Recognition features for Old Slavic letters: Macedonian versus Bosnian alphabet

Cveta Martinovska Bande[*], Mimoza Klekovska[**], Igor Nedelkovski[**], Dragan Kaevski[***]

[*] Faculty of Computer Science, University Goce Delcev, Stip, Macedonia
[**] Faculty of Technical Sciences, University St. Kliment Ohridski, Bitola, Macedonia
[***] Faculty of Elecrical Engineering and Information Technology, University St. Cyril and Methodius, Skopje, Macedonia

Abstract- This paper compares recognition methods for two Old Slavic Cyrillic alphabets: Macedonian and Bosnian. Two novel methodologies for recognition of Old Macedonian letters are already implemented and experimentally tested by calculating their recognition accuracy and precision. The first method is based on a decision tree classifier realized by a set of rules and the second one is based on a fuzzy classifier. To enhance the performance of the decision tree classifier the extracted rules are corrected according to their accuracy and coverage. The fuzzy classifier consists of rules constructed by fuzzy aggregation of letter features. Both classifiers use the same set of discriminative features, such as number and position of spots in outer segments, presence and position of horizontal and vertical lines and holes, compactness and symmetry. We argue that the same feature set can be used for recognition of Old Bosnian letters. Moreover, due to the similarity of the graphemes and fewer letters we expect better efficiency of the recognition system for Bosnian letters.

Index Terms- classifier, decision tree, fuzzy logic, historical manuscripts, precision and recall.

I. Introduction

The work presented in this paper is performed as part of a project for digitalization of historical collections found in Macedonian monasteries, institutes and archives that originate from different periods.

Commercial character recognition systems are not applicable for Old Slavic Cyrillic handwritten manuscripts due to their specific properties. For example, ABBYY FineReader supports several old languages but does not provide support for Old Slavic languages. The vast majority of the historical documents have low quality hence some pre-processing is necessary to enhance their readability [1].

In the field of character recognition there are two main research directions that lead to online [2] and offline [3, 4] recognition systems. Recognition of handwritten cursive letters is a complex procedure due to the inconsistent and conjoined manner of writing. The recognition of digits [5] is usually simpler compared to letter recognition.

In the last two decades a number of handwritten recognition systems have been proposed [6] and some of them are used in commercial products [7, 8]. Different approaches to letter recognition have been reported, such as fuzzy logic [9, 10], neural networks [11] and genetic algorithms [12].

Several steps have to be performed in the process of letter recognition, like pre-processing, segmentation, feature extraction and selection, classification, and post-processing [13]. Generally, pre-processing methods include image binarization, normalization, noise reduction, detection and correction of skew, estimation and removal of slant. There are two segmentation approaches [14]: explicit where the image is decomposed into separate letters and implicit where whole words are recognized without decomposition. A survey of methods for selection of discriminative features, based on wrapper and filter approaches is presented in [15]. Wrapper algorithms are less general since the feature selection is related to the learning algorithm, while filter algorithms execute faster and are more appropriate for classification problems with large number of features.

Methods for feature extraction depend on the representations of the segmented characters, like contours, skeletons, binary images or gray level images. In letter recognition systems different types of methods for feature extraction are used, such as statistical, structural and global transformations. The goal of feature selection in recognition methodologies is to find the most relevant features that maximize the efficiency of the classifier. Post-processing is related to word recognition. The language context can reduce the ambiguity in recognition of words and letters.

Our previous work [16] presents experimental results of novel recognition methodologies applied to Old Macedonian Cyrillic manuscripts. Two classifiers are created using the same set of features. The first classifier is based on a decision tree and the second one uses fuzzy techniques. The features that are selected as the most discriminative in the conducted experiments are: number and position of spots in the outer segments, presence and position of vertical and horizontal lines and holes, compactness and symmetry. The efficiency of the classifiers is tested experimentally and their performance is compared through the measures recognition accuracy and precision.

In this paper we analyze the potential of these methodologies for recognition of Old Bosnian letters. The proposed recognition techniques are applicable only to manuscripts written in Cyrillic alphabet with Constitutional script and some variants (not very slanted) of Semi-Constitutional script.

Set of features for recognition of Old Macedonian letters is tolerant of variations in different graphemes. We expect that the proposed set of features is appropriate for Old Bosnian letters,

too. There are only three Old Bosnian letters that do not have prototype and therefore are not properly covered with the created recognition system. Letter □ is not present in Macedonian alphabet, while letters ѣ and ѵ have different graphemes. Macedonian grapheme for ѣ is t and for ѵ is v. In Macedonian alphabet letters 5 or 3 might be used as a substitute of letter □.

The next section presents the properties of Old Slavic Cyrillic letters in the context of methodologies used for their digitalization and recognition. The main differences between Old Macedonian and Bosnian alphabets are outlined. After that, the pre-processing techniques applied to letter images are described followed with the process of feature extraction and selection. Applicable techniques for pre-processing of these historical documents include converting the row data to black and white bitmaps, normalization and segmentation.

In the remainder of the paper the decision trees for classification of Old Macedonian and Bosnian letters are presented. Then, fuzzy classifier and applied fuzzy aggregation methods are described, followed by section containing experimental results and evaluation of the efficiency of the proposed classifiers. Concluding remarks demonstrate the possible application of the presented methodologies for recognition of Old Bosnian letters.

