Feature Selection for Classification of Old Slavic Letters

C. M. Bande*, M. Klekovska, I. Nedelkovski**, D. Kaevski*****

**Computer Science Faculty, University Goce Delcev, Macedonia (Tel: +389 32 550 103; e-mail: cveta.martinovska@ugd.edu.mk)*
***Faculty of Technical Sciences,University St. Kliment Ohridski, Macedonia (e-mail:mimiklek@yahoo.com, igor.nedelkovski@uklo.edu.mk)*
**** Faculty of Electrical Engineering and Information Technology, University St. Cyril and Methodius, Macedonia (e-mail: d.kaevski@gmail.com)*

Abstract: This paper describes methodology for extracting discriminative features for fuzzy classification of Old Slavic characters. Recognition process is based on structural and statistical features, such as number and position of spots in outer segments, presence and position of horizontal and vertical lines and holes, compactness and symmetry. Preprocessing is divided into the following steps: conversion to black and white bitmaps, normalization, contour extraction and segmentation. Features are extracted from contour profiles, histograms and character intersections. C4.5 decision trees are used for feature selection. The same feature set is appropriate for different Old Slavic Cyrillic alphabets because of the similarity of their graphemes. The classification accuracy and precision are tested on Old Macedonian manuscripts and the decision trees are created for two alphabets Macedonian and Bosnian. The main advantage of the proposed method is saving processing resources and eliminating the need of large training sets necessary for Bayesian classifiers or neural networks.

Keywords: classifiers, decision tree, fuzzy logic, character recognition, precision and recall, historical manuscripts.

1. INTRODUCTION

In the field of letter recognition there are two main research directions that lead to online (Namboodiri and Jain, 2004) and offline (Vinciarelli, 2002; Arica and Yarman-Vural, 2001) recognition systems. Recognition of handwritten cursive letters is a complex procedure due to the inconsistent and conjoined manner of writing. The recognition of digits (Gader et al., 1991) is usually simpler compared to letter recognition.

In the last two decades a number of handwritten recognition systems are proposed (Bortolozzi et al., 2005) and some of them are used in commercial products (D'Amato et al., 2000; Gorski et al., 2001). Different approaches to letter recognition are reported, such as fuzzy logic (Malaviya and Peters, 2000; Ranawana et al., 2004), neural networks (Zhang, 2000) and genetic algorithms (Kim and Kim, 2000).

Several steps have to be performed in the process of letter recognition, like pre-processing, segmentation, feature extraction and selection, classification, and post-processing (Cheriet et al., 2007). 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 (Casey and Lecolinet, 1996): explicit where the image is decomposed into separate letters and implicit where whole words are recognized without decomposition. Methods for feature extraction depend on the representations of the segmented letters, like contours, skeletons, binary images or grey 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.

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 (Gatos et al., 2004). Applicable techniques for pre-processing of these historical documents include converting the row data to black and white bitmaps, normalization and segmentation.

This work describes a feature selection methodology based on decision tree algorithm for classification of Old Slavic Cyrillic letters. According to the performed analysis, the most discriminative features for these letters are: number and position of spots in the outer segments, presence and position of vertical and horizontal lines and holes, compactness and symmetry. The selected feature set is used for creating fuzzy classifier for recognition of Old Macedonian letters. Also the

potential of this methodology for recognition of other Old Slavic Cyrillic letters is analysed, as for example Old Bosnian letters, because the proposed feature set is tolerant to variations in different graphemes. The efficiency of the classifier is tested experimentally and the performance is measured computing its recognition accuracy and precision.

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.

The next section outlines different approaches to feature selection. Then the properties of Old Slavic Cyrillic letters in the context of methodologies used for their digitalization and recognition are introduced. After that, the process of creating decision trees for classification of Old Macedonian and Bosnian letters is presented. Then, fuzzy classifier and applied fuzzy aggregation methods are described, followed by a section containing experimental results and evaluation of the efficiency of the proposed classifier. Concluding remarks demonstrate the possible application of the presented methodologies for recognition of Old Slavic Cyrillic letters.

