Dating the Historical Documents from Digitalized Books by Orthography Recognition

Darko Brodić[1](✉) and Alessia Amelio[2]

[1] Technical Faculty in Bor, University of Belgrade, V.J. 12, 19210 Bor, Serbia
dbrodic@tfbor.bg.ac.rs
[2] DIMES, University of Calabria, Via Pietro Bucci Cube 44, 87036 Rende, CS, Italy
aamelio@dimes.unical.it

Abstract. This paper introduces a new method for automatically dating Serbian and Croatian historical documents. It is based on the concept that the documents in a certain script or language evolving in different historical periods are characterized by differences in orthography rules. Accordingly, we propose three stages of script coding, texture analysis and classification for capturing such a difference. Hence, the input document is transformed into a sequence of numerical codes, each representing an intensity value, determining an image. Then, texture analysis extracts features from the image to create a feature vector. Finally, it is classified for orthography recognition. Results obtained on two databases of historical documents in angular Glagolitic script and Slavonic-Serbian and Serbian languages extracted from digitalized books demonstrate the efficacy of the proposed method.

Keywords: Orthography recognition · Historical documents · Image processing · Digital book · Classification

1 Introduction

Digital libraries include the creation of digital counterparts of the original historical analog books and document materials. This digitization process has a few benefits: (i) digital preservation of the original analog material, (ii) easy way of accessing it through the Internet, (iii) creation of collections, which are previously not known, and (iv) additional potential of researching historical materials.

Documents are written in a certain language, incorporating different orthography rules. Orthography as a word has a background from the Greek words ($\acute{o}\rho\theta\acute{o}\zeta$) and ($\gamma\rho\acute{\alpha}\phi\epsilon\iota\nu$), which means "correct writing". Hence, it is deeply connected to the written language. Among the others, it takes care of capitalization, emphasis, hyphenation, punctuation and word breaks style [13]. In this way, it includes all methods and rules of correct writing. Accordingly, it is a sub-part of natural language processing. On the other side, linguistics sees the orthography as the methodology of writing a language. It considers the orthography as a standardization, which is given as a spectrum of conventions, i.e. set of rules. Some

C. Grana and L. Baraldi (Eds.): IRCDL 2017, CCIS 733, pp. 119–131, 2017.
DOI: 10.1007/978-3-319-68130-6_10

languages incorporate non-consistent orthography such as the English language, while the others such as Serbian language have remarkably consistent orthography. In the example of the Serbian language, it means that one letter in the written language has the equivalence of one sound in the spoken language. Also, it is worth noting that any language during a certain historical period has been evolved under different influences. Accordingly, its orthography was changed, too. Sometimes, the orthographic rules were so changed that they have acknowledged the new historical era of the language. In such cases, the differentiation between the orthographic rules can be used to correctly evaluate the historical period of the analyzed document or book. Basically, the differentiation of documents by their orthography can, in many cases, clearly identify documents' dating and printing origin.

In previous works about orthography, authors mainly tried to establish a link between the discrimination of some tokens during the transformation of the language [18] or among different languages [4]. Hence, they used typical linguistic tools like bi-grams, tri-grams, probabilistic mapping and variation in tokens and vocabulary [24]. The main limitation of these works is that they analyze orthography changes in the same language or among different languages. In this study, we overcome this limitation by proposing a new method for orthography discrimination inside the same language and during the evolution of the written language. In this way, we propose a system for recognition of the historical period of the Serbian and Croatian documents according to the use of the orthographic rules present inside the document. In such a context, the document is analyzed and further processed in its digital format, which is essential for the cultural heritage preservation. To the very best of our knowledge, it is the first time that a similar approach has been introduced in the literature. For evaluating the proposed approach, we conduct two quite different experiments: (i) on digitalized printed documents written by angular Glagolitic script dated between XIV and XIX century, which are written in Croatian language, and (ii) on digitalized printed documents written by Cyrillic script dated between XVIII and XIX century, which are written in Slavonic-Serbian and Serbian languages. In the first experiment of the angular Glagolitic script, the writing of the same script is changed, i.e. evolved over time. Hence we are talking about script evolving rules based on the orthography. In the second experiment, the change of the orthography is used as a change of the language and not only of script writing rules. Basically, the transformation of Slavonic-Serbian to modern Serbian language is established by more means: (i) change of the Cyrillic alphabet (differences in letters), (ii) change of language, i.e. rapid evolvement of the language, and (iii) change of the orthographic rules.

