Visual analysis of topological features for character recognition

Klekovska Mimoza*, Martinovska Bande Cveta**

* Faculty of Architecture, AUE FON University, Skopje, R. North Macedonia
mimiklek@yahoo.com
** Computer Science Faculty, University Goce Delcev, Stip, R. North Macedonia
cveta.martinovska@ugd.edu.mk

DOI: 10.29322/IJSRP.13.06.2023.p13810
http://dx.doi.org/10.29322/IJSRP.13.06.2023.p13810

Paper Received Date: 10th April 2023
Paper Acceptance Date: 21st May 2023
Paper Publication Date: 6th June 2023

Abstract- The text presented hereinafter is describing the process of extraction of several characteristics required for recognition of Old Slavic Letters. The analysis is conducted over old Slavic scripts written with Cyrillic alphabet named "Constitutional Script". It is a unique style of writing. The contour of the graphemes has been observed from several aspects: structural, graphic–aesthetical, geometrical and topological. As a result of such analysis, we have got semiotic style descriptions, harmonic proportions as well as characteristics of the letters that are used in the recognition process. The actual letter recognition is mainly founded on pixel recognition in particular segments like spots or blemishes, presence of vertical and horizontal lines as well as determination of compactness, complexity of the contour line and symmetry of the entire letter.

Index Terms- Feature extraction, Fuzzy Decision, Handwritten Character Recognition, OCR/ICR Application,

I. INTRODUCTION

The main goal of this research is set to building up a system for digitalization of manuscripts written with Old Slavic Alphabet acquired from Macedonian churches and monasteries along with the materials from Slavic institutes, libraries and archives. The presently available commercial OCR software cannot be used due to the specific characteristics of the letters in Cyrillic alphabet. Owing to the duality expressed in the nature of the Church Slavic Alphabet which is basically handwritten, but looks as printed one, a combination of methods for manuscript/handwritten and printed texts is to be used in the analysis. ABBYY FineReader software does not recognize Old Slavic languages besides having support for several other ancient languages. Aside the Latin and Greek, Old Slavic language is the fourth encoded language in Europe: gothic in the 4th century, Anglo – Saxon in 7-9th century, Old Germanic in the 7th and Old Slavic in the 9th [1]. Documents analyzed for the purpose of this article are actually written for the purpose of church services, using the Constitutional Script.

While different writing styles for the purposes of church – related documents are introduced in Europa, being influenced by the new styles in art emerging at the time (Roman, Gothic, Baroque and Rococo), Slavic graphemes have held their writing style unaffected by such trends.

The documents analyzed in this paper were written in Constitutional Script style for church-liturgical purposes. This writing style is similar with printed text using 'all caps' letters, and the contour lines of the letters can be drawn separately. The letters are well shaped, upright, stand-alone and decoratively designed. In this kind of old manuscripts there is no distinction between uppercase and lowercase letters. A specific feature of Church Slavic manuscripts is the joined writing or the so-called Scripta Continua. It does not mean co-joint line between the letters, but it means no blank space between the words, so using electronic dictionary book is helpless.

A standardized database as a computer font for Old Slavic Cyrillic letters does not exist. There are several variants of the computer sets of letters, depending of the linguistic redaction of the manuscript. The manuscripts used in the project for the recognition of Old Slavic Cyrillic characters are taken from the anthology of written monuments prepared by Macedonian linguists [2] and an electronic review published by Russian linguists [3]. Most of the digitized manuscripts used in this work were written for ecclesiastical purposes.

