

Automatic Letter Style Recognition of Churchslavic Manuscripts

Mimoza Klekovska¹, Igor Nedelkovski²
Vera Stojcevska-Antic³ and Dragan Mihajlov⁴

Abstract – In this paper the research on the churchslavic alphabet recognition in old hand-written papers is presented. The known OCR and ICR recognition methods in both printed and hand-written texts are analyzed. Their advantages and disadvantages are calculated by accepting some of their advantages and finally, two new modification methods are suggested. Churchslavic alphabet has a dual nature: it is a hand-written alphabet, but it looks like a printed one.

Keywords – Churchslavic alphabet, OCR, ICR, Picture processing, Font, Geometry analyses, ICEEST 2009.

I. INTRODUCTION

Paperless office, where paper documents are copied onto modern recording media such as CD/DVD, is no doubt an intention in the modern world. Many OCR (Optical Character Recognition), ICR (Intelligent Character Recognition) programs are developed accordingly. There are two basic methods used by those programs: a Matrix Matching and a Feature Extraction. The Matrix Matching compares what the OCR scanner is scanning as a character, comparing it with a library of character matrices or templates. The Feature Extraction is also known as Topological Feature Analysis or ICR. This method varies by how much “computer intelligence” is applied by the particular manufacturer. The computer looks for general features such as open areas, closed shapes, diagonal lines, line intersection etc. The newest versions of the OCR programs support a lot of natural (human) languages (about 180). The churchslavic language or alphabet is not included in those 180 languages. It is a “dead” alphabet not in active use today, but many libraries in the world keep a lot of papers written by this alphabet. There are common situations where two sheets of a particular old book are in a library in, say, Leningrad, 10 sheets in a library in Sofia, in France... The linguists in this field need to exchange information. They use microfilms, photos, descriptions in edited publications etc. It is very difficult to have an access to the originals and normally it is not suitable to operate with

¹Mimoza Klekovska is with the Technical Faculty of Bitola, address: ul. Ivo Ribar Lola bb, 7000 Bitola, Macedonia, e-mail: mimiklek@yahoo.com.mk

²Igor Nedelkovski is with the Technical Faculty of Bitola, address: ul. Ivo Ribar Lola, 7000 Bitola, Macedonia, e-mail: igor.nedelkovski@uklo.edu.mk

³Vera Stojcevska-Antic is Professor Emeritus, e-mail: veranino@gmail.com.

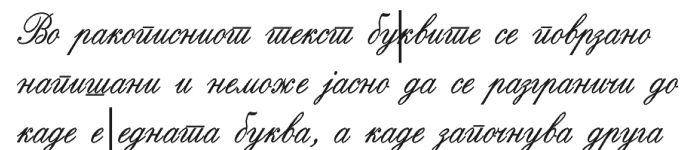
⁴Dragan Mihajlov is with the Faculty of Electrical Engineering and Information Technologies of Skopje, address: Karpos bb, 1000 Skopje, Macedonia, e-mail: dragan@feit.ukim.edu.mk

such old originals, but it is better for the originals to be scanned and available in electronic format. The number of originals is quite big and it is much better to keep them in text format. Reprinting old manuscripts in authentic form is a specific task, not very popular among computer engineers. Recognizing the manuscripts and treating them by the actual OCR programs is not popular too. The intention of this work is to initiate a process of including the churchslavic alphabet in the list of OCR programs.

Maybe this is not an urgent problem; the artifacts will be the same in the years to come, but those papers are keeping humanity's previous intelligence for many centuries and should be kept as a guide and a memory for its past and its history.

II. THE NATURE OF ALPHABET

The churchslavic alphabet is a handwritten one, but there are many differences in resolving the problems between other handwritten texts and this one. The most important difference is concerning the interconnection of the characters. The manuscript letters (characters) in other handwritten texts are usually joined or written connected (Fig. 1), but in the churchslavic manuscripts the letters (characters) are written separately (standalone), like in the printed texts.



*Во ракописности тексти буквите се поврзано
напишани и неможе јасно да се разграничи до
каде е еднашната буква, а каде започнува друга*

Fig. 1. Connected letters in handwritten manuscripts

So, the churchslavic manuscripts look like a printed document. But there are many additional problems in the handwritten manuscripts than in the printed texts.