II. OLD SLAVIC CYRILLIC SCRIPT

The work described in this paper focuses on the structural forms of Old Slavic Cyrillic graphemes. Many modern Cyrillic alphabets descend from this script. Documents that are analyzed in this work pertain to the liturgical manuscripts written on parchment with Constitutional script. Other types of scripts, Semi-constitutional and Cursive script were mainly used for legal and commercial documents.

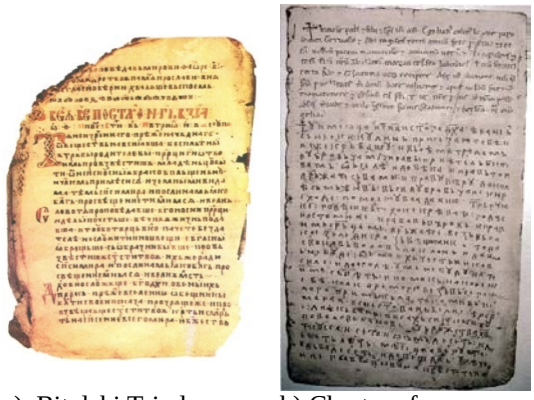
Constitutional script is handwritten but looks like printed text. The letters are well shaped, upright, separated and decoratively designed. In old manuscripts there is no distinction between uppercase and lowercase letters.

It is hard to ascertain the period of occurrence of the manuscripts because the graphemes were not affected by the style changes. While Latin letters undergone dramatic changes in their appearance, influenced by Romanic, Gothic or Baroque style, Old Slavic letters were only slightly changed in the period from 10th to 18th century. The manner of writing used in the Old Slavic Cyrillic manuscripts is called scripta continua because of the merged writing of the words.

A. Macedonian vs Bosnian alphabet

The recognition methodologies used in this paper are created for the Macedonian recension of manuscripts written with Constitutional script. The alphabet consists of 38 letters. The full set of Cyrillic letters in different alphabets consists of 43 letters. Figure 1a shows excerpt from Bitolski Triod, taken from the anthology of written monuments prepared by Macedonian linguists [17]. The manuscript is framed chronologically between 11th and 12th century.

The Bosnian Cyrillic alphabet, known as bosančica, is used in Bosnia from 10th to 20th century. This alphabet slightly differs from other Cyrillic alphabets under the influence of Glagolitic and Latin alphabets. The orthographical and phonetic systems are simplified and some Old Slavic letters are abandoned. Additionally, there are letters in Old Bosnian Cyrillic alphabet that have several different graphemes.



a) Bitolski Triod b) Charter of Bosnian Ban Kulin

Figure 1. Old Slavic Cyrillic Manuscripts

Figure 1b shows excerpt from Old Bosnian manuscript from 12th century, charter of Bosnian Ban Kulin, written with Constitutional script. This type of script has been used in Bosnia from 10th to 15th century primarily for legal documents. Constitutional script has also been used for marble made tombstone inscriptions, like Humac tablet from 10th or 11th century.

While this script was used for church purposes in Macedonia, it was used in public and legal documents in. Hence, in Bosnia Constitutional script was earlier replaced by Semi-Constitutional and Cursive script.

Table 1 shows parallel representation of Macedonian and Bosnian old alphabets. From the comparative analysis of the two alphabets is evident that 22 letters have same graphemes and pronunciation. Grapheme V has different appearance in these two redactions. Two letters (A, B) have slight differences in their graphemes, while 8 letters have completely different graphemes. Several graphemes ((`, 1, 5, 3) of the Macedonian alphabet are not used in the Bosnian alphabet.

In fact only 2 letter prototypes (□ and ѣ) have to be additionally defined in order to apply the created recognition system to Old Bosnian manuscripts.

The software support for recognition of Old Macedonian Cyrillic alphabet has already been made and tested and it is our intention to test this recognition system on other Old Cyrillic alphabets including Bosnian.

III. PRE-PROCESSING OF THE MANUSCRIPTS

In this section we present generally used recognition techniques that are applicable for Old Slavic manuscripts written in Cyrillic alphabet with Constitutional script.

Table 1. Graphemes of Macedonian and Bosnian old alphabets

	Macedonian grapheme	Macedonian phonetic sign	Latin sign	Bosnian grapheme
1	Aa	Az	A	a
2	b	Buki	B	b
3	v	Vedi	V	v
4	g	Glagoli	G	g
5	d	Dobro	D	d
6	e	Este	E	e
7	/	DZivejte	DZ	
8	\	ZCelo	ZC	6dz
9	z	Zemlja	Z	z
10	J	Idze		
11	i	I	I	i
12	k	Kako	K	k
13	l	LJudi	L	l
14	m	Mislete	M	m
15	n	Nasha	N	n
16	o	On	O	o
17	p	Pokoi	P	p
18	r	Raci	R	r
19	s	Slovo	S	s
20	t	Tverdo	T	t
21	U	Ouk	U	U
22	f	Fert	F	f
23	H	Ksita		
24	h	Hara	H	h
25	w	Omega	W	w
26	]	SHta	SH	Q
27	c	Ci	C	c
28	;	CHerv	CH	□
29	[	SHa	SH	
30	q	Jor(jeri)	Half-voice	.
31	Q	Jata	KJ	Y
32	2	Ju	J	x
33	`	Ja		
34	1	Je		
35	5	Jen-big		
36	3	Jon-small		
37	u	Idzhica		
38			GJ	□

Existing commercial computer software for translating scanned documents into machine-editable texts cannot be used because of the specific characteristics of these letters.