2. FEATURE SELECTION APPROACHES

Feature selection problems are investigated for many years (Blum et al., 1997; Kohavi et al., 1997; Ladha and Deepa, 2011; Cateni et al., 2013). Based on the nature of the features there are different approaches for feature selection. For evaluation of the symbolic or nominal features entropy based measures and analysis of the contingency tables are proposed (Rauber et al., 1993). Two methods are used for evaluation of the numerical features. The first method is based on the error rate of the classifier for different feature subsets. The subset with minimal misclassification rate is selected as appropriate feature subset. The second method selects the subset of features for which the overlapping of the classes is minimal measured by the distance between the classes.

The evaluation of all possible feature subsets is computationally expensive problem. Different approaches are proposed to overcome this problem, such as methods based on neural networks (Setiono and Liu, 1997; Lee et al., 2001; Kwak et al., 2002; Verikas and Bacauskiene, 2002), genetic algorithms (Siedlecki and Sklansky, 1989; De Stefano et al., 2014) and search techniques.

In classification systems researchers make a distinction between selecting and extracting features from the original feature set (Rezaee et al.,1999). Extracted features are usually obtained by transformation or association of the original features. They increase the discriminative ability of the classifier especially when the original features are not well distinguishable. The main problem with these combined features is that their physical meaning might be lost.

In many studies a distinction is made between three approaches for feature selection: filter, wrapper and embedded approach (Ladha and Deepa, 2011). In filter approaches feature selection is done as a preprocessing step. Filters are computationally efficient and independent of the learning algorithm. Several methods are used for selecting discriminative features: evaluation of the features based on their correlation with the class, Chi-square approach and information gain (Cardie, 1993).

Wrapper approaches depend on classification accuracy of a learning system for selected features. Some commonly used wrapper strategies are two greedy strategies: sequential forward and backward selection and genetic algorithms (Kohavi et al., 1997).

In embedded approach features are selected in the learning phase during the training of the system. This method is similar with wrapper but requires iterative updates and the selection is based on the performance of the system. For example, Wang et al. (2006) described a neural network model where feature selection is incorporated in the training process while Weston et al. (2000) proposed evaluation of the features which is meaningful only for SVM classifiers.

Feature selection and extraction have also been studied for recognition of letters. A survey of methods for extraction of discriminative features, based on wrapper and filter approaches is presented in (Trier, 1996). 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.

3. 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 analysed 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.

3.1 Macedonian vs. Bosnian alphabet

The recognition methodologies used in this paper are created for the Macedonian manuscripts written with Constitutional script. The alphabet consists of 38 letters. The full set of Cyrillic letters in different alphabets consists of 43 letters. Fig. 1a shows excerpt from Bitolski Triod, taken from the anthology of written monuments prepared by Macedonian linguists (Velev et al., 2008). 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.

a) Bitolski Triod b) Charter of Bosnian Ban Kulin

Fig. 1. Old Slavic Cyrillic Manuscripts.

This alphabet slightly differs from other Cyrillic alphabets under the influence of Glagolitic and Latin alphabets. The orthographic 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.

Fig. 1b shows excerpt from Old Bosnian manuscript from 12th century, called 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 Bosnia. 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 alphabets. 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 (Ð and t) 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.

4. PRE-PROCESSING OF THE MANUSCRIPTS

In this section generally used recognition techniques that are applicable for Old Slavic manuscripts written in Cyrillic alphabet with Constitutional script are presented. Existing commercial computer software for translating scanned documents into machine-editable texts cannot be used because of the specific characteristics of these letters.

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	ž
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	č
29	[	SHa	SH	Š
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	Đ

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.

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.