Each normal printed text has the same h-high value for the letters in the font. It has a horizontal (or straight) line as the base line. The ascenders, descenders and capitals are always in the equal straight line. In the churchslavic alphabet these rules are not valid.

**Prikaz na oddalecenosta vo
pecaten tekst kade se gleda deka
tekstot ima horizontalna bazna linija |
se naogja svetol megjuliniski prodor**

Fig. 2. Horizontal, straight “white” line in the printed texts

It is possible to find a horizontal or nearly horizontal "white" line between the rows in the printed texts (Fig. 2).

That is not applicable in the churchslavic language (Fig. 3) because of two reasons.

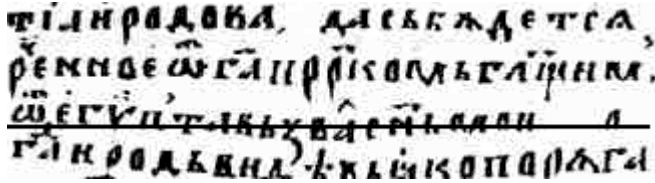


Fig. 3. There is no horizontal straight "white" line

The first one is the orthographic rule to put many upper letters and upper signs (even upper sign above upper letter) in the space between "normal" rows. The second one is because of very frequent use of initials (Fig. 4) or other decorative elements at the beginning of the rows.

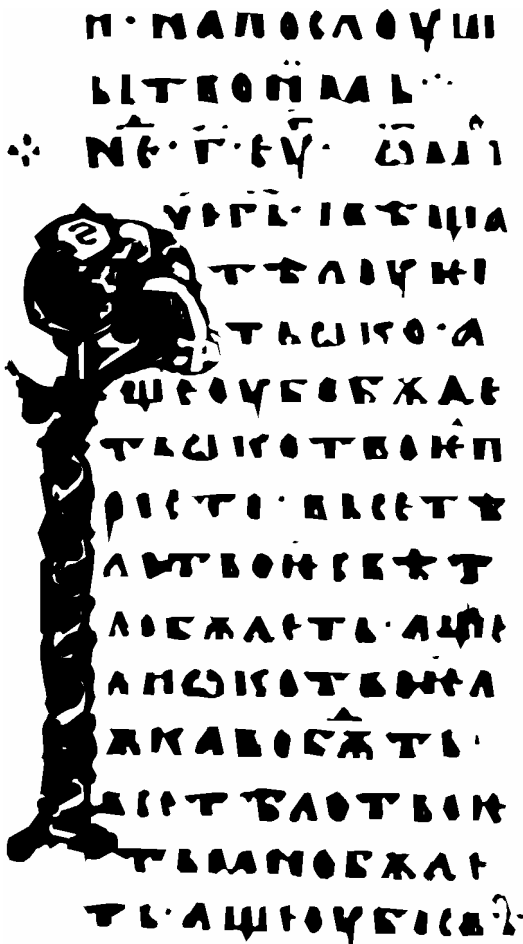


Fig. 4. Initial letter extended over many rows

The decorative elements are extended over many rows (between 5 and 12) and connect the rows in a block.

The decorative element should be cut and deleted to continue the processing intended to find the rows.

In modern printed texts there are spaces between the words; in some cases there is also a thin white space even between the letters in the word (Fig. 5).

Некаде се најдува вертикален празен простор, некаде нема

Fig. 5. White space between words, somewhere between letters

The churchslavic rule is "scripta continua", meaning that there is no space between words (Fig. 6). It is not known where the previous word ends and where the next word begins. This produces a problem in searching the whole word in the corresponding dictionary, like other OCR programs do.