The accuracy of recognition techniques applied to old manuscripts is affected by a number of factors, such as noise due to scanner quality or degradation as a result of parchment aging and fading of ink. A number of pre-processing techniques are often used to enhance the readability of the documents. As previously mentioned generally used pre-processing techniques are binarization, noise reduction, normalization, detection and correction of skew, estimation and removal of slant.

One of the properties of the Constitutional script is merged writing of the words (Fig. 2) or scripta continua in Latin. Hence, implicit segmentation methods for letter extraction are more relevant than explicit ones that are used for extraction of words.

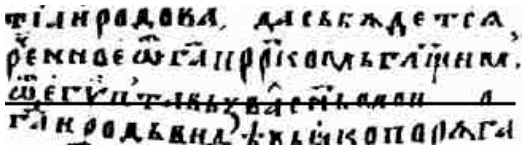


Figure 2. Merged writing of words.
Base line is not well determined.

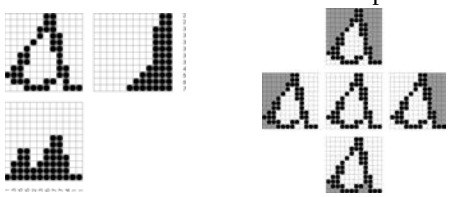
Another characteristic of the manuscripts written with Constitutional script is that the base line is not well determined, as shown by the straight line on Fig. 2. Hence, skew detection and correction are not applicable, which additionally complicates the process of letter extraction. Upright writing which is typical for Constitutional script eliminates the need for slant detection and correction. However, the created recognition system is robust for small slant angles. Thus, this recognition system can also be used for Semi-Constitutional script that has slight slant angles.

Several pre-processing steps are performed to the letter images, such as converting to black and white bitmaps, normalization and extracting letter contours using contour following function.

During the segmentation procedure vertical projections (histograms) can serve to separate adjacent letters and to detect multiple horizontal lines (Fig. 3).

Similar to histograms contour profiles (image residues) count the number of pixels or distance between bounding box and the edge of the letter. Contour profiles describe the external shapes of the letters and are used in topological analysis to determine the existence of certain features.

During the normalization process letter width is determined proportionally to the height. The values are transformed as multiples of number 12 because Old Slavic script is uncial script.



a) b)

Figure 3. a) Histograms b) Contour profiles

IV. FEATURE EXTRACTION AND SELECTION

The role of the pre-processing is to separate the letters and prepare the images for further steps. This step also defines letter representations in the form of contours, skeletons, binary images or gray level images. Feature extraction methods depend on the representations of the segmented letters.

The purpose of the analysis performed on the Old Slavic Cyrillic manuscripts is to determine style descriptions, harmonic proportions, structural and statistical features that are relevant for recognition. Figure 4 shows application module that determines features specific to a certain letter. Features used in the recognition process should be insensitive to variations and distortions within the samples of the same letter.

The objective of feature selection in the recognition methodologies is to find the most discriminative features that maximize the efficiency of the classifiers. In our approach the most resilient features to different variations within the samples

are obtained by testing the features on a number of letter samples.

As previously mentioned there are two types of feature selection methods: filter and wrapper. Filter 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 since the feature selection is related to the learning algorithm and are less appropriate for classification problems with large number of features.

General methods for feature extraction, like moments, contour profiles, histograms and Hough transformation are not applicable for extracting these types of features. These methods use large set of samples contrary to our recognition system based on evidence of prof and more suitable for if-then rules and fuzzy classification.

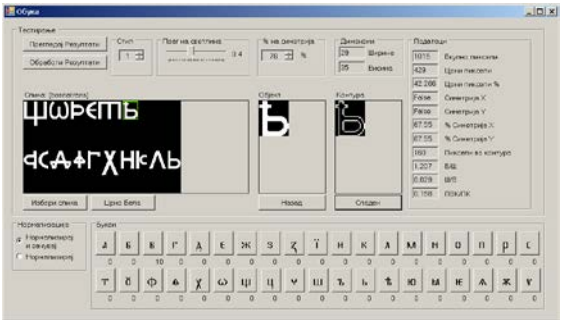


Figure 4. Feature extraction

Additionally, using sophisticated features saves processing resources and eliminates the need of large training sets necessary for Bayesian classifiers, neural networks or support vector machines. Neural networks are not capable to extract these features because of the presence of noise in the letter images.

Standardized database for Old Slavic Cyrillic letters does not exist. The manuscripts that are used in the project for recognition of Old Church Slavic Cyrillic characters are taken from the anthology of written monuments prepared by Macedonian linguists [17] and electronic review published by Russian linguists [18]. The majority of the digitalized manuscripts used in this work were written for church purposes.

A. Discriminative Features

Letter bitmaps are examined in order to extract features that are used in the process of classification. Letter prototypes are built as combinations of features. Each prototype might have several variations due to the inconsistency in writing manner. The prototype variations are apparent in Table 2, were uncertain features are denoted with light (yellow) fields.

Initially, 22 features were considered as discriminative for creating letter prototypes, such as dimensions of the bitmap image, height vs. width ratio, harmonic relationship of the height and the width, black vs. white pixels ratio and black vs. total number of pixels, percentage of pixels symmetric to x and y axes, the length of the outer contour expressed in pixels, outer contour length vs. area occupied.