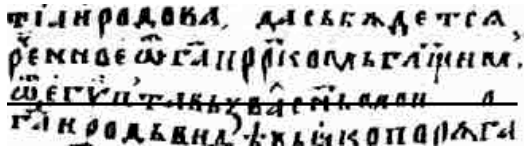
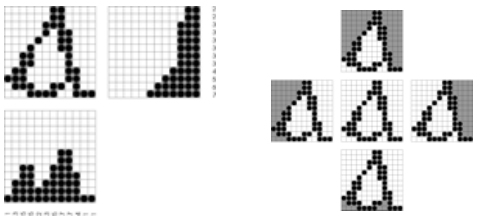


Fig. 2. Merged writing of words. Base line is not well determined.

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).



a) b)

Fig. 3. a) Histograms b) Contour profiles.

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.

5. FEATURE 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. Fig. 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.

Different features are extracted from letter representations, such as dimensions of the 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, and outer contour length vs. area occupied.

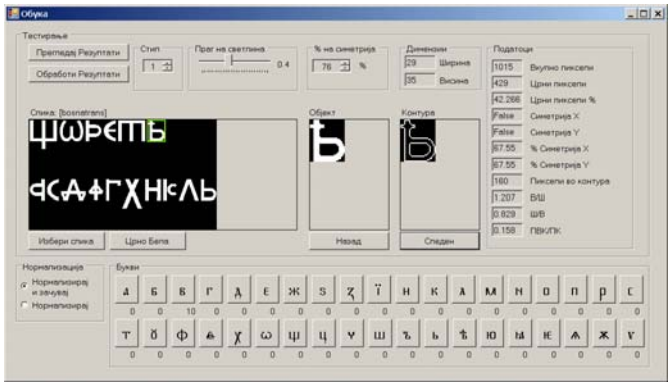


Fig. 4. Feature extraction module.

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. The recognition system for Old Slavic Cyrillic letters is build using 22 features. Decision tree algorithm C4.5 is used to select the most relevant features that are incorporated in the fuzzy classifier.

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 more suitable for if-then rules and fuzzy classification.

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.

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 letters are taken from the anthology of written monuments prepared by Macedonian linguists (Velev et al., 2008) and electronic review published by Russian linguists (Russian Review). The majority of the digitalized manuscripts used in this work were written for church purposes.

5.1 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 the 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, and outer contour length vs. area occupied.

Some of the features that are relevant for printed texts are not applicable for handwritten documents, even though they look like printed 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.

Fig. 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).

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. The human visual system recognizes that perceived image is line if it is more than 5/8 filled. Position of the line is also important for the process of classification.

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

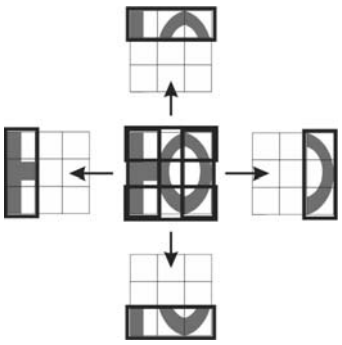


Fig. 6. Intersections of a letter.

Results show 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.

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
ž	*		+	*		*					2	3	2	3
6dz	*		+								2	1	2	1
z	*							+			1,2,3	1	1,2	1
Đ	*			+				+	+		1,2	1	1,2	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	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
č	*			+							1	2	1	1
Š	*			*	*	*	*			*	1	3	1	1
.		1			*					*	1	1	1	1
Y		1						+	+	+	1	1	1	1

5.2 Feature Selection using Decision Trees

The first step for building recognition system for Old Slavic Cyrillic letters was the decision tree represented in Fig. 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. As Fig. 7 shows, three spots in the lower segment is very rear feature not present in Latin alphabets but expressive for Cyrillic alphabets.

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 Fig. 8.

