Fuzzy Classifier for Church Cyrillic Handwritten Characters

Cveta Martinovska[1], Igor Nedelkovski[2], Mimoza Klekovska[2] and Dragan Kaevski[3]
[1]*Computer Science Faculty, University Goce Delcev, Tosho Arsov 14, Stip, R. Macedonia*
[2]*Faculty of Technical Sciences, University St Kliment Ohridski, Ivo Ribar Lola bb, Bitola, R. Macedonia*
[3]*Faculty of Electrical Engineering and Information Technologies, University St Cyril and Methodius, Rugjer Boshkovik bb Skopje, R. Macedonia*

Keywords: Handwritten Character Recognition, Historical Manuscripts Recognition, Fuzzy Decision Techniques, Feature Extraction, Recognition Accuracy and Precision.

Abstract: This paper presents a fuzzy methodology for classification of Old Slavic Cyrillic handwritten characters. The main idea is that the most discriminative features are extracted from the outer character segments defined by intersections. Prototype classes are formed using fuzzy aggregation techniques applied over the fuzzy rules that constitute the descriptions of the characters. Recognition methods use features like number and position of spots in outer segments, compactness, symmetry, beams and columns to assign a pattern to a prototype class. The accuracy and precision of the fuzzy classifier are evaluated experimentally. This fuzzy recognition system is applicable to a large collection of Old Church Slavic Cyrillic manuscripts.

1 INTRODUCTION

Recognition of handwritten characters has been a subject of intensive research in the last 20 years (Arica and Yarman-Vural, 2001); (Vinciarelli, 2002). Different approaches for developing handwritten character recognition systems are proposed, like Fuzzy Logic (Malaviya and Peters, 2000); (Ranawana et al., 2004), Neural Networks (Zhang, 2000) and Genetic Algorithms (Kim and Kim, 2000).

This paper describes a character recognition system developed for digitalization of a large Old Cyrillic manuscripts collection found in Macedonian churches and monasteries. This process cannot be performed using the existing computer software due to the specific properties of Old Slavic characters.

A novel classification methodology based on the fuzzy descriptions of characters is proposed. Number and position of spots, beams and columns that appear in the outer segments of the topological character map are considered as significant features. This character recognition system is applicable to a large historical collection of manuscripts that originate from various periods and locations. The manuscripts used for church liturgical purposes are unaffected by style changes. They are written in Constitutional Script. This Script looks like printed text, where character contour lines can be easily extracted.

2 CHARACTER ANALYSIS AND FEATURE EXTRACTION

Manuscripts are converted to black and white bitmaps. The first step of processing is extracting the characters using contour following function (Fig. 1). Visual prototype of a normalized character is analyzed to determine character features and their membership functions. Several features are examined, such as compactness, x-y symmetry, presence of beams and columns in three horizontal and vertical segments and number of spots in outer segments.

According to visual features, the characters of the Church Slavic alphabet can be grouped in several subsets. There is a subset whose members are Г, В and Б that have emphasized vertical lines on the left-side or left column. Another subset contains characters such as П and Ш that have a right-side and left-side column. The third subset consists of characters like П, Г and Б that have noticeable horizontal line in the upper segment (upper beam). The fourth subset consisting of characters as Ш and

Martinovska C., Nedelkovski I., Klekovska M. and Kaevski D..
Fuzzy Classifier for Church Cyrillic Handwritten Characters.
DOI: 10.5220/0003968403100313
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (ICEIS-2012), pages 310-313
ISBN: 978-989-8565-10-5

q has horizontal line in the bottom segment (bottom beam).

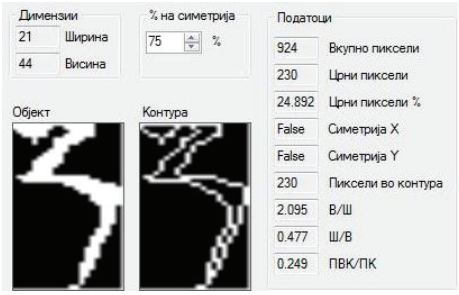

Figure 1: Extracting a character contour.

These features are illustrated in Fig. 2. Particular character can be a member of several of these subsets.

Figure 2: Vertical lines and horizontal lines of characters.

The character can be intersected in such a way that 6 segments are formed. Four outer segments provide useful information for the proposed character recognition system (Fig. 3).

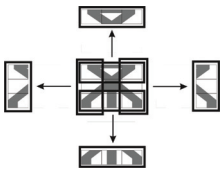

Figure 3: Intersections of the characters that form the upper, down, left and right segments.