Some of the features that are relevant for printed texts are not applicable for handwritten documents. So the resilient set of features is restricted to features presented in Table 2. First three features are related to the appearance of the whole letter, thus the letter is compact, or airy (with one hole) or double airy (with two holes), as shown in Fig. 5.

Figure 5. Comapct, airy and double airy letter

Second group of features also refers to the whole letter bitmap and is connected to the letter symmetry. The symmetry is observed either to x or y axis. Criteria for symmetry are softened with a threshold values. Thus, letters which are registered as symmetrical by human visual system are considered as symmetrical.

Several features are related to letter segments as topological parts of bitmaps. The segments are formed by two vertical and two horizontal intersections (Fig. 6).

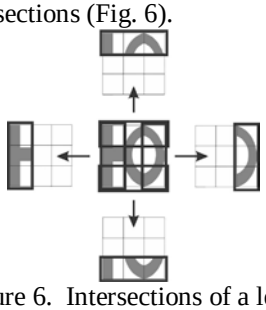


Figure 6. Intersections of a letter

Vertical lines or columns and horizontal lines or beams are optional features. It is considered that letter has vertical or horizontal line if more than 5/8 of the segment is filled. Position of the line is also important for the process of classification.

We found that one of the most discriminative features is the number of spots in the four outer segments. The number of spots can vary from one to three and these four groups of features are obligatory.

Table 3 presents the discriminative features of Bosnian alphabet. Again uncertain features are denoted with light (yellow) fields.

V. RECOGNITION METHODOLOGIES

The system uses sophisticated features that incorporate previous knowledge based on artistic, constructive and analytical principles. Harmonic ratios of letters, as part of artistic analysis, bring evidence for the positioning of vertical and horizontal intersections of letter bitmaps. For example, the human visual system recognizes that perceived image is line if it is more than 5/8 fulfilled.

Different variations of constructive elements are used to form the descriptions of letters, such as number and position of vertical and horizontal lines together with letter compactness or presence of holes.

Table 2. Features of Macedonian graphemes

	Compactness	Holes	x-symmetry	y-symmetry	Column left	Column central	Column right	Row up	Row middle	Row down	Spots left	Spots up	Spots right	Spots down
A a		1									1	1	1	1,2
b		1			*			*		+	1	1	2	1
v		2	*		*			+		+	1	1	2	1
g	*				*			*			1	1	1	1
d		1		+					+		1	1	1	1,2
e	*		+		*			+	+	+	1	1	3	1
/	*		+	*		*					2	3	2	3
\	*		+								2	1	2	1
z	*							+			2,3	1	2	1
J	*		*	*	+	*	+	+	*	+	1	1	1	1
i	*		+	+	*		*		+		1	2	1	2
k	*		+		*		+				1	2	2	2
l	*			+			+				1	1	1	2
m	*			+	*		*				1	2	1	1,2,3
n	*		+	+	*		*				1	2	1	2
o		1	*	*	+		+	+		+	1	1	1	1
p	*			*	*		*	*			1	1	1	2
r		1			*						1	1	1	1
s	*		*		*			+		+	1	1	2	1
t	*			*		*		*			1	1	1	1
U		1		*	+		+		+	+	1,2	2	1,2	1
f		2	*	*		*					1	1	1	1
H		2		*							1	1	1	1,2,3
h	*										1,2	1,2	1,2	1,2
w	*			*	+		+			+	1	1,2,3	1	1,2
]	*			*	*	*	*				1	3	1	1
c	*				+		*				1	2	1	1
;	*			*		+					1	2	1	1
[	*			*	*	*	*			*	1	3	1	1
q		1			*				+	*	1	1	1	1
Q		1			+	*		+	+	+	1	1	1	1
2		1	*		*	+	+				1	2	1	2
`		1			*					+	1	2	1	2,3
1	*		+		*			+	+	+	1	2	3	2
5	+	1		*		+		+			1	1	1,2	3
3	+	1		*		+			+		1,2	1	1	3
u	*						+				1,2	2	1	1

Table 3. Features of Bosnian graphemes

	Compactness	Holes	x-symmetry	y-symmetry	Column left	Column central	Column right	Row up	Row middle	Row down	Spots left	Spots up	Spots right	Spots down
a		1					+				1	1	1	1
b		1			+			*		+	1	1	2	1
v		1	+	+	*		*	*		*	1	1	1	1
g	*				*			*			1	1	1	1
d		1		+					+		1	1	1	1,2
e	*		+		+			+	+	+	1	1	3	1
	*		+	*		*					2	3	2	3

6d	*		+								2	1	2	1
z	*							+			1,2,3	1	1,2	1
□	*			+				+	+		1,2	1	1,2	1
i	*			+	*		*		+		1	2	1	2
ƙ	*										1	2	2	2
l	*			+							1	1	1	2
m	*			+							1	2	1	1,2,3
n	*								*		1	2	1	2
o		1	*	*	+		+	+		+	1	1	1	1
p	*			*	*		*	*			1	1	1	2
r		1			+						1	1	1	1
s	*		*		*			+		+	1	1	2	1
t	*			+	*	*	*	*			1	1	1	3
U		1		*	+		+		+	+	1,2	2	1,2	1
f		2		*		*					1	1	1	1
x		1	*		*	+	+				1	2	1	2
h	*										1,2	1,2	1,2	1,2
w	*			*	+		+			+	1	1,2,3	1	1,2
Q	*			*	*	*	*				1	3	1	1
c	*				+		+				1	2	1	1
□	*			+							1	2	1	1
	*			*	*	*	*			*	1	3	1	1
.		1			*					*	1	1	1	1
Y		1						+	+	+	1	1	1	1

Features that are used in the recognition process are built in the system thus eliminating the need for learning with large training set.