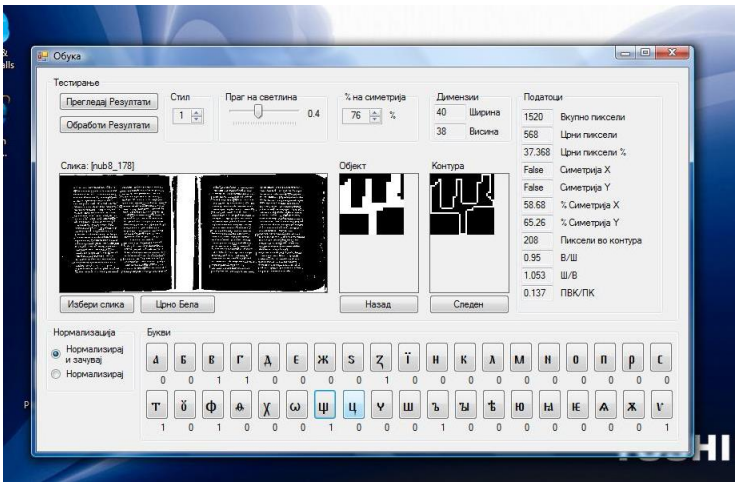
In the last decades, a number of handwriting recognition systems have been proposed [4], and some of them are used in commercial products [5, 6]. Different approaches have been used for letter recognition, such as fuzzy logic [7, 8], neural networks [9], and genetic algorithms [10].
The goal of feature selection in letter recognition methodologies is to find the discriminative features among entire set of letters that maximize the efficiency of the classifiers. In our approach, the features which are the most robust to in-sample variation of letters are obtained by testing a large number of letter samples from different manuscripts written by 'different hands'.
There are usually two types of feature selection methods in the books: filter and wrapper [11]. Filtering algorithms use some prior knowledge to select the best features and are independent of the classification algorithm or its error criteria. Wrapper algorithms are less general because feature selection is related to the learning algorithm and are less suitable for classification problems with a large number of features.

II. LETTER PROCESSING METHODS

The total letter recognition process consists of several steps: pre-processing, segmentation, feature extraction and selection, classification and post-processing [12]. In general, preprocessing methods include: image binarization, normalization, noise reduction, detection and correction of skew, estimation and removal of slant.
The role of pre-processing is to separate the letters, each letter as a separate image and to prepare the images for further steps. This step also defines the representations of the letters in the form of: outlines, skeletons, binary images or grayscale images.
The feature extraction methods depend on the representation of the segmented letters. Different types of feature extraction methods are used in letter recognition systems, such as: statistical, structural and global transformations.
Various techniques and methods for feature extraction are found in the literature, such as: histograms, contour profiles, median axis transformation (MAT) or "thinning" [13]. With thinning (MAT) the shape is transformed into a line drawing or skeleton.
Representing the shape of a letter with separate features or primitives is done in order to find the most adequate subset of features that will be most effective in describing the shape so that the learned prototypes can be used for recognizing new, similar to learned forms.
Post-processing is related to word recognition. Linguistic context can reduce ambiguity in word and letter recognition, but this method is not very useful here, due to 'Scripta Continua' rule.

III. PREPARING THE LETTERS FOR FEATURE EXTRACTION PROCEDURE

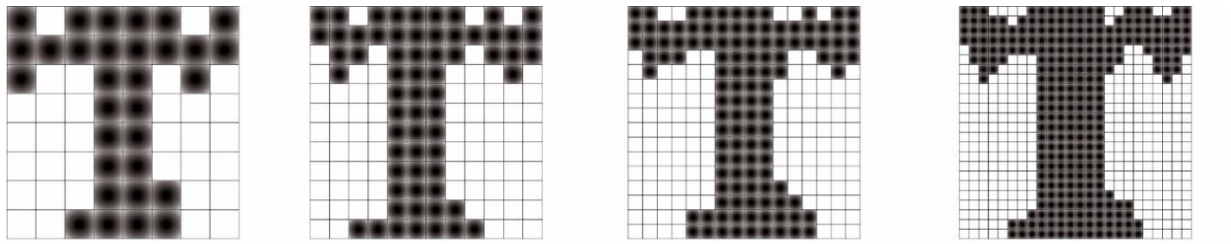
An application was created for visual monitoring and checking the feature extraction process of letters that uses scanned documents written with Old Slavic letters as input. It helps to compare the human versus computer reasoning of the dynamics of the process. The screen is used as a medium to check and monitor the correctness of human's thoughts. The first - Training module of the application, forms a reference folder, where the characters (letters) are placed together with all the selected characteristics for each character.



Screenshot from the TRAINING module

A. Preparing the letters to be process in the application

At the beginning, a contour extraction method is applied to the letters, which determines the minimum matrix (rectangle) of pixels that preserves the shape of the letter. The scaling is with the same coefficient in both directions (***x*** and ***y***) to preserve the original aspect ratio of the matrix, that is, to obtain a matrix "similar" to the original one. This process is called letter 'normalization' in dimensions suitable for computer application



"Normalization" of the matrix does not change the basic stylistic features (proportion, blackness)

The sets of normalized displayed samples (one set, for one letter), standardization of the values of each characteristic is performed, whereby those characteristics that drastically deviate from the mean values are rejected. From the remaining samples, reference samples are selected for each class, in four variants: the minimum (graphically, the intersection of the set of pixels), the maximum (graphically, the union of the set of pixels) and 2 additional states (fuzzy intersection and fuzzy union).