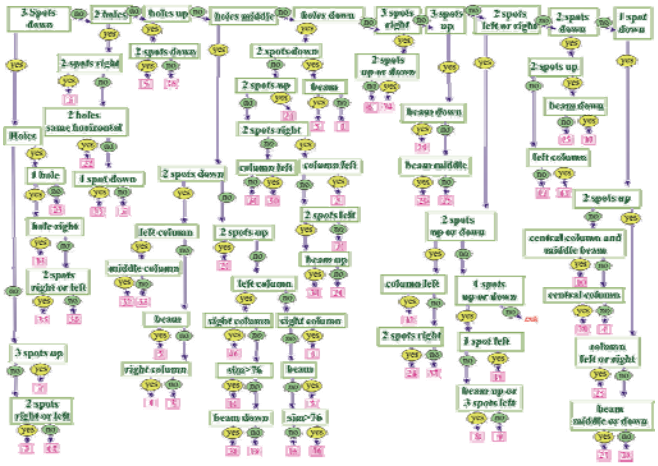


Fig. 7. Decision tree for Macedonian alphabet.

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 which 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).

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 reordering of the rules.

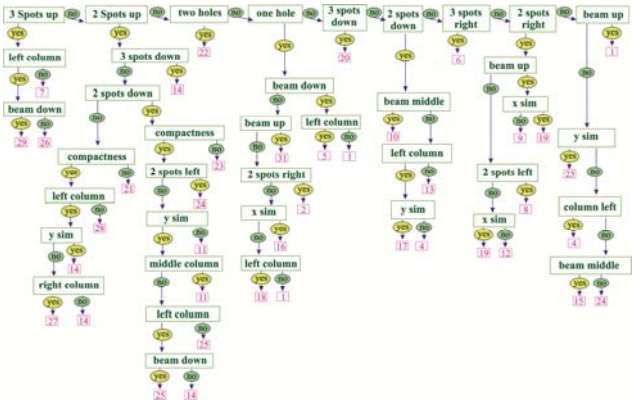


Fig. 8. Decision tree for Bosnian alphabet.

C4.5 algorithm (Tan et al., 2006) is used as a method of feature selection. It is a greedy algorithm that selects the best attribute according to entropy based measure and splits the current set of samples according to values of the test attribute. The process of splitting stops if all samples in a subset belong

to the same class or if the information gained is under a specified threshold.

The measure Gain Ratio is used to select the best attribute when splitting the tree node. This measure is quotient of Information Gain and Split Information. Split Information of an attribute is proportional to the number of values the attribute can take. The measure Information Gain is used in ID3 and has a disadvantage of preferring attributes with many values as splitting attributes. To overcome the problem when Split Information becomes very small C4.5 uses Gain Ratio to select the best attribute from the set of attributes whose Information Gain is above the average value.

First the complete tree is created allowing overfitting which leads to several prototypes for the same letter. Then some of the nodes or subtrees are removed by human expert. This post-pruning is necessary to improve the accuracy of the classifier. The training set of letters consists of imprecise and noisy data so overfitting is an expected effect. Post pruning is performed on the rules obtained from the decision tree to remove some insignificant features from the letter descriptions.

6. FUZZY CLASSIFIER

Fuzzy classifier uses features that are selected with C4.5 decision tree created with samples from a training set. Thus fuzzy classifier can eliminate some of the original attributes which can greatly reduce the size of the training and test data, as well as the running time. In addition the training procedure will take less time due to a fewer features and the classifier would obtain higher generalization ability.

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. Fuzzy methods are introduced 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

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

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

$$\mu_A = \cup_{n=1}^{N} \mu_n \,. \tag{2}$$

6.1 Applied fuzzy techniques

In the feature extraction phase two types of features are distinguished: 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 $(\mu_1, \mu_2, \dots, \mu_N)$ for the letter features

$$\mu_G = \mathrm{Agg}(\mu_1, \mu_2, \dots, \mu_N) \tag{3}$$

where Agg is a fuzzy aggregation operator. Let $w_1, w_2, \dots, w_N$ represent weights associated with fuzzy sets $A_1, A_2, \dots, A_N$. The weighted median aggregation is computed by the following formula (Malaviya and Peters, 1995):

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

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

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 in (Yager, 1990)

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

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 $\min(a_1, a_2, \dots, a_N)$.

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.

