comparison of these methods is represented in Table 1.

Table 1. Comparison of methods for the learning sample.

characteristic method	Search time	Accuracy
pattern matching	3 sec. (depending on the size of the symbol)	< 30%
the comparison with the skeleton of the pattern	1-2 sec. (depending on the size of the symbol)	~40%
method of projections	~0.5 sec.	~40%
the boundary distance method	< 0.01 sec.	> 60%

The low accuracy of the comparison with the pattern is caused by binarization. The character could have different thickness depending on the size of the shorthand records selected. While comparing the skeletons of characters, for skeletolization we used Zhang Suen's algorithm [5]. This algorithm for finding characters in the database works as follows: using the method of classification symbols of height to width ratio, the characters are divided into three classes: wide, high, square. Next, is determined to which class the current character belongs. And then there is a search for similar character in this class using the boundary distances method. The boundary distances method consists of choosing symbols with a similar height to width ratio. Then the current symbol is measured $\{l_1, l_2, \ldots, l_8\}$ (Fig. 3).

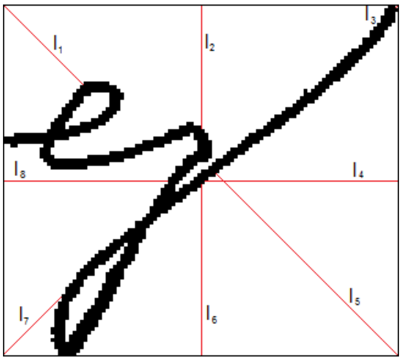


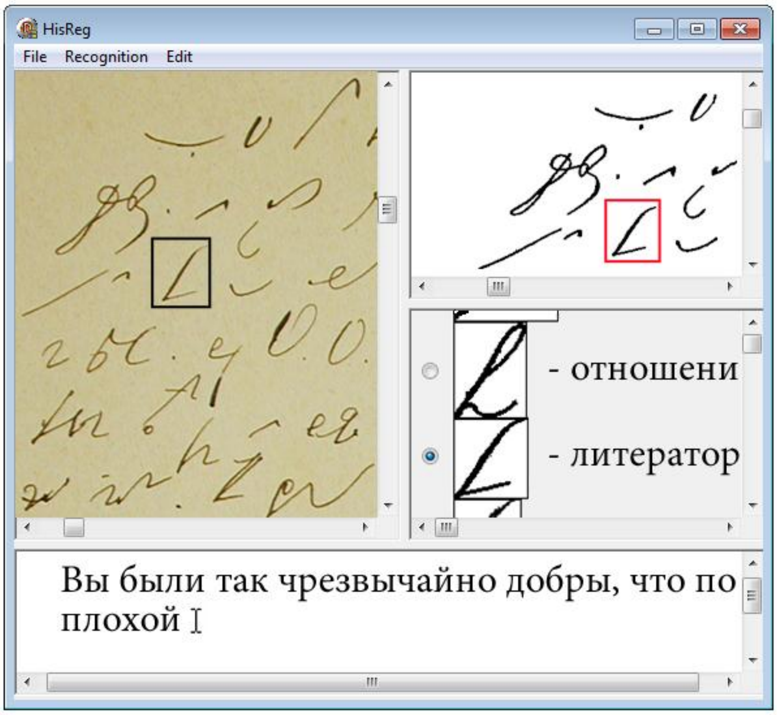
Fig. 3. The boundary distance method

The characters, that are selected from the database, have the distance $\{l_1', l_2',\ldots, l_8'\}$ and are located in the interval ($l_1 \cdot k$ - ε, $l \cdot k$ + ε), where k is the height to width ratio of the current symbol, $\varepsilon = k \cdot l \cdot \alpha$ where $\alpha = 0.1$.

During processing of 29 sheets of shorthand records, more than 2,500 diagrams of the original characters have been allocated. After that we met a problem of matching the distance parameters derived from the book Snitkina and partially transcribed records. The prototype interface of this system is presented below.

6 The prototype of computerized recognition system of historical manuscripts

Fig. 4 shows the interface of the prototype of our computerized recognition system for historical manuscripts. This system has four main areas: the area with the original image, the range of possible values of deciphered symbols or groups of characters.



○ - отношени

● - литератор

Вы были так чрезвычайно добры, что по
плохой

Fig. 4. Interface of the computerized system for historical manuscript recognition.

When the user selected a character in the original image, the image of the symbol is located in the 2nd panel, at the same place where it is on the original shorthand record. The fourth area displays the decoded symbols. The system analyzes components of the word in the process of keying it in, and offers the user the closest interpretations in meaning from the database. The system produces an automatic decoding of similar characters or groups of characters by analyzing the original image while the characters are keyed in. The main advantages of this system are the following:

1. the ability to use "hot keys" which accelerates the keying in of a historical shorthand record;
2. the connection of graphic and text representation of a shorthand record;

3. intellectualized typing;
4. the ability to automatically recognize similar combinations of characters, words in the text;
5. the opportunity to work with a multi-user dictionary.

This system is designed to accelerate the process of deciphering handwritten historical shorthand records. The system's future realization as a Web-service for organizations working with shorthand records is possible.

7 Conclusion

The threshold method of image binarization is used in the module; with the selected parameters the binarization goes to the utmost. These parameters are specific to each type of document, so there is a need to use a more adaptive method of binarization as described in [6].

Current methods for image search on database do not provide high accuracy; because of that original symbol diagram redundancy may occur. As a result, it is necessary to use a method that gives a higher accuracy. This will be analyzed by the methods of individual signatures verification [7].

Acknowledgments. This research supported by a grant RGNF № 11-01-12026v (head Rogov AA). We would like to thank Dmitry Ignatov and Jonas Poelmans for language improvements of the final submission.

References

1. Rogov, A.A., Talbonen, A.N., Varfolomeev, A.G.: Automated recognition of handwritten historical documents. Digital libraries: advanced methods and technologies, digital collection: Proceedings of the XII All-Russian Scientific Conference RCDL'2010, Kazan. Univ of Sciences, P. 469-475 (2010) (in Russian)
2. Olkhin, P.: Guide to the Russian shorthand. Printing Dr. M. Khan, St. Petersburg (1866) (in Russian)
3. Gorski, N., Anisimov, V., Gorskaya, L.: Mountain handwriting recognition tech-hundred: from theory to practice. Polytechnics, St. Petersburg (1997) (in Russian)
4. Nagabhushan, P., Anami, B. S.: A knowledge-based approach for recognition of handwriting Pitman shorthand language storkes. In: P. Nagabhushan, Basavaraj S. Anami. Sadhana., Vol. 27, Part 5, P. 685–698 (2002)
5. Zhang, T.Y.: A fast parallel algorithm for thinning digital patterns / T. Y. Zhang, C. Y. Suen. Commun. ACM. Vol. 27, №3, P. 236-239 (1984)
6. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). In ICDAR, P. 1506–1510 (2011)
7. Kukharev, G.A.: Biometric Systems: Methods and means of identification of human personality. Polytechnics, St. Petersburg (2001) (in Russian)