Visual prototype of a character is formed applying fuzzy intersection and fuzzy union operators over a set of character samples (Fig. 4).

a) Fuzzy Intersection b) Fuzzy Union

Figure 4: a) Fuzzy intersection and b) fuzzy union.

3 FUZZY CLASSIFIER

Fuzzy classifier for Church Slavic characters is based on character prototypes created in the form of fuzzy linguistic rules. Fuzziness emerges from the fact that texts are written by individuals with different ways of writing and manuscripts originate from different historical periods characterized with certain styles of writing.

3.1 Operators for Fuzzy Aggregation

The precision of the character recognition system to a certain extent depends on the proper selection of features. This is done by calculating the overall measure for the features applying the fuzzy aggregation techniques. The general fuzzy membership function μ_G that combines the fuzzy information $(\mu_1, \mu_2, \dots, \mu_N)$ for the character features can be represented as:

$$\mu_G = Agg(\mu_1, \mu_2, \dots, \mu_N) \quad (1)$$

where Agg is a fuzzy aggregation operator.

This approach uses operators defined by Yager (Yager, 1990) for the union and for the calculation of weighted median aggregation.

Let $w_1, w_2, \dots, w_N$ represent weights associated with fuzzy sets $A_1, A_2, \dots, A_N$. Yager defines the union using the following formula:

$$U(a_1, a_2, \dots, a_N) = \min\left\{1, \left(\sum_{i=1}^{N} (a_i)^{\alpha}\right)^{\frac{1}{\alpha}}\right\} \quad (2)$$

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

The weighted median aggregation is defined by the following formula (Malaviya and Peters, 1995):

$$Med(a_1, \dots, a_N, w_1, \dots, w_N) = \left(\sum_{i=1}^{N} (w_i a_i)^{\alpha}\right)^{\frac{1}{\alpha}} \quad (3)$$

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

3.2 Fuzzy Descriptions of Characters

The Church Slavic character recognition system operates in two working regimes: building the prototypes and character recognition. The first regime creates a matrix of combined characteristics. Using this matrix, fuzzy rules are generated in the form of linguistic descriptions of the characters.

The rules contain only the features that are relevant for the character classification and identification. For example, the character "B" is described by the following combination of features: two vertical holes, x symmetry, left column, one spot in the left segment, one spot in the upper segment, two spots in the right segment, and one

spot in the lower segment. In the fuzzy description of this character less significant features are upper beam and lower beam.

Let the number of significant features for the particular character is S and the number of segments is C. The importance of every feature in the aggregation process is represented by a certain weight. The input values in the system are the features of the character segments for which the fuzzy values are calculated.

Generally, the input matrix with the character features has dimension C x G, where G is the total number of features that can be of structural nature (symmetry, compactness, number of spots) or to denote position:

$$I = \{i_{cg} | c = [1, C], g = [1, G]\} \quad (4)$$

From the elements of the above matrix, a K_j matrix with dimensions p x C can be formed for each segment, where p is a number of significant features for the segment and $j = 1, \dots, C;\ p \leq G$.

$$K_j = \{\bar{k}_{p_1 j}, \bar{k}_{p_2 j}, \dots, \bar{k}_{pCj}\} \quad (5)$$

Using the weighted median aggregation operator, the feature vector for a particular character is calculated

$$\mu_j = \mathrm{Med}(a_1, \dots, a_N, w_1, \dots, w_N) \quad (6)$$

Then, using the union operator, the most significant features from the list of features obtained in the previous step, are selected:

$$\mu_p = \min\left\{1, \left(\sum \mu_{pj}\right)\right\} \quad (7)$$

Weight matrices implicitly represent fuzzy rules that describe the character prototypes. The weights are obtained by statistical calculations from the training samples. The number of appearances of a particular feature is measured for every character. Higher frequency of a feature decreases its recognition importance. Smaller weights are assigned to more frequent and hence less important features.

Weight matrices are used to reduce the number of features that are considered for each character. Different features are considered at each step in the recognition phase and character prototypes that possess these features are activated. Finally, only the most similar character prototype is winner.

4 CHARACTER RECOGNITION

Procedure for character recognition consists of several steps: 1. Determining the membership functions for the global features of the unknown symbol. 2. Calculating the membership functions of an unknown symbol for all the prototypes according to formula:

$$\mu_n = \frac{\sum_{c=1}^{C} w_c . \mu_c}{C} \quad n = 1, .. N \quad (8)$$

3. Selecting the possible prototypes that are most similar to the unknown character, following the formula:

$$\mu_A = \bigcup_{n=1}^{N} \mu_n \quad (9)$$

The result of the classification process is a list of prototypes that have the most similar features with the features of the unknown character.