6.2 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 (5):

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

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

6.3 Assigning non-membership functions to features

Non-membership functions are introduced for intuitionistic fuzzy sets (Atanassov, 1986; Atanassov et al., 2010). 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, respectively.

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 the other letter. For example, features of the letter G are proper subset of the features of B, P and E. Moreover, features of G partially overlap with the features of V, R, S and O.

Fig. 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.

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		■		■				

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. 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.

7. EXPERIMENTAL RESULTS

In this section the experimental results obtained with the fuzzy classifier applied to Old Macedonian Cyrillic manuscripts are presented.

Several measures are computed to evaluate the precision and recall of the classifier. 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 fuzzy classifier obtained for Macedonian manuscripts are presented in Table 5.

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.

8. CONCLUSIONS

The fuzzy classification approach for Old Slavic manuscripts written with Macedonian alphabet discussed in this work achieve acceptable results with 76% of recognized letters.

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.

Features that are used in the recognition process are built in the system thus eliminating the need for learning with large training sets. Decision trees used for feature selection reduce the training and the running time of the fuzzy classifier.

Based on the similarity of the letter graphemes it is expected that fuzzy classifier is applicable to Bosnian manuscripts with comparable results.

Table 5. Precision and recall of the fuzzy classifier

Letter	Recall	Precision	Letter	Recall	Precision
Aa	0.71	0.59	t	1	0.79
b	0.75	0.79	U	0.33	1
v	0.89	1	f	0.8	0.89
g	0.56	0.83	H	0.5	0.6
d	1	0.9	h	1	0.75
e	0.43	1	w	0.33	1
/	0.63	0.83	]	1	0.69
\	0.9	0.86	c	0.77	0.59
z	0.25	1	;	0.83	0.71
J	0.5	0.5	[	0.82	1
i	0.8	0.92	q	0.67	0.57
k	0.8	0.8	Q	0.5	1
l	0.94	0.85	2	0.77	0.91
m	1	1	`	0.4	1
n	0.95	0.82	1	1	1
o	1	0.6	5	0.33	0.5
p	0.94	0.88	3	0.25	1
r	0.67	0.8	u	0.63	0.71
s	0.6	0.5			

REFERENCES

Arica, N. and Yarman-Vural,F.T. (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics*. 31(2). pp.216-233.

Atanassov, K. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*. 20(1). pp.87-96.

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

Blum, A.L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, pp. 245-271.

Bortolozzi, F., Britto, J., Oliveira, L. and Morita, M. (2005). Recent advances in handwriting recognition. In Umpala Pal et al. (eds.), *Document Analysis.* pp. 1-31.

Cardie, C. (1993). Using decision trees to improve case-based learning. *In Proc. of 10th Int. Conf. on Machine Learning*, pp. 25-32, Mogran Kaufman.

Casey, R.G. and Lecolinet E. (1996). A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 18(7). pp. 690–706.

Cateni, S., Vannucci, M., Vannocci, M. and Colla, V. (2013). Variable selection and feature extraction through artificial intelligence techniques. In Leandro Valim de Freitas and Ana Paula Barbosa Rodrigues de Freitas (eds.), *Multivariate Analysis in Management, Engineering and the Sciences,* InTech.

Cheriet, M., Kharma, N., Liu, C. and Suen, C. (2007). *Character Recognition Systems, A Guide for Students and Practioners*. Wiley and sons.

D'Amato, D., Kuebert, E. and Lawson, A. (2000). 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, pp. 189–198.

De Stefano, C., Fontanella, F., Marrocco, C. and Scotto di Freca, A.(2014). A GA-based feature selection approach with an application to handwritten character recognition. *Pattern Recognition Letters*.35.pp.130-141.

Gader, P., Forester, B., Ganzberger, M., Gillies, A., Mitchell, B., Whalen, M. and Yocum, T. (1991). Recognition of

handwritten digits using template and model matching. *Pattern Recognition*. 5(24). pp. 421-431.

