Analysis of the South Slavic Scripts by Run-length Features of the Image Texture

Darko Brodic[1], Zoran N. Milivojevic[2], Alessia Amelio[3]

[1]*University of Belgrade, Technical Faculty in Bor,*
Vojske Jugoslavije 12, 19210 Bor, Serbia
[2]*College of Applied Technical Sciences,*
Aleksandra Medvedeva 20, 18000 Niš, Serbia
[3]*Institute for High Performance Computing and Networking, National Research Council of Italy,*
CNR-ICAR, Via P. Bucci 41C, 87036 Rende (CS), Italy
dbrodic@tf.bor.ac.rs

Abstract—**The paper proposes an algorithm for the script recognition based on the texture characteristics. The image texture is achieved by coding each letter with the equivalent script type (number code) according to its position in the text line. Each code is transformed into equivalent grey level pixel creating an 1-D image. Then, the image texture is subject to the run-length analysis. This analysis extracts the run-length features, which are classified to make a distinction between the scripts under consideration. In the experiment, a custom oriented database is subject to the proposed algorithm. The database consists of some text documents written in Cyrillic, Latin and Glagolitic scripts. Furthermore, it is divided into training and test parts. The results of the experiment show that 3 out of 5 run-length features can be used for effective differentiation between the analysed South Slavic scripts.**

Index Terms—**Classification, coding, image analysis, script, texture.**

I. INTRODUCTION

The Balkan region, which is populated by South Slavs, is very rich in cultural heritage elements dated from medieval age. One of the most important cultural achievements represents the variety of used scripts. In the medieval age, South Slavs had spoken the old Church Slavonic language. It was written with the Glagolitic alphabet called round Glagolitic script, but later it was replaced by the Cyrillic script in the east region of Balkan, i.e. in Bulgaria and Macedonia. In Bosnia, the local version of Cyrillic alphabet was used, while in Croatia, a variant of the Glagolitic alphabet called squared or angular Glagolitic script was preserved. Accordingly, all books from medieval age were written by the aforementioned scripts. Currently, Serbian language is the only European standard language with complete synchronic diagraphia, which uses both Cyrillic and Latin alphabets.

The Serbian language has been studied in a natural speech analysis as a part of the South Slavic language group [1]. Similarly, the recognition of the South Slavic scripts can be used in document image analysis (DIA) and optical character recognition (OCR) [2]. Recognition of different script characters in an OCR module is a difficult task [3]. It is so, because the character recognition depends on different features like the structural properties, style and nature of writing, which generally differ from one script to another. It is especially true for digitalization and script recognition of the old books from the Balkan region, which are written by different scripts, i.e. Cyrillic, Glagolitic and Latin scripts.

The script recognition techniques have been classified as global or local. Global methods treat the document image as a group of big image blocks. Then, the image blocks are statistically analysed [4]. The drawback of these methods corresponds to processing the noise images, which can decay the recognition results [5]. Local approaches segment document images into small blocks representing text pieces, i.e. connected components. Then, they are subject to the statistical analysis like the run-length, co-occurrence, local binary patterns, etc. [6]. The efficiency of the local methods is unaffected by the noise in the document image. Unfortunately, they are more computer time intensive than global methods.

This paper presents a method similar to local methods. It is based on coding each letter in the script by its position in the text line [5]–[7]. Accordingly, the method takes into account energy profile of each script sign in a certain text line and classifies it into four different groups [3]. Hence, the efficiency of the method is linked to the basis of the printed text. In this way, the number of current variables is considerably reduced compared to the initial ones. The received codes are translated into the grey level pixels of the image. Then, the image texture is subject to the run-length analysis. Extracted run-length features are the basis for distinguishing the scripts. This goal is accomplished by the classification tool GA-ICDA, an extension of GA-IC framework [8], which segments the data into clusters representing different scripts.

The remainder of the paper is organized as follows. Section II addresses all aspects concerning the proposed algorithm. It includes the text-line definition and script modelling. The result represents the coded text which

corresponds to 1-D four grey level image. Furthermore, it is subject to run-length analysis of the image texture. Then, the obtained results are analysed and classified by the clustering algorithm. Section III describes the experiment and custom oriented database consisting in training and test document sets written in different scripts. Section IV gives the results of the experiment and discusses them. Section V makes conclusions and points out further research direction.