Printed medieval documents were written in the angular Glagolitic script by the Croatian recension of the Church Slavonic language. Angular Glagolitic was mainly used in the regions along the east, i.e. Croatian side of Adriatic coast. Up to XV century, the holy books were written by hand printing. After inventing printing machine by Gutenberg, the printed books spread over Europe. Up to 1561, the printing offices were opened along Croatian Adriatic coast, i.e. in

Senj (from 1491 to 1508) and Rijeka (from 1530 to 1531) [16]. Accordingly, many books were printed by the angular Glagolic script, but using the old orthographic rules [2]. These rules incorporate the writing of capital letters as descendent ones, which is typical for hand printed Glagolitic medieval books. Hence, the use of the old Glagolitic orthography was linked with the printing offices in Rijeka, Senj, Kosinje or Venice, if the editors were Croats. Outside Croatia, the Glagolitic books were mainly printed in Venice, Rome [20], Urach, Tubingen and Prague [3] around and after 1535. However the editors, who were mainly Italians, did not know the old orthography rules. Instead, they used Latin-based orthography in Glagolitic books. This means the use of capital letters as ascending ones. Hence, the change of orthography rules was clearly limited by the historical period before and after 1535.

Slavonic-Serbian language was the literary language used in Habsburg Empire by the Serbs. In the beginning, the Habsburg Empire tried to assimilate the Serbian population by changing their culture and heritage. Accordingly, it wanted to change the Serbian alphabet from Cyrillic to Latin one. Serbian Church opposed to this change by asking help to Russian Empire. Consequently, in that period, all books, especially liturgical ones, as well as literary, were printed in the Russo-Slavonic language. During the XVIII century, the writers began to blend Russo-Slavonic with Serbian language resulting with a mixed language, which is called Slavonic-Serbian. In 1818, Vuk Stefanović Karadžić with the help of Jernej Kopitar and Sava Mrkalj concluded his book Serbian Dictionary in Vienna. This book was the beginning of reforming the Slavonic-Serbian language with its old orthography according to the widespread modern language of the Serb population. It reformed the Slavonic-Serbian language by Cyrillic alphabet using strict phonemic principles. In this way, the modern Serbian language was distanced from Slavonic-Serbian language in different alphabet and new orthography rules. In the meantime, the last notable work in Slavonic-Serbian was published in 1825 [19]. After that, the modern Serbian language with the new orthography style spread in printed books.

To identify orthography changes defining old and new orthography, the proposed approach uses our previous method for language and script discrimination [5–8]. Then, obtained results are subjected to classification by Naive Bayes. In both examples, recognition of different orthography rules used in the digital documents can distinctively identify the printing date of these documents.

The paper is organized as follows. Section 2 introduces the methodology for the orthography differentiation. Section 3 presents the experiments. Section 4 gives the results of the experiments and makes a discussion. Section 5 draws conclusions and outlines future work directions.

2 Methodology

The methodology for orthography recognition is composed of the following steps: (i) script coding, (ii) texture analysis, and (iii) classification.

In script coding, a document is transformed into a sequence of four numerical codes. It is based on character classification in four script types according to their height in the text line area. Hence, the obtained coded text is considered as a 1-D image, where each code represents an intensity level. Then, the texture analysis is employed on the 1-D image for feature extraction. It performs successfully because characters height and their disposition in the text may change according to the orthography rules. Codes envelope the height information. Linear patterns of codes detected by texture analysis capture the characters disposition. For this reason, extracted features exhibit a discriminant capability. At the end, the document represented as a feature vector is classified as given in old or new orthography by Naive Bayes approach. Next, we describe the main steps of the method.