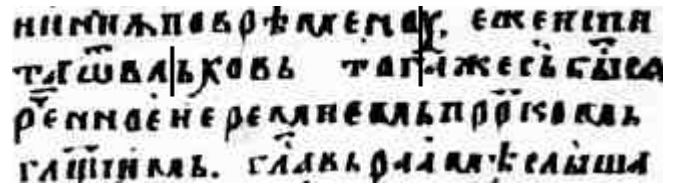


Fig. 6. "Scripta continua" rule in churchslavic manuscripts

But there are still some advantages, used in recognizing printed and handwritten texts that can also be used in processing the churchslavic alphabet. Preprocessing the picture, cleaning it, correcting the slant angle are common problems already treated by the scientific articles of other authors

The aim is to continue in analyzing the geometrical proportions of churchslavic alphabet letters. The decision was made to work in a graphic mode. Monitor's surface is the working area (Fig.7). In that way there is always a visual touch with the evaluation of the author's guiding idea. The Berger's thought: "Seeing comes before speaking" was the driving idea in choosing this operating mode. The pixel is the smallest working measurement. The starting base can be a page of any churchslavic manuscript. The accepted criteria was that the scanned pictures to work with have to be with the same resolution.

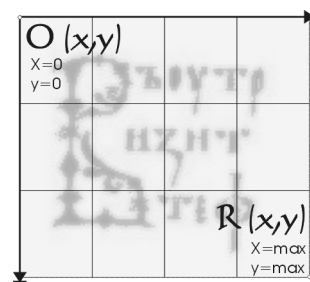


Fig. 7. Rectangular coordinate system over the manuscript

The upper left corner of the manuscript shown on the screen is at the coordinate position (0, 0) in the rectangular coordinate system and the lower right corner at the maximum position (x, y). In determining the average height of the row, the number of rows on the page as a data in the analysis is entered.

The first step was to make a classification of the manuscripts according to their appearance within the

historical period between the 10-Th and the 18-Th century. By some logic, the estimate was that different art styles should have a trace in the written style of the particular century. But it was encountered that there are no dramatic style differences during that period. The next step was to follow the logic of interpolating the manuscripts by the style of scriptor's font (handwriting font). It is similar to the today's light, normal, bold or italic font style. Graphologists would say: "It depends on the writer's personality".

The basic determining measurement is the h-high value. The number of rows on the page based on the value on the y-axis in the letter matrix (in pixels) will direct the activities to calculate h-high of the text row in the range between the minimum and the maximum limit.

The first black pixel on the picture is identified. Then the contour starting with that pixel is followed (tracked down). The number of pixels is counted and their x and y coordinates are compared. The smallest rectangle keeping the form of the sign-letter matrix is determined. High or y_value of the matrix is indicating that either the sign is a valid letter or not. If it is too high there is probably an initial letter or accidentally connected rows. If it is too small, that sign is probably upper letter or upper sign. If it is a valid letter, it is transferred in a new domain to keep it for further analyses of their characteristics; otherwise it is deleted. At the end of this process only the proper signs (letters) on the page (Fig.8) are kept. Initials, other decorative elements, upper letters, upper signs and some undefined small features are removed.

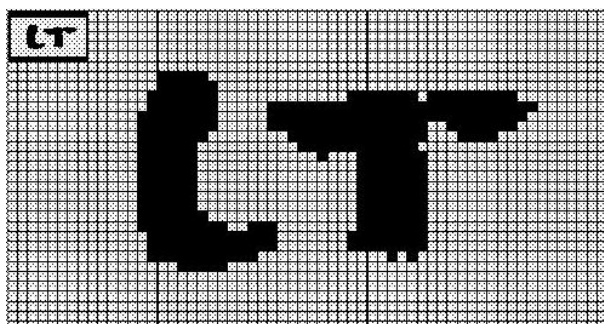


Fig. 8. Proper letters in the matrix are kept

Then each matrix of every particular letter that was preserved (memorized) as a valid sign is analyzed. The fields of interest are:

- Number of matrix rows,
- Number of matrix columns,
- Number of total pixels in the matrix (square surface),
- Number of black pixels and its percentage according to the total pixel numbers,
- X : y ratio,
- Y : x ratio,
- Whether the matrix is symmetrical or not,
- Ratio between contour length and square surface,
- Number of "dictuses" in the letter - primitive features ("dictus" is the trace (or black pixels) from the moment when the pen have touched the paper until the moment when the writer picked up the hand (the pen)),
 - Dictus positions (horizontal, vertical, upper-left slanting stroke, upper-right slanting stroke or undefined),

- The weigh (thickness) of the dictus track (line),
- The existence of white holes inside the contour,
- The existence of high percentage black pixels in some specific topological areas in the matrix, such as several left side columns or several upper rows.