II. ALGORITHM

The proposed algorithm consists in the following three stages: (i) script coding, (ii) texture analysis, and (iii) feature classification. Script coding translates each letter into the grey-level pixel of the image in accordance to the position in the text line. Then, the texture analysis process derives a 1-D image to extract run-length texture features. These features are subject to the classification process in order to cluster classes representing documents written in different scripts. Fig. 1 shows the stages of the proposed algorithm.

Fig. 1. The stages of the proposed algorithm.

A. Script Coding

Document text can be segmented into text lines. Furthermore, each text line can be separated taking into account the energy of the script signs [4] into the following virtual lines [9]: (i) top-line, (ii) upper-line, (iii) base-line, and (iv) bottom-line. These lines divide the text line area into three vertical zones [9]: (i) upper zone, (ii) middle zone, and (iii) lower zone.

The letters can be classified depending on their position in vertical zones of the line, which reflects theirs energy profile. The short letters (S) occupy the middle zone only. The ascender letters (A) spread over the middle and upper zones. The descendent letters (D) enlarge into the middle and lower zones. The full letters (F) outspread over all vertical zones. Hence, all letters can be grouped into four different script types [7]. Figure 2 shows the script characteristics according to the letter baseline position.
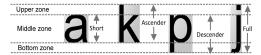

Fig. 2. Illustration of the virtual lines and vertical zones in the text line.

Each of the script types, i.e. S, A, D and F, can be exchanged with different number codes, i.e. 0, 1, 2 and 3, respectively. Accordingly, there exist only four script types, which lead to four number codes. To create a texture, these numbers are transformed into different levels of grey. Figure 3 shows the equivalence between script type number codes and grey levels.

Fig. 3. Script type equivalent number codes and corresponding grey levels.

In this way, each text is transformed into the set of number codes and then into the pixels of only four grey levels. Obtained image is a 1-D image $\mathbf{I_m}$ representing a texture, which can be subject to the texture analysis.

B. Texture Analysis

Texture is a measure of the intensity variation in the image surface. Hence, it is used to extract information from the images that can quantify the properties like smoothness, coarseness, and regularity [10]. Accordingly, the texture is characterized by calculating statistical measures obtained from the greyscale intensities in the image. The aforementioned statistical measures can be used for classification and segmentation of the image [11].

Run-length statistical analysis is used to evaluate and quantify the coarseness of a texture [12]. A run can be defined as a set of consecutive pixels, which are characterized by the same grey level intensity in a specific direction. Typically, coarse textures contain long runs with different grey level intensities, while fine textures include short runs with similar grey level intensities.

Let's suppose that we have an image $\mathbf{I_m}$ featuring X rows, Y columns and M levels of grey intensity. The starting point is the extraction of the run-length matrix $p(i, j)$. It is defined as the number of runs with pixels of grey level $i = 1,..., M$ and run length $j = 1,..., N$, where M is the number of grey levels, while N is the number representing the maximum run length. In our case, each element of the run-length matrix $p(i, j)$ represents the grey level run-length of the 1-D matrix $\mathbf{I_m}$ that gives the total number of occurrences of grey-level runs of length j and of intensity value i. A set of consecutive pixels with identical or similar intensity values constitutes a grey level run. Furthermore, various texture matrices and vectors representing semi-features can be extracted from the run-length matrix $p(i, j)$ [13]: (i) grey level run-length pixel number matrix p_p, (ii) grey level run-number vector p_g, (iii) run-length run-number vector p_r, and (iv) grey level run-length-one vector p_o.

Grey level run-length pixel number matrix p_p is defined as:

$$p_p(i, j) = p(i, j) \times j. \qquad (1)$$

Grey level run-number vector p_g is calculated as:

$$p_g(i) = \sum_{j=1}^{N} p(i, j). \qquad (2)$$

Run-length run-number vector p_r is given as:

$$p_r(j) = \sum_{i=1}^{M} p(i, j). \qquad (3)$$

Grey level run-length-one vector p_o is:

$$p_o(i) = p(i, 1). \qquad (4)$$

Using aforementioned matrix and vectors, the following run-length features were originally proposed in [14]: (i) Short run emphasis (SRE), (ii) Long run emphasis (LRE), (iii) Grey-level non-uniformity (GLN), (iv) Run length non-

uniformity (*RLN*), and (v) Run percentage (*RP*).