2.1 Script Coding

Script coding represents the first and crucial stage in the proposed methodology. It converts each letter according to its position in the text-line into different script types. The given letters are mapped to a certain script type based on their horizontal energy profiles. In this way, all letters are grouped into: (i) the base letters, (ii) the ascender letters, (iii) the descendent letters, and (iv) the full letters [29]. Figure 1(a) illustrates this grouping.

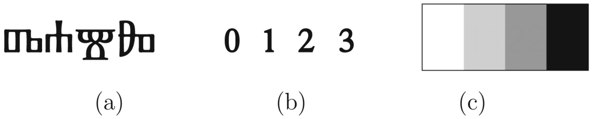
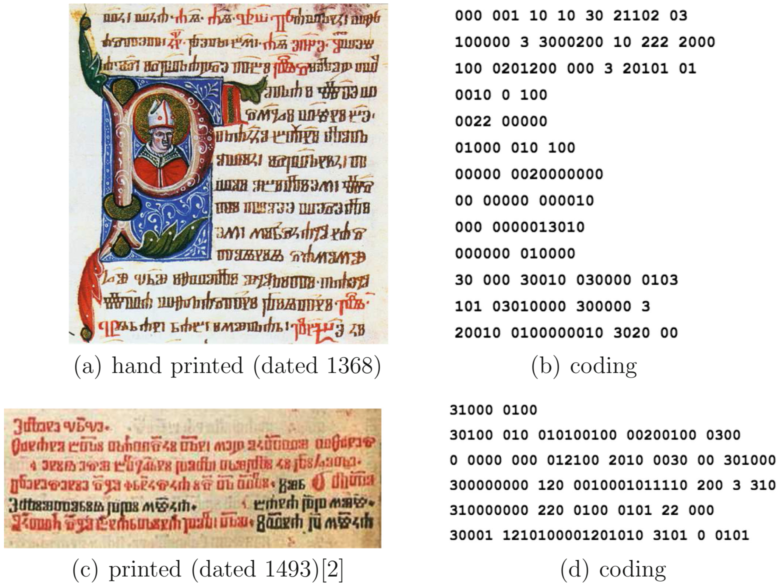


Fig. 1. The basis of the script mapping: (a) initial text, (b) script coding, and (c) transformation into image

Furthermore, this mapped text is coded. The coding is performed as follows: (i) the base letter to 0, (ii) the ascender letter to 1, (iii) the descendent letter to 2, and (iv) the full letter to 3 [7,8]. Figure 1(b) shows the coding. In this way, the initial text is coded into a string corresponding to a combination of the set of codes {0, 1, 2, 3}. The final coded text can be seen as an image $\mathbf{I}(1,j)$ with the pixel values taken from the set {0, 1, 2, 3}. Figure 1(c) shows the transformation to the image. The given transformation enables the reduction of variables from the number of letters in the alphabet to 4. Still, because of the specific transformation, additional information is added such as differentiating between capital and small letters. Furthermore, the transformation of coded text into a 1-D image allows a wide variety of additional analyses to be performed.

Orthography. In the following, we present an example of orthography differentiation by using script coding. Old style Glagolitic orthography implicates the writing of the capital letters as descendent letters. On the contrary, new Glagolitic orthography follows the rule of writing a capital letter as ascender one like in Latin alphabet. Coding will be performed on the old style orthography used in hand printed and printed Glagolitic documents as well as on the new style Glagolitic orthography used in the newer printed Glagolitic books. Figure 2 illustrates hand printed and printed excerpts in old orthography and their coding. Figure 3 depicts an excerpt using the new orthography style and its coding. Figure 4 shows the distribution of the script types of the hand printed and printed Glagolitic documents written with the old orthography rules, and printed Glagolitic document written with the new orthography rules. The graphical results show that differentiation between old and new orthography in Glagolitic documents is clearly visible at the level of use of base, descendent and full letters.