38 strings (sets) of signs were produced, because there are 38 letters in the churchslavic alphabet. Every set contains 100 members (elements, samples) of a sign "a" or "b" and so on.

By using the statistical methods the most frequent (common) characteristics of all the samples belonging to the string are defined, keeping those characteristics as a "master letter". Additionally, this "master letter" represents a template in the process of recognizing other letters.

III. LABYRINTH METHOD

By analyzing the style of every particular letter in the churchslavic alphabet the decision criteria for every sign about what letter it is determined. Although each scriptor has his own writing style, there are some rules - the main characteristics of the letter that every scriptor has to respect. For example, the letter "i" will always be with high percentage in black pixels, no matter who wrote it. The most dominant value in geometry analyses of letters is the first criteria in the decision system of the character recognition treated. By the geometry analyses the expectation is to obtain the set of unordered values. To make an order, the values in a range determined during the work should be "normalized" (Fig.9). The H-high is a starting position. All the obtained values are transformed, expressed as a number of pixels, in a system measure that is expressed in a multiple of number 12. The h-high value is either 12 (or a number in a ratio (proportion) of 12, like: 6, 24, 36, ...144 ...) so as to find a proper coefficient in other measurements.

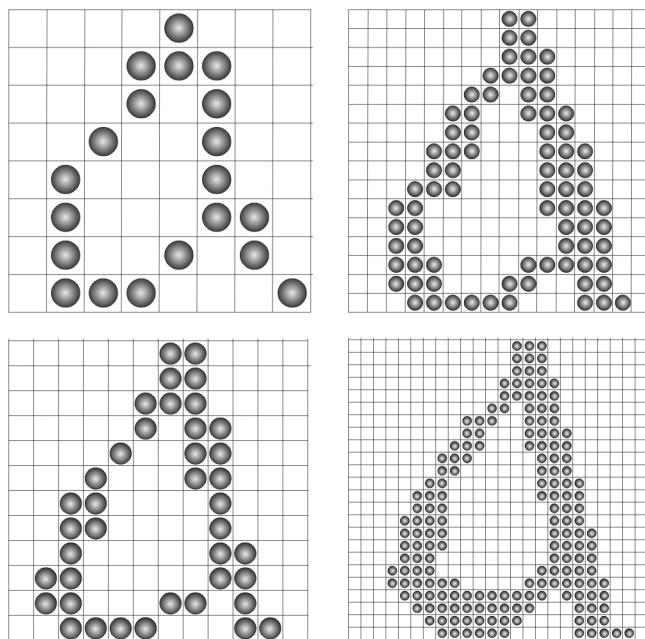


Fig. 9. "Normalizing" the values in a range

The "Labyrinth" method is conceived as a method that can be continually upgraded and refined during its extended use and testing. It is a system of filters and elimination that, based on the total set of 38 characters in every filter, are decreasing the possible number of correct answers in the subset containing smaller number of members-letters. Every query, every criteria represents one filter determining whether the sign is accepted or rejected from the membership in that set. The sign-letter can in the same time be a member of several subsets. When the tests are made by all the criteria, the result is obtained as a combination or the frequency of "membership" of the particular letter in different subsets. If a situation arises when two or more letters have the same frequency in the subsets by all criteria, the two or more results are recommended.

This is a method that is operating by the matrices and is mostly using a mathematical logic and use of logical operators NOT, OR, AND, XOR, i.e. its graphical application in the value tables of truth. Mathematically, the logical operators are shown in the table by the values of 0 and 1. The graphic interpretation of those values represents a black or white dot (Fig.10 - First column), i.e. the presence or absence of a pixel.