5 EXPERIMENTAL RESULTS

Several experiments are performed to test the performance of the proposed fuzzy classifier. Table 1 shows the recall and the precision measures for each character. Recall (R) is computed as a fraction of the number of retrieved correct characters divided by the total number of relevant characters:

$$R = TP/(TP + FN) \quad (10)$$

Precision (P) is computed as a fraction of the number of retrieved correct characters, divided with the number of retrieved characters:

$$P = TP/(TP + FP) \quad (11)$$

In formulas (10) and (11), the TP (True Positive) is the number of correctly predicted examples, FP (False Positive) is the number of negative examples wrongly predicted as positive, and FN (False Negative) is the number of positive examples wrongly predicted as negative. The sum of precision and recall i.e. F1 metric is computed as

$$F1 = 2RP/(R + P) \quad (12)$$

The proposed fuzzy classifier recognizes the characters with an average recall of 0.69, average precision of 0.72 and an overall average measure of precision and recall F1 of 0.70.

6 CONCLUSIONS

In this paper a novel methodology for recognition of Old Slavic Cyrillic handwritten characters based on fuzzy prototypes is described. Fuzzy descriptions of

the characters are represented as fuzzy rules. Fuzzy aggregation techniques are used to combine different character features, such as number and position of spots in outer segments, compactness, symmetry, beams and columns.

Table 1: Precision and recall of the fuzzy classifier.

	Number of characters	recall	precision
Aa	10	0.4	1
b	7	0.67	0.67
v	6	0.83	1
g	10	1	0.58
d	12	0.75	0.56
e	7	0.43	1
/	5	0.4	1
\	6	0.67	0.36
z	4	0.25	0.5
J	12	0.83	1
i	5	0.4	1
k	1	1	o.5
l	10	0.9	0.75
m	4	0.25	1
n	11	1	0.73
o	8	1	0.61
p	4	0.5	0.67
r	5	0.2	1
s	9	0.89	0.73
t	7	1	0.78
U	4	0.75	1
f	7	1	0.87
H	6	0.67	0.8
h	9	0.28	0.67
w	5	0	0
]	7	1	1
c	11	0.91	0.58
;	10	0.8	0.89
[	7	0.86	0.75
q	14	0.86	0.63
Q	9	0.67	0.67
2	9	1	0.9
`	5	0.2	1
1	1	1	1
5	2	1	0.67
3	5	0	0
u	3	1	0.43
Total	257	0.69	0.72

Higher weights are assigned to features that are more discriminative. For example, three spots right/left/up or down and two holes are the most indicative for the recognition process.

The accuracy and precision of the proposed fuzzy classifier are acceptable and motivational for future work and improvement.

Presented experimental results of this visual methodology are comparable to the recognition of the human visual system. Characters that are misclassified are also unrecognizable for the humans. Besides the fuzzy classifier a decision tree classifier is designed. The recognition results of the two classifiers are comparable. Both classifiers use the same set of discriminative features.

For future work a combination of these two classifiers is planned to achieve more accurate and precise recognition of the Old Slavic Cyrillic characters.

REFERENCES

Arica, N., Yarman-Vural, F. T., 2001. An Overview of Character Recognition Focused on Off-Line Handwriting. *IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Rev.*, vol. 31, no. 2, pp. 216-233.

Kim G., Kim S., 2000. Feature Selection using Genetic Algorithms for Handwritten Character Recognition, *In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proc. of the 7th Int. Workshop on Frontiers in Handwriting Recognition,* pp 103-112.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On Combining Classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3.

Malaviya A., Peters L., 1995. Extracting Meaningful Handwriting Features with Fuzzy Aggregation Method, *Proc. of the 3rd Int. Conf. on Document Analysis and Recognition*, Montreal, pp. 841-844

Malaviya A., Peters L., 2000. Fuzzy Handwritten Description Language: FOHDEL, *Pattern Recognition*, 33, pp. 119-131.

Ranawana, R., Palade V., Bandara, GEMDC, 2004. An efficient Fuzzy method for Handwritten Character Recognition, *In M.Gh. Negoita et al. (eds.), KES 2004, LNAI 3214*, Springer-Verlag, pp.698-707.

Vinciarelli, A., A Survey on Off-line Cursive Word Recognition, 2002. *Pattern Recognition* 35, pp.1433-1446.

Yager, R., 1990. On the Representation of Multi-Agent Aggregation using Fuzzy Logic, *Cybernetics and Systems* 21, pp.575-590.

Zhang, G. P., 2000. Neural Networks for Classification: A Survey. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 30 no.4, pp.451-462.