On the contrary, Slavonic-Serbian and Serbian languages are related to different alphabet, different evolving language and according to that, different words in use.



```
000 001 10 10 30 21102 03
100000 3 3000200 10 222 2000
100 0201200 000 3 20101 01
0010 0 100
0022 00000
01000 010 100
00000 0020000000
00 00000 000010
000 0000013010
000000 010000
30 000 30010 030000 0103
101 03010000 300000 3
20010 0100000010 3020 00
```

(a) hand printed (dated 1368)

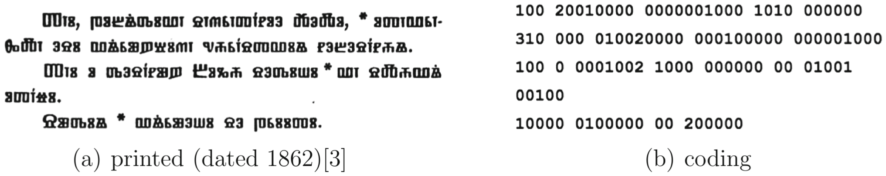
(b) coding

```
31000 0100
30100 010 010100100 00200100 0300
0 0000 000 012100 2010 0030 00 301000
300000000 120 0010001011110 200 3 310
310000000 220 0100 0101 22 000
30001 1210100001201010 3101 0 0101
```

(c) printed (dated 1493)[2]

(d) coding

Fig. 2. Glagolitic hand printed and printed excerpts written in the old orthography and their equivalent coding



```
100 20010000 0000001000 1010 000000
310 000 010020000 000100000 000001000
100 0 0001002 1000 000000 00 01001
00100
10000 0100000 00 200000
```

(a) printed (dated 1862)[3]

(b) coding

Fig. 3. Glagolitic printed excerpt written in the new orthography and its equivalent coding

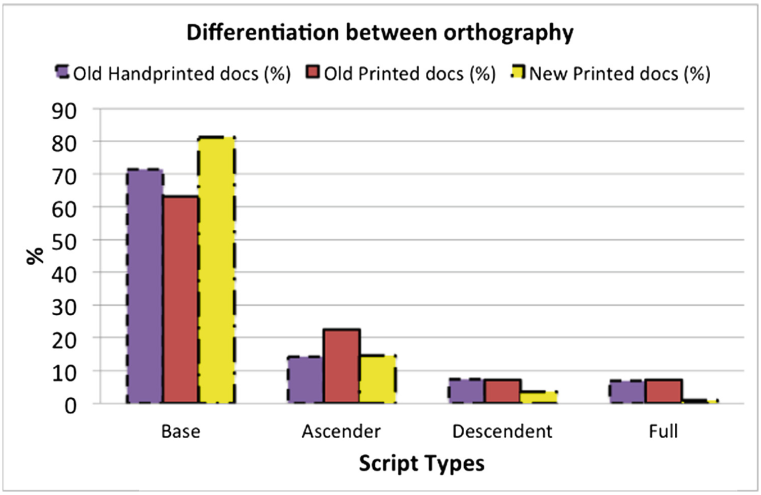


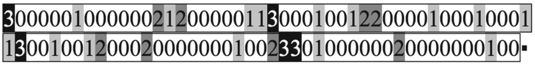
Fig. 4. Distribution of the script types in old and new Glagolitic orthography

2.2 Texture Analysis

Texture analysis is employed on the 1-D image obtained by the script coding phase, and representing the codified document. Texture envelopes information of spatial positioning of intensities in the image, determining the features. Two typical methods adopted for texture feature extraction are: (i) run-length statistics, and (ii) (A)LBP. Texture features extracted from these methods are the input to the classification algorithm.