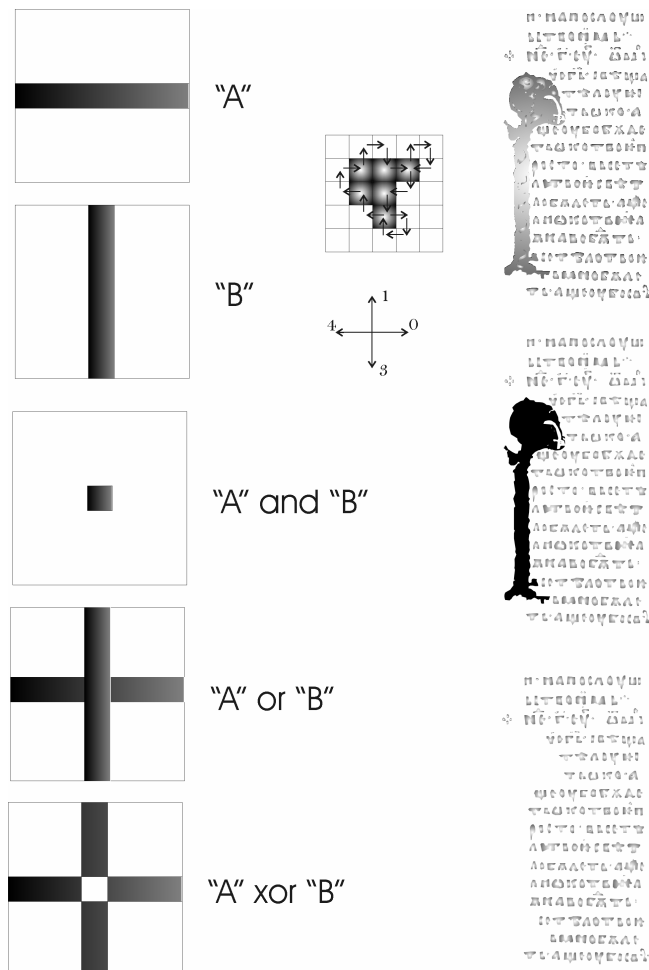


Fig. 10. First column - Graphic result of logical operators application. Second column - An example of contour tracing and an outline of a code . Third column - An outline of the initial elimination.

The programming method of tracing the contour is applied for the sake of limiting (fencing off) the white area from the black area. The graphic interpretation of this method for one approximate black-and-white fenced area is given in the outline (Fig.10 - Second column), together with the selected codes for the direction, the course of movement around the contour.

When one contour will be rimed, the presence of "black dots" in it is counted. Then, the function of filling or replenishment is used. By this function we are sure that all the dots (pixels) that are inside the contour are black. Its recount is done again. If the result (the number of black pixels) in the first and in the second case do not differ dramatically, then there is no presence of white areas or holes in the letter. If there is a real numeric difference (larger than the predefined value of tolerance) in the first and second counting, then we diagnose a presence of white area (hole) in the letter. Besides the counting, we can also make the presence of white holes in the letter by the combined use of logical operators and by filling the contour.

We also use the contour tracing and filling in the graphic elimination of the initial signs of the manuscript, which are not subject of our domain of work (Fig.10 – Third column).

Because the work is in the initial preparation phase, it is hardly possible to talk about satisfactory recognition results. An initial schemata of decision making based on some of the mentioned criteria, can give satisfactory results for one experimentally created font; but for the real world manuscripts additional branching and broadening of this schemata is necessary.

IV. CONCLUSION

A schematic construction suitable for recognizing the text written by standalone churchslavic characters is presented in this paper. If there are some accidentally connected letters they are treated as initial letter or some decorative element(s) and are extracted as a sign that is not a subject of interest. As a starting point, the effort is to concentrate the research toward 38 main letters in the churchslavic alphabet. The supposition in this work is that there is a clean paper with no slant written letters, because other authors are treating those additional problems. The second groups of additional problems, like upper letters, upper signs, decorative elements or stylized crosses are not currently treated.

REFERENCES

- [1] Chen Tsun Chuang and Lin Yu Tseng, "A Heuristic Algorithm for the Recognition of Printed Chinese Characters", IEEE Transactions on Systems, Man and Cybernetics, vol. 25, no. 4, pp. 710-717, 1995.
- [2] David Earls, "Designing Typefaces", Roto Vision, 2002.
- [3] Vera Stojcevska Antic, "Makedonska srednovekovna knizevnost", Skopje, 1997.
- [4] <http://www.dataid.com/aboutocr.htm>