A. Decision Trees

The first step for building recognition system for Old Slavic Cyrillic letters was the decision tree represented in Figure 7. Letter prototypes are constructed in the form of if-then rules extracted from this classifier.

The feature set used for building the classifier consists of compactness, symmetry, number and position of holes, spots, vertical and horizontal lines. The features that are most discriminative for letter classification are positioned in the nodes closer to the root of the decision tree created using the measure information gain. As Figure 7 shows, three spots in the lower segment is very rear feature not present in Latin alphabets but expressive for Cyrillic alphabets.

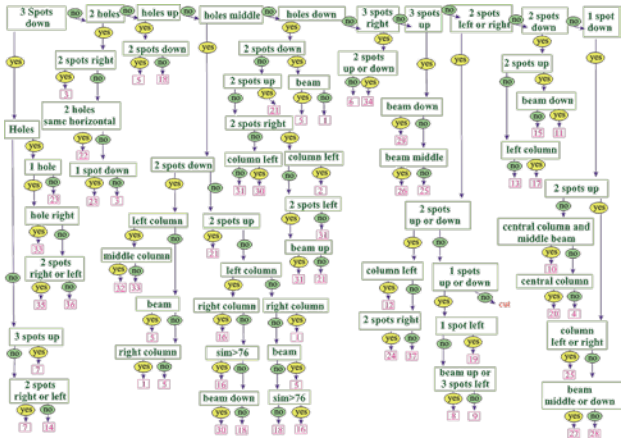


Figure 7. Decision tree for Macedonian alphabet

Comparing Old Macedonian and Bosnian Cyrillic alphabets is evident that this feature is present in letter / (in both alphabets), t (in Bosnian), 5 and 3 (in Macedonian). The decision tree for Bosnian alphabet is shown in Figure 8.

The proposed recognition system is flexibly designed allowing several prototypes for the same letter, to cope with the imperfection of the bitmap images of handwritten letters. Several prototypes are created for letters that do not possess expressive features that will distinguish them from the others and for the letters that do not have consistency in the manner of writing. For example:

-letter r is described by the following features: one hole (in the upper or middle segment), left vertical line (presence or absence)

-letter U is represented by the following descriptions: one hole (in the lower or middle segment), one or two spots (in the upper, left or right segment).

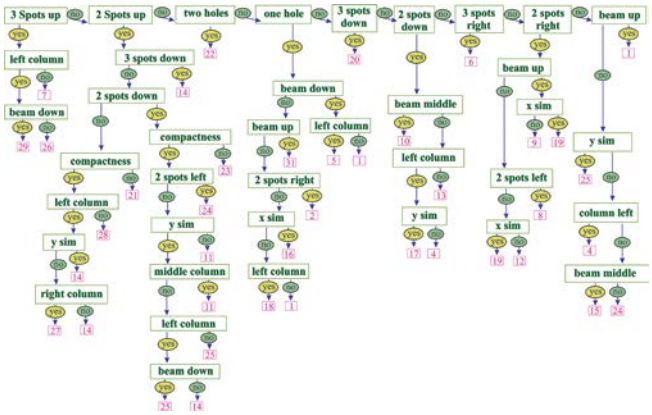


Figure 8. Decision tree for Bosnian alphabet

Decision tree classifiers for Old Macedonian and Bosnian alphabets are realized through the set of if-then rules. The measures accuracy and coverage are computed for the extracted rules. The performance of the decision tree classifiers is enhanced by recomposing (combining or pruning the conditional part) and preordering of the rules.

C. Fuzzy Classifier

Fuzzy classifier consists of fuzzy linguistic rules that form the classification rule base. The recognition system uses fuzzy aggregation of the letter features to construct letter prototypes.

Human visual system recognizes letters even when they possess some vagueness or imprecision. The recognition of various patterns is done by selecting discriminative features which are combined to identify the given letter. We introduce fuzzy methods to cope with an imprecision which is a result of writing manners of different scriptors and other variations that arise from differences in historical periods or regions where manuscripts originate.

The classifier must determine the most likely identity for a given letter. This is achieved by applying fuzzy rules to a letter presented as an input and computing membership values for letters. The first step in letter recognition phase, is calculating the membership functions for every feature of the letter. Then, the membership of a particular letter to all letter prototypes is evaluated, using the following formula

$$\mu_n = \frac{\sum_{c=1}^{C} w_c \cdot \mu_c}{C} \quad n = 1,..N. \qquad (1)$$

The last step is selection of the prototype with the highest compatibility, or selection of more than one class when there are several most similar classes that pass the threshold

$$\mu_A = \bigcup_{n=1}^{N} \mu_n \,. \qquad (2)$$

Applied fuzzy techniques

In the feature extraction phase we distinguish two types of features: global features extracted from the non-segmented letters, like symmetry and compactness and local features, such as vertical and horizontal lines, spots and holes. The letters are represented by a combination of these basic features.

Weight coefficients are used to express the importance of the features for a particular letter in the process of classification. Higher weights are assigned to the features that are rare and more discriminative.