B. Marginal values maximum and minimum

The experience shows that the limit values (maximum and minimum) in a very small number of cases give a true picture of the representative of the class. Only in very practiced hand (person) these values are recognizable. So, two additional fuzzy-states (30% and 60%) or signs are included, which in an imaginary three-dimensional graph of the given position from the matrix have 30% and 60% presence of pixels at a particular position, which would correspond to a partial intersection, i.e., a partial union.

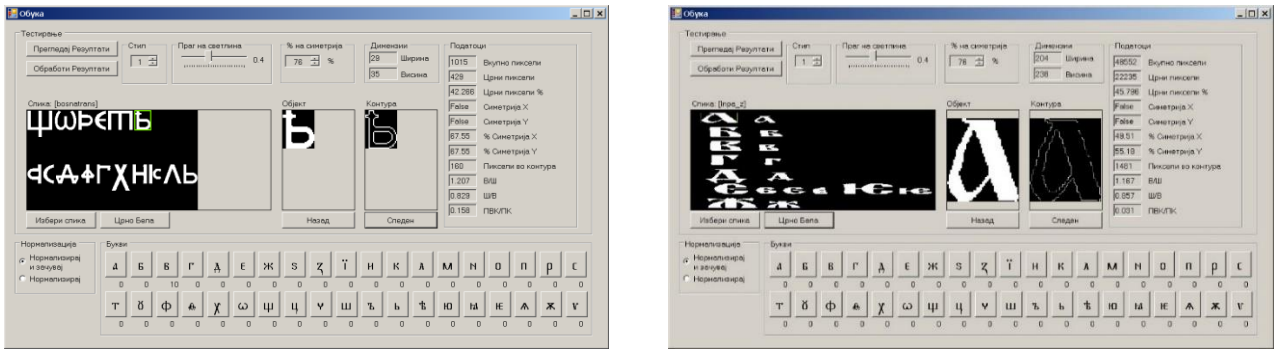
			
Intersection (classic)	*Intersection 30%*	*Union 60%*	*Union (classic)*

Application extracted reference graphemes from the training letter samples

Graphic records (bitmapped images) are stored in a separate folder (database) as a visual control for the course of the analyses. Several statistical and structural characteristics are distinguished from the contour of each letter, such as: dimension of the smallest surface in which the contour can be inscribed; proportion (harmonic scale) of contour matrix dimensions; percentage of blackness of the image (light-dark ratio); contour symmetry (does it exist and to what extent is it achieved); determining the ratio between the length of the contour line and the area encapsuled within; then: monolithicity, presence and position of a group of pixels (spots), vertical and horizontal lines, etc.

IV. FEATURE EXTRACTION PROCEDURE -TEST MODULE

A variety of characteristics have been considered for the process of characters/letters recognition such as size of the bitmap picture, proportion of height and width, harmonic proportion of height and width, proportion of black – white pixels and the ratio of black to total number of pixels, the percentage of pixels symmetrical to the x and y axes, the length of the outer contour expressed in pixels, and the ratio of the length of the outer contour to the occupied area.



Test of feature extraction

Some features that are not significant for the recognition process, but they have a role in distinguishing the stylistic features of the letter [14]. For example, the ratio of height to width for a particular code-sign is used to determine whether a letter belongs to one of the basic styles: Roman or Gothic. The length of the contour line for a code-sign also, provides information about the curvature of the letter and the presence of decorative elements.

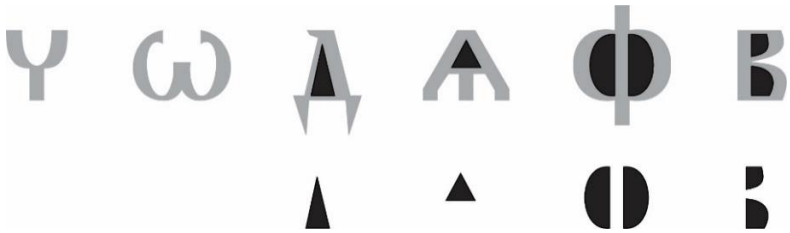
A. Contour line

Simple, plain forms have smaller length of contour lines per standard surface matrix compared to decorative letters. The percentage of black pixels vs total matrix area shows if the letter/character is light, regular or emphasized (bolded).