SRE measures the distribution of short runs. It is highly dependent on the occurrence of short runs and is expected to be large for fine textures. It is given as:

$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2} = \frac{1}{n_r} \sum_{j=1}^{N} \frac{p_r(j)}{j^2}. \qquad (5)$$

LRE measures the distribution of long runs. It is highly dependent on the occurrence of long runs and is expected to be large for coarse structural textures. It is calculated as:

$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \times j^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j) \times j^2. \qquad (6)$$

GLN measures the similarity of grey level values throughout the image. Its value is expected to be small if the grey level values are alike throughout the image. It is computed as:

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i)^2. \qquad (7)$$

RLN measures the similarity of the length of runs throughout the image. It is expected to receive smaller values if the run lengths are alike throughout the image. It is defined as:

$$RLN = \frac{1}{n_r} \sum_{j=1}^{N} \left(\sum_{i=1}^{M} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j)^2. \qquad (8)$$

RP measures the homogeneity and the distribution of runs of an image in a specific direction. It receives the largest values when the length of runs is 1 for all grey levels in a specific direction. It is given as:

$$RP = \frac{n_r}{n_p}. \qquad (9)$$

In (5)−(9) n_r represents the total number of runs, while n_p is the number of pixels in the image $\mathbf{I_m}$.

C. Classification

Obtained features can be analysed in order to extract only run-length features that can distinguish different scripts. For this purpose, we adopt an extension of the GA-IC tool in ref. [8], called GA-ICDA (Genetic Algorithms Image Clustering for Document Analysis). GA-IC is used for clustering images from a database. It creates a weighted graph. The nodes of the graph correspond to images. An edge of the graph connecting two images is established if the images are similar. Each edge has its weight, which expresses the level of similarity extracted from the feature vectors of each image. The graph of images is then clustered by applying a genetic algorithm that divides the graph into groups of nodes. GA-ICDA introduces three new aspects from GA-IC. First of all, each graph node is a document represented as a

run-length feature vector. Secondly, an edge connects two nodes if related documents are similar and also the node distance is less than a threshold *T*, established a node ordering. Finally, a refinement procedure at the end of the genetic algorithm merges pairs of clusters with minimum distance to each other, until a fixed number is reached.

III. EXPERIMENT

The goal of the experiment is to prove the effectiveness of the proposed algorithm for the script recognition. Serbian or Croatian languages can be written by three different scripts: Cyrillic, Latin and Glagolitic. Hence, it is suitable for the experiment. For the purpose of the experiment, a custom oriented database of text documents written in Cyrillic, Latin and Glagolitic scripts is created. The database includes two parts: training and test set of documents. Training set consists in a total of 100 documents, while test set consists in a total of 15 documents. Each set incorporates Cyrillic, Latin and Glagolitic scripts in a similar number of documents. Both sets are subject to the run-length statistical analysis. The obtained results are given and discussed below.

IV. RESULTS AND DISCUSSION

First, the training set is subject to the run-length statistical analysis. The results are given in Table I for the Cyrillic, Latin and Glagolitic scripts.

TABLE I. TRAINING SET OF CYRILLIC SCRIPT.

	SRE	*LRE*	*GLN*	*RLN*	*RP*
Cyrillic					
Minimum	0.59	21.00	40.00	38.00	0.32
Maximum	0.63	25.00	200.00	175.00	0.36
Latin					
Minimum	0.70	6.00	80.00	90.00	0.52
Maximum	0.72	7.00	400.00	390.00	0.56
Glagolitic					
Minimum	0.66	10.00	60.00	65.00	0.39
Maximum	0.68	15.00	230.00	230.00	0.46

Then, the test set is subject to the run-length statistical analysis. The results are given in Table II for the Cyrillic, Latin and Glagolitic scripts.

TABLE II. TEST SET OF CYRILLIC SCRIPT.