Run-length statistics is based on the concept of run. It is a sequence of consecutive pixels of the same intensity in a certain direction of the texture. Let **I** be the image with M intensity levels, and N be the maximum run length. A matrix **p** is computed from the runs in **I** along an established direction. Position (i, j) of **p** contains the number of runs of intensity level i and of run length j. From matrix **p**, 11 features are computed: (i) Short run emphasis (SRE), (ii) Long run emphasis (LRE), (iii) Gray-level non-uniformity (GLN), (iv) Run length non-uniformity (RLN), (v) Run percentage (RP) [17], (vi) Low gray-level run emphasis (LGRE) and (vii) High gray-level run emphasis (HGRE) [10], (viii) Short run low gray-level emphasis (SRLGE), (ix) Short run high gray-level emphasis (SRHGE), (x) Long run low gray-level emphasis (LRLGE), and (xi) Long run high gray-level emphasis (LRHGE) [15].

Figure 5 shows a 1-D image with the corresponding numerical codes, and the associated run-length matrix **p**. For example, element at position $(1, 1)$ has a value of 11, which indicates that the number of runs, i.e. consecutive pixels, of intensity level 1 and length 1 in the 1-D image is 11.



(a) 1-D image pattern (a unique sequence, ■ marks the end of the text)

p*(*i*,*j*)** ***j

i

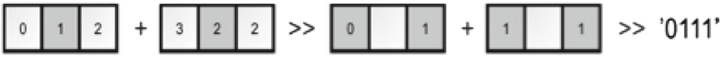
	1	**2**	**3**	**4**	**5**	**6**	**7**
0	1	5	4	1	2	2	2
1	11	2	0	0	0	0	0
2	6	1	0	0	0	0	0
3	3	1	0	0	0	0	0

(b) run-length matrix

Fig. 5. Run-length matrix computed on 1-D image

For each center pixel of the 1-D image, Local Binary Pattern (LBP) examines its neighbor pixels (left and right pixels) to find if their intensity is above or below the center pixel intensity and thresholds them according to this intensity. Then, it creates binary numbers and generates a histogram of the corresponding decimal labels for the overall image [23]. However, LBP determines in our case only 4 different elements, which are not enough to be employed for discrimination [22]. Consequently, the extension to Adjacent LBP (ALBP) is performed [9]. Considering the two horizontal pixels in the neighborhood (LBP(+)), a binary number is created from LBP(+) and adjacent LBPs(+) are combined to create ALBP. This combination creates 4-bit binary numbers between '0000' and '1111' representing 16 features [22].

Figure 6 shows the thresholding procedure for two center pixels of intensity levels 1 and 2 (in gray) of the 1-D image (on the top left), the adjacent LBP(+) combination determining the 4-bit binary number '0111' (on the top right), and the resulting histogram of the decimal labels computed for the overall image, realizing the 16 features (on the bottom).



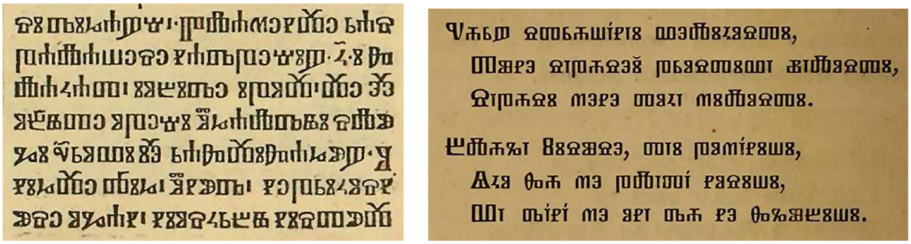
0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
27	0	8	1	0	0	0	0	8	0	0	0	0	0	0	0

Fig. 6. ALBP features computed from 1-D image

3 Experiment

The experiment includes two different parts of the orthography test.

In the first part of the experiment, the text written in the Glagolitic script by old and new orthography style is analyzed. It is performed on a database of thirteen Glagolitic printed documents. Five out of thirteen documents are written using the old orthography. They are text excerpts from pages of the book entitled *Missale Romanum Glagolitice* dated from 1483 [20,21]. Figure 7(a) shows a sample excerpt from the book. It represents the first printed Glagolitic book and the most beautiful of all printed Glagolitic books at all. Eight of thirteen documents are written in the new orthography. Six of them are text excerpts from pages of the book entitled *The Confession and Knowledge of the True Christian Faith* dated from 1564 [20]. The last two documents are text excerpts from pages of the book entitled *Foundations of the Old Slavic language* dated from 1862 [3]. Figure 7(b) illustrates a sample excerpt from the book.