Contour line for the decorative letter

Characteristics of the letter are derived from bitmaps created as prototypes for each letter. Prototypes are obtained by application of logical operators on the samples from the digitalized original scriptures. Graphical interpretation of logical operators provides fast results and is visually easer for tracking. Contour line might be drawn at once and it makes a compact letter. Otherwise, the letter is airy or double airy, means one-hole or two-holes letter. The term of compactness is used for characterization the letters drawn with a single contour line and no holes. Should the letter have one or more holes, it is considered as aerated. For example, the G letter is compact while the A letter is aerated, having one hole.

Contour line might be drawn symmetrically in *x* or *y* axis, too. The compactness and symmetry as characteristic features are determined from a contour line of the letter's bitmap in a non–segmented bitmap form. Some other features need to segment the bitmap surface in several segments.

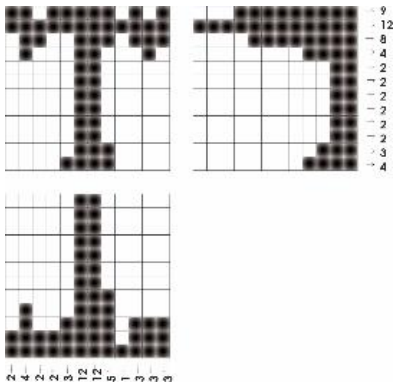


Compact, airy and double airy letter

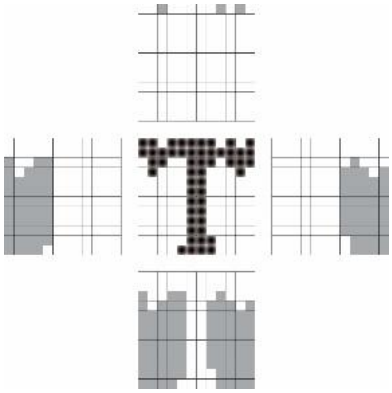
Compact letter G and one-hole aerated letter A

B. Histograms and contour profiles

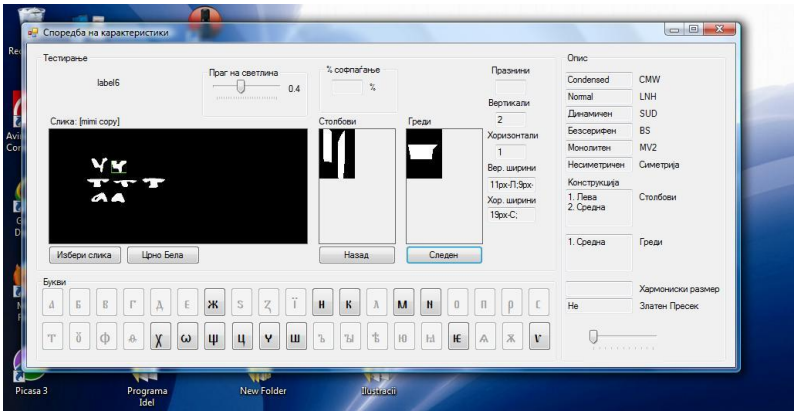
Designed histograms are actually replications from two-dimensional to one-dimensional function, having values that represent cumulative of pixels along one direction, horizontal or vertical. Contour profiles represent the remaining part (residues), from the matrix border to the contour line of the letter. Contour profiles attack the letter from four sides. Histograms and contour lines are illustrated on the Figure only in perpendicular directions. Variations of diagonal histograms and contour profiles are not taken into consideration, because they have not provided reliable results.



Histograms



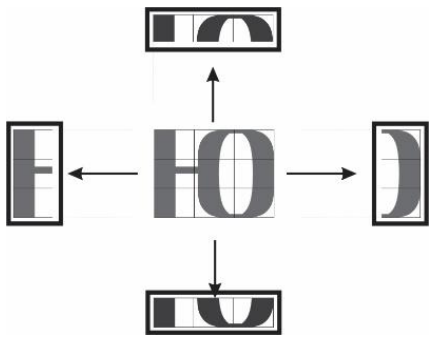
Contour profiles



CHARACTERISTICS work module screen

C. Spot analyses

Another approach for derivation of vital characteristics has been inspired by the methods for letter recognition that use histograms and contour profiles [12]. They are normalized in a way that the height is set to 24 units while the width is determined proportionally, hence maintaining the original shape. Each bitmap is segmented by two horizontal and two vertical cross sections. These segments are topologically surveyed to determine the number of dots/spots and presence of vertical and horizontal lines in each sector/segment. Some of the characteristics extracted from the normalized black-white bitmaps with height and width set proportionally, actually divide the full set of letters into several subsets, such as letters with emphasized left, right or both verticals, subsets with emphasized horizontal line on top, on the bottom or both, middle horizontal or vertical line and so on. This method we named Spot analysis method.