	SRE	*LRE*	*GLN*	*RLN*	*RP*
Cyrillic					
Minimum	0.6036	21.3960	42.2632	40.2281	0.3314
Maximum	0.6230	23.9298	188.8730	170.5437	0.3470
Latin					
Minimum	0.7043	6.3769	84.0352	92.8693	0.5344
Maximum	0.7183	6.7704	350.5431	377.5103	0.5517
Glagolitic					
Minimum	0.6600	10.2716	62.7531	69.0247	0.4036
Maximum	0.6779	13.9696	214.4705	226.0555	0.4525

The above results represent the extracted five run-length features for each script. Currently, the classification problem

is mandatory. First, a comparison of the training and test set results is important. If the test set results are a subset of the training set results, then the results are valid. Furthermore, we have to explore the run-length features that can be separated according to the given script. From Table I and Table II, the values of *GLN* and *RLN* features are mutually overlapped between the scripts. However, *SRE*, *LRE* and *RP* characterize each script differently. Figure 4 shows an *SRE* test set of the scripts.

Fig. 4. *SRE* for test set of the scripts.

The *SRE* range given from minimum to maximum obtained values for each script is clearly distinct between each script. Hence, *SRE* is suitable for differentiation between the scripts. Figure 5 shows an *LRE* test set of the scripts.

Fig. 5. *LRE* for test set of the scripts.

The *LRE* range for each script can distinguish between each script. Accordingly, *LRE* is also suitable for distinction between scripts. Figure 6 shows an *RP* test set of the scripts.

The *RP* range for each script is quite distinct from each other. Hence, the combination of the *SRE*, *LRE* and *RP* can be used to freely distinct the characteristics of different scripts like South Slavic scripts: Cyrillic, Latin and Glagolitic. It represents a much easier solution than those given in [15]. GA-ICDA is used to classify the texture features and cluster data.

Our goal is to classify a model and correctly predict the classes that represent documents written by different scripts. Hence, *precision*, *recall* and *f-measure* are preferred metrics

to evaluate the proposed algorithm. *Precision* is the fraction of retrieved and relevant documents with respect to all retrieved documents. *Recall* is the fraction of relevant documents, which are retrieved with respect to all relevant documents. *F-Measure* is the harmonic mean of precision and recall. Classification by GA-ICDA on the test set established three communities with the correct prediction of each document in the right class as given in Table III.
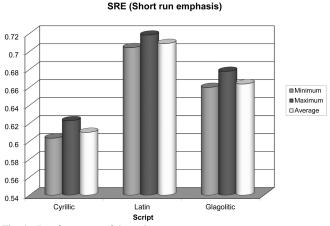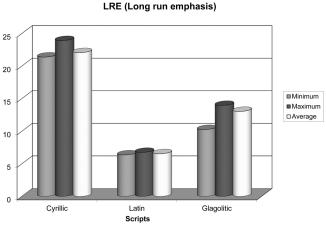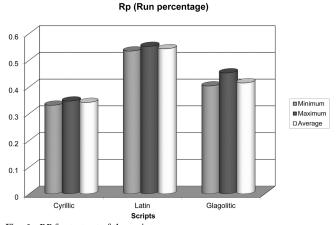

Fig. 6. *RP* for test set of the scripts.

TABLE III. CLASSIFICATION RESULTS ON TEST SET.

	Script	*Precision*	*Recall*	*F-Measure*
GA-ICDA	Cyr.	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	Latin	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	Glag.	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Expectation-Maximization	Cyr.	0.7479 (0.2840)	0.9780 (0.0986)	0.8196 (0.2140)
	Latin	0.6564 (0.2374)	0.9700 (0.1040)	0.7515 (0.1529)
	Glag.	0.4774 (0.0414)	0.9800 (0.0603)	0.6408 (0.0449)
Hierarchical Clustering	Cyr.	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	Latin	0.5000 (0.0000)	1.0000 (0.0000)	0.6667 (0.0000)
	Glag.	0.5000 (0.0000)	1.0000 (0.0000)	0.6667 (0.0000)

Hence, *precision*, *recall* and *f-measure* receive a value of 1. In this way, the proposed algorithm correctly predicts the classification of each document from a database in the adequate class representing Cyrillic, Latin or Glagolitic scripts (three classes). Comparisons with other two well-known classifiers, Hierarchical Clustering and Expectation-Maximization, on the same run-length coded test set revealed the superiority of our approach. Each algorithm has been executed 100 times and the average values together with standard deviation (in parenthesis) are reported.