(a)

(b)

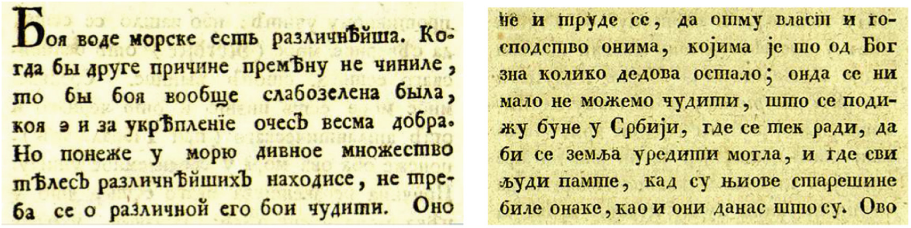
Fig. 7. Documents written in: (a) old Glagolitic and (b) new Glagolitic orthography

The second part of the experiment covers the documents written in Slavonic-Serbian and Serbian languages. A database of fifteen documents is analyzed. Five out of fifteen documents are written in Slavonic-Serbian language, while ten out of fifteen documents are written in modern Serbian language. The documents written in Slavonic-Serbian language are text excerpts from pages of the book entitled *Fisika* by Atanasije Stojković dated from 1802 [27]. Figure 8(a) shows a sample excerpt from the book *Fisika*. The rest of the ten documents is a collection of text excerpts from pages of the book *Zabavnik* dated from 1826 by Vuk Stefanović Karadžić. Figure 8(b) illustrates a sample excerpt from this book. These documents are written in so-called reformed Serbian language, i.e. modern one [26].

Historical Glagolitic books are available in their digital format at National Library of Zagreb[1]. On the contrary, serbian books are collected in their digital format at National Digital Library of Serbia[2]. Hence, documents have been selected from the digital books and processed in digital format.

[1] http://stari.nsk.hr/home.aspx?id=24.

[2] http://digitalna.nb.rs/.



Боя воде морске есть различнѣйша. Когда бы друге причине премѣну не чиниле, то бы боя вообще слабозелена была, коя э и за укрѣпленїе очесъ весма добра. Но понеже у морю дивное множество тѣлесъ различнѣйшихъ находисе, не треба се о различной его бои чудити. Оно

(a)

нѐ и труде се, да отму власти и господство онима, којима је то од Бог зна колико дедова остало; онда се ни мало не можемо чудити, што се подижу буне у Србији, где се тек ради, да би се земља уредити могла, и где сви људи памте, кад су њиове старешине биле онаке, као и они данас што су. Ово

(b)

Fig. 8. Documents written in (a) Slavonic-Serbian and (b) Serbian

4 Results and Discussion

Experiment has been performed in Matlab R2015a on a laptop computer with Quad-Core CPU at 2.2 GHz, 16 GB RAM and UNIX operating system. Our aim is to recognize the historical period of the digitalized printed documents.

Firstly, each document in the Serbian and Croatian databases is represented as a feature vector derived from the aforementioned feature coding. In particular, documents in Serbian database are represented as 27-dimensional feature vectors of run-length (the first 11 features) and ALBP (the last 16 features) statistics. Documents in Croatian database are given as 11-dimensional feature vectors of run-length statistics. In fact, we found that they perform considerably better than ALBP in the classification process of Glagolitic documents.

Secondly, Naive Bayes classifier is employed on each database for solving the binary recognition problem of documents as given in new or old orthography [25]. Because adopted run-length and ALBP features determine numerical values, the classifier uses the normal distribution for probabilities computation. Different classifiers, i.e. K-Nearest Neighbors (KNN) [1] and Support Vector Machine (SVM) [12] have also been tested on the databases for solving the classification task. Because Naive Bayes performs considerably better than the other classifiers, it has been definitively adopted in this context.