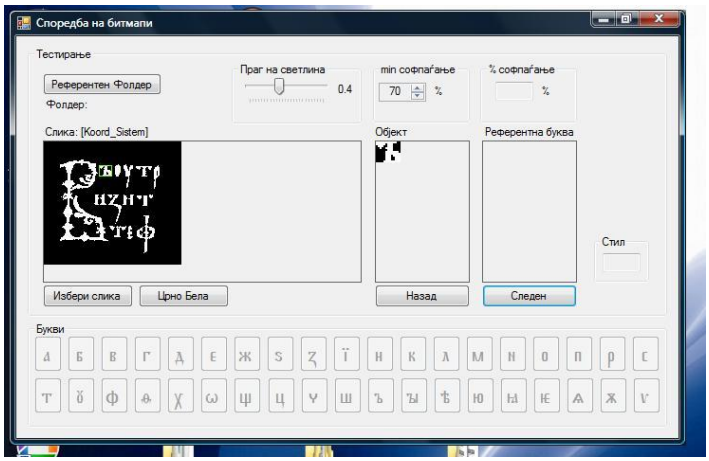


Intersections of a letter to apply spot analysis method

The number of dots/spots (one, two or three) in the outer segments is yet another characteristic of importance for the classification of the letters.

V. OCR - BITMAP MODULE

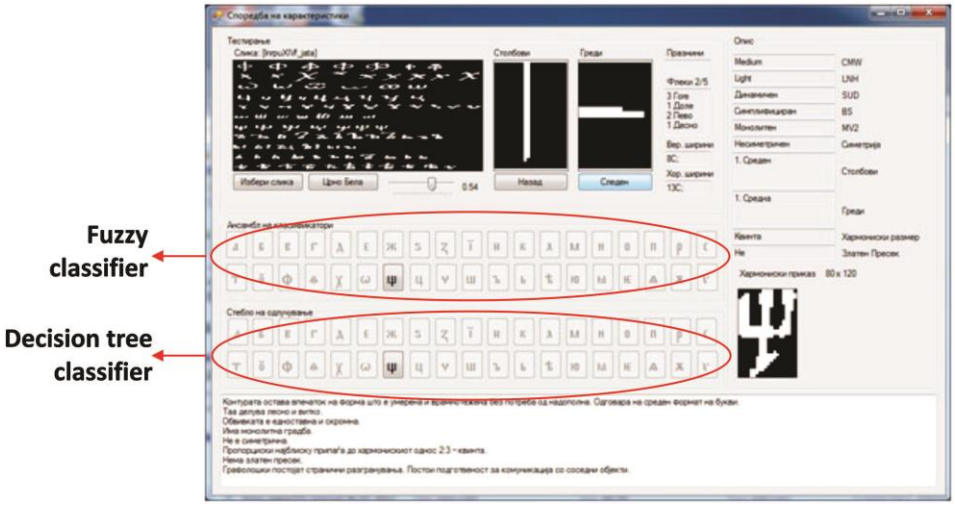
In the second module (Bitmap comparison) a set of test letters is compared to the reference forms and the reference letters with the highest match are found. The module works as an OCR system and serves to check the correctness of the base of reference models.



Screenshot of the BITMAP module

VI. ICR – CLASSIFIERS MODULE

The third module works as an ICR (Intelligent Character Recognition) without comparisons. Here, a comparison is made as a competition of two methods for the classification of Old Slavic Cyrillic characters. The first classification is based on a decision tree method, and the second uses a fuzzy-classifier. Both methods can apply the same set of feature selections extracted from character bitmaps in previous steps. A semiotic comment or graphological interpretation for the shape of the sign is also given in the lower part of the screen, based on selected features.