V. CONCLUSIONS

The manuscript proposed a methodology for the script recognition in the South Slavic documents. It uses the run-length statistical analysis of the document based on the status of each script element in the text line. Due to the difference in the script characteristics, the results of the statistical analysis show significant dissimilarity. It represents a starting point for the feature classification. The proposed method is tested on documents written in Cyrillic, Latin and Glagolitic scripts. The experiments show encouraging

results. Further research direction will be toward the statistical analysis of other scripts.

REFERENCES

[1] S. Sovilj-Nikic, V. Delic, I. Sovilj-Nikic, M. Markovic, "Tree-based phone duration modelling of the Serbian language", *Elektronika Ir Elektrotechnika*, vol. 20, no. 3, pp. 77–82, 2014. [Online] Available: http://dx.doi.org/10.5755/j01.eee.20.3.4090

[2] R. Kasturi, L. O'Gorman, V. Govindaraju, "Document image analysis: A primer", *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002. [Online] Available: http://dx.doi.org/10.1007/BF02703309

[3] D. Ghosh, T. Dube, A. P. Shivaprasad, "Script recognition - a review", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010. [Online] Available: http://dx.doi.org/10.1109/TPAMI.2010.30

[4] G. D. Joshi, S. Garg, J. Sivaswamy, "A generalised framework for script identification", *International Journal of Document Analysis and Recognition*, vol. 10, no. 2, pp. 55–68, 2007. [Online] Available: http://dx.doi.org/10.1007/s10032-007-0043-3

[5] A. Busch, W. W. Boles, S. Sridharan, "Texture for script identification", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1720–1732, 2006. [Online] Available: http://dx.doi.org/10.1109/TPAMI.2005.227

[6] U. Pal, B. B. Chaudhury, "Identification of different script lines from multi-script documents", *Image and Vision Computing*, vol. 20, no. 13–14, pp. 945–954, 2002. [Online] Available: http://dx.doi.org/10.1016/S0262-8856(02)00101-4

[7] D. Brodi , Z. N. Milivojevi , . A. Maluckov, "Recognition of the script in Serbian documents using frequency occurrence and co-occurrence analysis", The Scientific World Journal, vol. 2013, article ID 896328, pp. 1–14, 2013. [Online] Available: http://dx.doi.org/10.1155/2013/896328

[8] A. Amelio, C. Pizzuti, "A new evolutionary-based clustering framework for image databases", in *Image and Signal Processing* (Eds. A. Elmoataz et al.), ICISP 2014, LNCS 8509, pp. 322–331, Springer International Publishing Switzerland, 2014. [Online] Available: http://dx.doi.org/10.1007/978-3-319-07998-1_3

[9] A. W. Zramdini, R. Ingold, "Optical font recognition using typographical features", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 877–882, 1998. Online] Available: http://dx.doi.org/10.1109/34.709616

[10] G. N. Srinivasan, G. Shobha, "Statistical texture analysis", in *Proceesings of World Academy of Science, Engineering and Technology*, Bangkok, Thailand, 2008, vol. 36, pp. 1264–1269.

[11] R. Jain, R. Kasturi, B. G. Schunck, *Machine Vision*. McGraw-Hill: U.S.A., 1995, pp. 234–248.

[12] D. Xu, A. Kurani, J. Furst, D. Raicu, "Run-length encoding for volumetric texture", in *Proceesings of International Conference on VIIP*, Marbella, Spain, 2004, pp. 452–458.

[13] X. Tan, "Texture information in run-length matrices", *IEEE Transaction on Image Processing*, vol. 7, no. 11, pp. 1602–1609, 1998. [Online] Available: http://dx.doi.org/10.1109/83.725367

[14] M. M. Galloway, "Texture analysis using grey level run lengths", *Computer Grahics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975. [Online] Available: http://dx.doi.org/10.1016/S0146-664X(75)80008-6

[15] D. Brodi , Z. N. Milivojevi , . A. Maluckov, "An approach to the script discrimination in the Slavic documents", *Soft Computing*, (In Press). [Online] Available: http://dx.doi.org/10.1007/s00500-014-1435-1