Document feature vectors for each database have been normalized before the classification process. In particular, we normalize every value x_i^k of k-th feature for the i-th document feature vector x_i in a certain database by using the following min-max approach:

$$\overline{x_i^k} = \frac{x_i^k - min_k}{max_k - min_k}, \tag{1}$$

where min_k is the minimum value of the k-th feature and max_k is the maximum value of the k-th feature in that database.

Confusion matrix for the binary classification problem (new and old orthography) is used for performance evaluation, from which precision, recall and f-measure are computed [11]. Evaluation is performed by dividing each database into training and test sets. To make the evaluation process independent from the selected training and test sets, we adopt the K-fold cross validation strategy [14]. Accordingly, performance measures are computed K times, and the average value together with the standard deviation are reported. In our case, the K value is fixed to 5 and 10.

Table 1 shows the classification results obtained by our approach on the database of Glagolitic documents using old and new orthography, and on the database of Slavonic-Serbian and Serbian documents. In the case of Glagolitic database, it is worth noting that f-measure is very high, reaching a value of 0.9333 for both classes in 5-fold and a peak of 0.9667 for new orthography recognition in 10-fold. In the case of Serbian database, our approach perfectly recognizes the new (modern Serbian) and old (Slavonic-Serbian) orthography in 5 and 10-folds, with an f-measure value of 1.

To confirm the efficacy of our method, a comparison is performed with bi-gram and tri-gram language models. In particular, documents in Slavonic-Serbian and Serbian languages are represented by bi-gram and tri-gram frequency vectors [28], on which the same Naive Bayes classifier is employed. Because Glagolitic documents are in the same language but in different script, comparison with bi-grams and tri-grams is not possible. It is a clear limitation of the methods which is overcome by our approach. Table 2 shows the classification results obtained by bi-gram and tri-gram language models. We can observe that our approach outperforms both the competitor methods in terms of f-measure. In fact, bi-gram language model performs poorly in old orthography (Slavonic-Serbian) recognition, with an f-measure of 0.20 in 5-fold and 0.60 in 10-fold. In any case, the f-measure is always below 0.90. On the other hand, tri-gram language model totally misses the old orthography (Slavonic-Serbian) recognition, with an f-measure value of 0 in 5-fold and 0.50 in 10-fold. Also, the f-measure value for new orthography (modern Serbian) recognition is always below 0.85. Again, standard deviation values are pretty high, demonstrating that the methods do not obtain a stable solution.

Table 1. Classification results of our method for old and new orthography recognition

Class		Glagolitic database			Serbian database		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
5-fold	*new orth.*	0.9500	1.0000	0.9714	0.9600	1.0000	0.9778
		(0.1118)	(0.0000)	(0.0639)	(0.0894)	(0.0000)	(0.0497)
	old orth.	1.0000	0.9000	0.9333	1.0000	0.9000	0.9333
		(0.0000)	(0.2236)	(0.1491)	(0.0000)	(0.2236)	(0.1491)
10-fold	*new orth.*	0.9667	1.0000	0.9800	0.9667	0.9500	0.9467
		(0.1054)	(0.0000)	(0.0632)	(0.1054)	(0.1581)	(0.1167)
	old orth.	0.9000	0.9000	0.9000	0.8500	0.9000	0.8667
		(0.3162)	(0.3162)	(0.3162)	(0.3375)	(0.3162)	(0.3220)

Table 2. Classification results of n-gram language model for old (Slavonic-Serbian) and new (modern Serbian) orthography recognition in Serbian language