Let μ_G denote the membership function that aggregates the fuzzy information $(\mu_1, \mu_2, \ldots, \mu_N)$ for the character features

$$\mu_G = \mathrm{Agg}(\mu_1, \mu_2, \ldots, \mu_N) \qquad (3)$$

where Agg is a fuzzy aggregation operator. Let w1, w2, …,wN represent weights associated with fuzzy sets A1, A2, …, AN. The weighted median aggregation is computed by the following formula [19]:

$$\mathrm{Med}(a_1, \ldots, a_N, w_1, \ldots, w_N) = \left(\sum_{i=1}^{N} (w_i a_i)^{\alpha}\right)^{\frac{1}{\alpha}} \qquad (4)$$

where $\sum_{i=1}^{N} w_i = 1$, α is a real non-zero number with values between $\max(a_1, a_2, \ldots, a_N)$ and $\min(a_1, a_2, \ldots, a_N)$.

Weighted median aggregation operator (4) is used to create a matrix that contains associated features from the complete set of features I. With this step structural features (vertical and horizontal lines, spots and holes) are combined with location related features (position, orientation). Overall measure of feature importance is computed using a union operator proposed by Yager [20]

$$U(a_1, a_2, \ldots, a_N) = \min\left\{1, \left(\sum_{i=1}^{N} (a_i)^{\alpha}\right)^{\frac{1}{\alpha}}\right\} \qquad (5)$$

where α is a real non-zero number and the value that can be obtained as a result of the union ranges between 1 and $\min(a_1, a_2, \ldots, a_N)$.

Fuzzy aggregation techniques are used to compute the overall measure and to arrange the features by the degree of importance.

Extracting the most relevant features from the total set and ordering the features by the degree of importance is essential to achieve high efficiency of the letter recognition system.

Calculation of aggregated features

Letters are segmented in S segments. In our approach 6 segments are obtained by two vertical and two horizontal intersections. The total set of features is divided in two categories. The first category G comprises global features like symmetry and compactness that are extracted from the non-segmented characters. The second category L contains local

features, such as structural features.

Associations of features are formed by combining the structural features with their position or size. The number of calculations that have to be performed during the recognition process is reduced by creating the associations of features with the operators (4) and (5). The importance of the features for the recognition process is represented using weight matrix.

Let $\bar{I}_s$ denote the L x S matrix of local features extracted from S segments:

$$\bar{I}_s = \{i_{sl} | l = [1, L]\}, s = [1, S] \quad (6)$$

and $\bar{I}_g$ denote the global feature set.

Combined feature vectors Vs for each segment are obtained associating the local features of each segment with position and size related features:

$$\bar{V}_s = \{\bar{v}_{sc} | i = [1, C]\}, s = [1, S] \quad (7)$$

where C is the number of combined features for each segment. Then, the set of combined feature vectors is extended with the global features. Using estimation function E only the combined features that are relevant for the recognition process are extracted:

$$\bar{v}_{sc} = E(\bar{I}_{sj}). \quad (8)$$

The number of combined features C is less than or equal to the number of combination of L+G choose P elements, where P is the number of relevant features. The weight matrix $\bar{W}_s$ related to the feature importance for the process of recognition is computed through statistic evaluation of the prototype samples:

$$\bar{W}_s = \{\bar{w}_{s1}, \dots, \bar{w}_{sC}\} \quad (9)$$

The feature vectors for each segment are computed using the weighted median aggregation by formula (4):

$$\bar{\mu}_s = \mathrm{Med}(\bar{w}_{sc}, \bar{v}_{sc}) \quad (10)$$

The most important features from the previously generated feature list are selected using Yager's union connective:

$$\{\mu_p\} = \min\{1, (\sum \mu_{ps})\} \quad (11)$$

Finally, from the computed subset $\{\mu_p\}$ of meaningful features fuzzy descriptions of the letters are constructed.

Assigning non-membership functions to letter features

Non-membership functions are introduced for intuitionistic fuzzy sets [21, 22]. The intuitionistic fuzzy set S in U is defined as

$$S = \{(x, \mu_S(x), \nu_S(x)) | x \in U\} \quad (12)$$

where $\mu_S: U \to [0,1]$ and $\nu_S: U \to [0,1]$ represent the degree of membership and the degree of non-membership of the element x to the set S. In our fuzzy recognition system non-membership functions are used to overcome the ambiguity during the process of letter classification. Several Old Slavic letters contain a grapheme of other letter as a subset (Fig. 9). Thus the features of a certain letter form a proper subset of the features of other letter. For example, features of a letter g are proper subset of the features of b, p and e. Moreover, features of g partially overlap with the features of vV, r R, s and o.

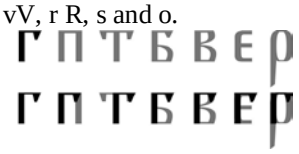


Figure 9. Letters that contain a grapheme of other letter

These ideas are implemented in the fuzzy classifier using non-membership functions. The absence of a feature is characterized with a value of non-membership function. As Table 4 shows marked features are used to compute the values of non-membership functions. When the features of a letter (g) are proper subset of the features of other letter (b) this measure is used to make distinction between the letters.