Gatos, B., Pratikakis, I. and Perantonis, S. (2004). Locating text in historical collection manuscripts. *Lecture Notes in Artificial Intelligence.* 3025. pp. 476-485.

Gorski, N., Anisimov, V., Augustin, E., Baret, O. and Maximor, S. (2001). Industrial bank check processing: the A2iA check reader. International Journal of Document Analysis and Recognition. 3. pp. 196-206.

Hahn-Ming, L., Chih-Ming, C., Jyh-Ming, C. and Yu-Lu, J. (2001). An efficient fuzzy classifier with feature selection based on fuzzy entropy, *IEEE Transactions on Systems, Man, and Cybernetics,* part b: Cybernetics, 31(3), pp. 426-432.

Kim, G. and Kim S. (2000). Feature selection using genetic algorithms for handwritten character recognition. *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, Nijmegen: International Unipen Foundation. pp.103-112.

Kohavi, R. and John,G.H.(1997).Wrappers for feature subset selection. *Artificial Intelligence*,97(1-2),pp.273-323.

Kwak N. and Choi C. (2002). Input feature selection for classification problems. *IEEE Transaction on Neural Networks*, 13(1), pp. 143-159.

Ladha, L. and Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*. 3(5). pp. 1787-1797.

Malaviya, A. and Peters, L. (1995). Extracting meaningful handwriting features with fuzzy aggregation method. *Proc. 3rd Int. Conf. on Document Analysis and Rec*ognition. Montreal. pp. 841-844.

Malaviya, A. and Peters, L. (2000). Fuzzy handwritten description language: FOHDEL. *Pattern Recognition*. 33. pp.119-131.

Namboodiri, A. and Jain, A. (2004). Online handwritten script recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 26 (1). pp.124-130.

Ranawana, R., Palade, V. and Bandara, G.E.M.D.C. (2004). An efficient fuzzy method for handwritten character recognition. In Negoita M. Gh. Et al. (eds.): *KES 2004, LNAI 3214*. pp.698-707. Springer-Verlag.

Trier, Ø.D., Jain, A.K. and Taxt, T. (1996). Feature extraction methods for character recognition - A survey. *Pattern Recognition*. 29(4). pp. 641–662.

Rezaee, M. R., Goedhart, B., Lelieveldt, B.P.F. and Reiber, J.H.C. (1999). Fuzzy feature selection. *Pattern Recognition.* 32. pp. 2011-2019.

Rauber, T. W. and Steiger-Garcao, A.S. (1993). Feature selection of categorical attributes based on contingency table analysis. *In Proc. of the 5th Portuguese Conference on Pattern Recognition*, Porto, Portugal.

Russian Review of Cyrillic Manuscripts, http://xlt.narod.ru/pg/alpha.html

Setiono, R. and Liu, H. (1997). Neural-network feature selector, *IEEE Transactions on Neural Networks.* 8 (3). pp.654-662.

Siedlecki, W. and Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters.* 10(5). pp.335-347.

Tan, P.N., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining.* Addison-Wesley.

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

Verikas, A. and Bacauskiene, M. (2002). Feature selection with neural networks. *Pattern Recognition Letters*. 23(11). pp. 1323–1335.

Vinciarelli, A. (2002). A survey on off-line cursive word recognition. *Pattern Recognition*. 35. pp.1433-1446.

Wang L. and Jiao L. (2006). Multi-layer perceptrons with embedded feature selection with application in cancer classification. *Chinese Journal of Electronics*. 15. pp. 832-835.

Weston J., Mukherjee S. Chapelle O., Pontil M., Poggio T. and Vapnik V. (2000). Feature selection for SVMs. In S.A. Solla, T.K.Leen and K-R Muller (eds.), *Advances in Neural Information Processing Systems.* 12. pp. 526-532. MIT Press, USA.

Yager, R. (1990). On the Representation of multi-agent aggregation using fuzzy logic. *Cybernetics and Systems*. 21. pp. 575-590.

Zhang, G. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics*. 30(4). pp. 451-462.