Screenshot of the ICR - Classifiers module

VII. CONCLUSION

Feature selection described in this paper may provide successful results if the following preconditions are assumed:

- Church Slovenian manuscripts are written by a Constitutional Script in which the letters are upright, free-standing, beautifully formed. Manuscripts written in Scripta Continua style are not applicable for recognition;
- A digital substrate can be derived from the manuscripts that is sufficiently readable in printed or screen form in black and white technique;

- The visual reasoning approach converts or dissects letters to basic geometric proportional elements, which as logic is applicable to practically any organized system of signs. This type of logic reasoning might be used for recognition to any other alphabet, i.e., language or symbolism (traffic, geographical, meteorological symbols...)
- There is a need to expand the efforts made to read the Latin alphabet and by that example other letters should receive the same treatment. Such efforts are also recorded among the Chinese, although the number of symbols in their writing is measured by thousands, not tens of characters;
- The visual reasoning approach is reduced to basic geometric proportional elements, which as logic is applicable to practically any organized system of signs, that is, to any other alphabet, i.e. language or symbolism (traffic, geographical, meteorological symbols...)

REFERENCES

[1] Antic, V.: Macedonian Medieval Literature. Institute for Macedonian Literature. Skopje, Macedonia (1997) (in Macedonian)

[2] Velev, I., Makarijoska, L., Crvenkovska, E.: Macedonian Monuments with Glagolitic and Cyrillic Handwriting. 2nd August, Stip, Macedonia (2008) (in Macedonian)

[3] Russian Review of Cyrillic Manuscripts: http://xlt.narod.ru/pg/alpha.html

[4] Eastwood, B., Jennings, A., Harvey, A.: A Feature Based Neural Network Segmenter for Handwritten Words. International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'97). Australia, pp. 286-290 (1997)

[5] Ntzios, K., Gatos, B., Pratikakis, I., Perantonis, S.J.: An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach. Int. Journal on Document Analysis and Recognition, vol. 9, no. 2, pp.179-192. Springer, Heidelberg (2007)

[6] Chen, C. H., Curtins, J.: Word Recognition in a Segmentation-free Approach to OCR. Second International Conference on Document Analysis and Recognition (ICDAR'93), pp. 573-576 (2003)

[7] A. Malaviya, and L. Peters, "Fuzzy handwritten description language: FOHDEL", Pattern Recognition, vol. 33,2000, pp. 119-131.

[8] R. Ranawana, V. Palade, and GEMDC Bandara, "An efficient fuzzy method for handwritten character recognition", M.Gh. Negoita et al. (eds.), KES 2004, LNAI 3214, Springer-Verlag, 2004, pp.698-707.

[9] G. Zhang, "Neural networks for classification: ASurvey". IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, vol. 30(4), 2000, pp.451-462.

[10] G. Kim, and S. Kim, "Feature selection using genetic algorithms for handwritten character recognition", In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Nijmegen: International Unipen Foundation, 2000, pp 103-112.

[11] D. Calmakov, and B. Younes, "Feature selection for pattern recognition", Publisher Informa, Skopje, Macedonia, 2002.

[12] Cheriet, M., Kharma, N., Liu, C. L., Suen, C. Y.: Character Recognition Systems, A Guide for Students and Practioners. John Wiley and Sons, New Jersey (2007)

[13] K. Ntzios, B. Gatos, I. Pratikakis and S.J. Perantonis: "An Old Greek Handwritten OCR System based on an Efficient Segmentation-free Approach", Int. Journal on Document Analysis and Recognition, Vol. 9 No. 2, pp.179-192, , 2007

[14] Klekovska, M., Nedelkovski, I., Stojcevska-Antic, V., Mihajlov, D.: Automatic Letter Style Recognition of Church Slavic Manuscripts. Proc. of 44th Int. Scientific Conf. on Information, Communication and Energy Systems and Technologies. Veliko Tarnovo, Bulgaria, pp.221-224 (2009)

AUTHORS

First Author – Klekovska Mimoza, associate professor, Faculty of Architecture, AUE FON Universizitet, Skopje, R. North Macedonia, mimiklek@yahoo.com
Second Author – Martinovska Bande Cveta, full-time professor, Computer Science Faculty, University Goce Delcev, Stip, R. North Macedonia, cveta.martinovska@ugd.edu.mk

Correspondence Author – Klekovska Mimoza, mimiklek@yahoo.com, cveta.martinovska@ugd.edu.mk, + 389 (0)70 273275.