Table 4. Letters for which non-membership functions are computed

ID number	Grapheme	Name of the letter	Middle right	Up right	Up middle	Up left	Middle left	Down left	Down middle	Down right
4	g	Glagoli	■							■
5	d	Dobro		■		■				
11	i	I			■				■	
12	k	Kako	■							
13	l	LJudi				■				
18	r	Raci								■
20	t	Tverdo						■		■
24	h	Hara							■	
28	;	Cherv			■			■		■
30	q	Jer		■						
36	3	Jon-small		■		■				

Without non-membership functions class g is always fired together with class b when letter b is present at the input of the recognition system. Thus, the value of non-membership function is used to eliminate the misclassification of certain letters.

VI. EXPERIMENTAL RESULTS

In this section we present the experimental results obtained with the fuzzy classifier applied to Macedonian recension of Old Slavic Cyrillic manuscripts. Figure 10 shows a screenshot of the application used in the experiments with both classifiers: decision tree and fuzzy classifier.

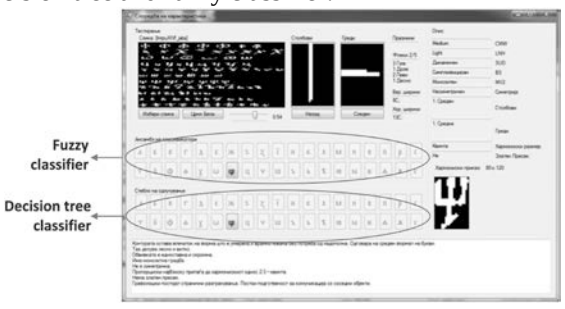


Figure 10. Screenshot of the application: results of the simultaneous classification of letter] using two different classifiers

We argue that recognition of Old Bosnian letters will be comparable or even more accurate than recognition of Old Macedonian letters. Bosnian alphabet consists of 33 letters but 3 of them have same graphemes and only their phonetic interpretation is different. From the whole set of Bosnian letters

22 have same graphemes as Macedonian but we expect differences in the recognition accuracy and recall because the rest of the sets are different. Recognition efficiency of the fuzzy classifier depends on the overlapping features of the whole set of letters.

We expect less overlapping between the prototypes of letter graphemes in Bosnian documents because of the smaller number of 'confusing' graphemes. In this recognition system digraphs (Đl and Đn) can be discovered with post-processing using the context.

A. Recognition Accuracy and Recall

Several measures are computed to evaluate the precision and recall of the classifiers. The sensitivity or recall of the classifier denotes the probability that a letter of a current class is correctly classified. This measure is computed according to the formula

$$R = \frac{TP}{TP+FN} \quad (13)$$

where TP (True Positive) is the number of correctly labeled letters that belong to the current class and FN (False Negative) is the number of letters that belong to the current class incorrectly labeled as belonging to other classes. Precision of the classifier can be interpreted as probability of a letter classified in the current class actually to belong to that class and is defined as

$$P = \frac{TP}{TP+FP} \quad (14)$$

where FP (False Positive) is the number of letters incorrectly labeled as belonging to the current class and TP has the same meaning as defined in (13).

Both metrics recall and precision have to be combined in order to estimate the efficiency of the classifier. For that purpose measure F1 is computed as harmonic mean of precision and recall. F1 is calculated according to the following formula

$$F1 = \frac{2RP}{R+P} = \frac{2\,xTP}{2\,xTP+FP+FN}. \quad (15)$$

The values of these measures for decision tree (DTC) and fuzzy classifier (FC) obtained for Macedonian manuscripts are presented in Table 5. Both approaches, discussed in this work achieve acceptable results in classifying the Old Slavic letters with almost equal percent of recognized letters.

Table 5. Precision and recall of the classifiers

Letters	FC recall	FC precision	DTC recall	DTC precision
Aa	0.71	0.59	0.5	0.63
b	0.75	0.79	0.67	1
v	0.89	1	0.88	0.64
g	0.56	0.83	1	0.5
d	1	0.9	0.75	0.75
e	0.43	1	0.56	1
/	0.63	0.83	0.8	1
\	0.9	0.86	0.25	0.33
z	0.25	1	1	0.42
J	0.5	0.5	0.71	1
i	0.8	0.92	1	0.82
k	0.8	0.8	1	0.2
l	0.94	0.85	0.89	1
m	1	1	0.8	1
n	0.95	0.82	1	0.82
o	1	0.6	0.86	0.86
p	0.94	0.88	1	1
r	0.67	0.8	0.5	0.75
s	0.6	0.5	1	0.82
t	1	0.79	0.86	0.58
U	0.33	1	1	0.8
f	0.8	0.89	1	0.86
H	0.5	0.6	0.43	0.75
h	1	0.75	0.56	0.71
w	0.33	1	0.67	0.2
]	1	0.69	1	1
c	0.77	0.59	0.43	0.75
;	0.83	0.71	0.43	1
[	0.82	1	0.89	1
q	0.67	0.57	0.71	0.71
Q	0.5	1	0.79	0.65
2	0.77	0.91	0.89	0.57
`	0.4	1	0.33	1
1	1	1	0.4	1
5	0.33	0.5	0.5	0.33
3	0.25	1	0.4	1
u	0.63	0.71	0.44	0.67
Total	**0.71**	**0.82**	**0.73**	**0.76**

The decision tree classifier recognizes the Old Slavic letters with an average recall of 0.73, average precision of 0.76 and F1 measure of 0.74.

The fuzzy classifier recognizes Old Slavic letters with an average recall of 0.71, average precision of 0.82 and an overall average measure of precision and recall F1 of 0.76. Recognition accuracy of the fuzzy classifier is improved after incorporating intuitionistic fuzzy measures.

Moreover, the proposed methodology reduced the misclassifications that occurred between similar letters l and p, o and r, g and t.

The experiments reported in this paper use the same database of letters from Old Church Slavic manuscripts originating from 12th till 16th century.

VII. Conclusion

Both approaches, decision tree and fuzzy classifier, discussed in this work achieve acceptable results for Old Slavic manuscripts written with Macedonian alphabet with almost equal percent of recognized letters. Proposed classifiers are tested only on Macedonian manuscripts because we do not have access to the original Bosnian manuscripts.

The same set of features is used to build decision tree for Old Bosnian letters. We argue that the selected features and the proposed methodology are also appropriate for recognition of Old Bosnian Cyrillic letters. Based on previously elaborated evidence we expect that fuzzy classifier is applicable to Bosnian manuscripts with comparable results.

To further improve the recognition capabilities of the system we plan to include linguistic rules based on discourse analysis. There are several studies that elaborate orthography of Old Cyrillic script.

REFERENCES

[1] Gatos B, Pratikakis I, Perantonis S. "Locating text in historical collection manuscripts."LNAI, 2004; 3025: 476-485.

[2] Namboodiri A, Jain A, "Online handwritten script recognition." IEEE Trans. PAMI 2004; 26 (1): 124-130.

[3] Vinciarelli A. "A Survey on off-line cursive word recognition", Pattern Recognition 2002; 35: 1433-1446.

[4] Arica N, Yarman-Vural FT. "An Overview of Character Recognition Focused on Off-Line Handwriting."IEEE Trans. Systems, Man, and Cybernetics 2001; 31 (2): 216-233.

[5] Gader P, Forester B, Ganzberger M, et al."Recognition of handwritten digits using template and model matching." Pattern Recognition 1991; 5(24): 421-431.

[6] Bortolozzi F, Britto J, Oliveira L,et al. "Recent advances in handwriting recognition." Document Analysis 2005; 1-31.

[7] D'Amato D, Kuebert E, Lawson A. "Results from a performance evaluation of handwritten address recognition systems for the United States Postal Service.", Proc. Int. Workshop on Frontiers in Handwriting Recognition, Amsterdam, 2000;189–198.

[8] Gorski N, Anisimov V, Augustin E, et al. "A2iA check reader: a family of bank check recognition systems." Proc. Int. Conf. on Document Analysis and Recognition 1999; 1: 523–526.

[9] Malaviya A, Peters L."Fuzzy handwritten description language: FOHDEL." Pattern Recognition 2000; 33: 119-131.

[10] Ranawana R, Palade V, Bandara GEMDC."An efficient fuzzy method for handwritten character recognition." LNAI 2004; 3214: 698-707.

[11] Zhang G. "Neural networks for classification: Asurvey." IEEE Trans. on Systems, Man, and Cybernetics 2000; 30 (4): 451-462.

[12] Kim G, Kim S."Feature selection using genetic algorithms for handwritten character recognition." Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition, Nijmegen: International Unipen Foundation, 2000; 103-112.

[13] Cheriet M, Kharma N, Liu C, et al. Character recognition systems, A guide for students and practioners. Wiley and sons, 2007.

[14] Casey RG, and Lecolinet E. "A survey of methods and strategies in character segmentation." IEEE Trans. on Pattern Analysis and Machine Intelligence 1996; 18 (7): 690–706.

[15] Trier ØD, Jain AK, Taxt T. "Feature extraction methods for character recognition - A survey." Pattern Recognition 1996; 29 (4): 641–662

[16] Klekovska M, Martinovska C, Nedelkovski I, Kaevski D."Comparison of models for recognition of Old Slavic Characters." ICT Innovations 2012, AISC 207, Springer-Verlag 2013; 129-139.

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

[18] Russian Review of Cyrillic Manuscripts

[19] http://xlt.narod.ru/pg/alpha.html

[20] Malaviya A, Peters L. "Extracting meaningful handwriting features with fuzzy aggregation method." Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, 1995; 841-844.

[21] Yager R. "On the representation of multi-agent aggregation using fuzzy logic."Cybernetics and Systems 1990; 21: 575-590.

[22] Atanassov K."Intuitionistic fuzzy sets." Fuzzy sets and Systems 1986; 20(1):87-96.

[23] Atanassov K, Szmidt E, Kacprzyk J. "On some ways of determining membership and non-membership functions characterizing intuitionistic fuzzy sets." 6th Int. Workshop on IFSs, Banska Bystrica, Slovakia 2010; NIFS 16(4):26-30.

AUTHORS

Cveta Martinovska Bande – Full professor, Faculty of Computer Science, University Goce Delcev, Stip, Macedonia
cveta.martinovska@ugd.edu.mk

Mimoza Klekovska Ph.D. – Faculty of Technical Sciences, University St. Kliment Ohridski, Bitola, Macedonia
mimiklek@yahoo.com

Igor Nedelkovski – Full professor, Faculty of Technical Sciences, University St. Kliment Ohridski, Bitola, Macedonia
igor.nedelkovski@uklo.edu.mk

Dragan Kaevski, MSc - Faculty of Elecrical Engineering and Information Technology, University St. Cyril and Methodius, Skopje, Macedonia
d.kaevski@gmail.com

Correspondence author
Cveta Martinovska Bande,
Faculty of Computer Science,
University Goce Delcev,
Stip,Macedonia,
E-mail:cveta.martinovska@ugd.edu.mk