		Bigram features			Trigram features		
Class		Precision	Recall	F-Measure	Precision	Recall	F-Measure
5-fold	*new orth.*	0.7500	0.6167	0.6500	0.9333	0.5667	0.7010
		(0.4330)	(0.4394)	(0.4183)	(0.1491)	(0.0913)	(0.0984)
	old orth.	0.0667	0.2000	0.1000	0.4000	0.6000	0.4800
		(0.1491)	(0.4472)	(0.2236)	(0.3651)	(0.5477)	(0.4382)
10-fold	*new orth.*	0.5500	0.5500	0.5500	0.6000	0.5000	0.5333
		(0.4972)	(0.4972)	(0.4972)	(0.4595)	(0.4082)	(0.4143)
	old orth.	0.2500	0.3000	0.2667	0.4500	0.5000	0.4667
		(0.4249)	(0.4830)	(0.4389)	(0.4972)	(0.5270)	(0.5018)

Finally, our method is computer time non-intensive, taking 1 s for processing a document of 2 K characters. Differently, bi-gram method takes 4 s and tri-gram method takes 5 s on the same document of 2 K characters. Hence, our method clearly showed its advantage.

5 Conclusions

This paper presented a new method for dating Serbian and Croatian documents from historical books by orthography recognition, overcoming the limitations of the current methods. It considered that text written in a script or language evolved through different historical periods is characterized by difference in orthography rules. Such a difference is captured by script coding, mapping the document characters to four numerical codes. If each code is considered as intensity level, codified text represents an image, subjected to texture analysis for feature extraction. Document features are classified by Naive Bayes approach for recognition of old or new orthography. Results on two databases of historical documents demonstrated that our method obtains very accurate results and very good performances when compared with other state-of-the-art methods.

Future work will extent the databases with a much larger collection of documents from multiple sources. Also, it will test the robustness of the method by using a set of noisy documents. Finally, it will provide a test set of documents from unknown sources for extending the experiment.

References

1. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
2. Baromic's Breviary, Venice (1493)

3. Berčić, I.: Foundations of the Old Slavic language written by Glagolitic scripts to read the church books, Prague (1862)
4. Biller, O., El-Sana, J., Kedem, K.: The influence of language orthographic characteristics on digital word recognition. In: The 11th IAPR International Workshop on Document Analysis Systems, Tours, pp. 131–135 (2014)
5. Brodić, D., Amelio, A., Milivojević, Z.N.: Clustering documents in evolving languages by image texture analysis. Appl. Intell. **46**(4), 916–933 (2017)
6. Brodić, D., Amelio, A., Milivojević, Z.N.: An approach to the language discrimination in different scripts using adjacent local binary pattern. J. Exp. Theor. Artif. Intell. **29**(5), 929–947 (2017)
7. Brodić, D., Amelio, A., Milivojević, Z.N.: Identification of Fraktur and Latin Scripts in German historical documents using image texture analysis. Appl. Artif. Intell. **30**(5), 379–395 (2016)
8. Brodić, D., Amelio, A., Milivojević, Z.N.: Language discrimination by texture analysis of the image corresponding to the text. Neural Comput. Appl., 1–21 (2016)
9. Brodić, D., Maluckov, Č.A., Milivojević, Z.N., Draganov, I.R.: Differentiation of the script using adjacent local binary patterns. In: Agre, G., Hitzler, P., Krisnadhi, A.A., Kuznetsov, S.O. (eds.) AIMSA 2014. LNCS (LNAI), vol. 8722, pp. 162–169. Springer, Cham (2014). doi:10.1007/978-3-319-10554-3_15
10. Chu, A., Sehgal, C.M., Greenleaf, J.F.: Use of gray value distribution of run lengths for texture analysis. Pattern Recogn. Lett. **11**(6), 415–419 (1990)
11. Confusion Matrix. http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html
12. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
13. Coulmas, F.: The Blackwell Encyclopedia of Writing Systems, p. 379. Blackwell, Oxford (1996)
14. Cross Validation (1997). https://www.cs.cmu.edu/~schneide/tut5/node42.html
15. Dasarathy, B.R., Holder, E.B.: Image characterizations based on joint gray-level run-length distributions. Pattern Recogn. Lett. **12**(8), 497–502 (1991)
16. Febvre, L., Martin, H.J.: The Coming of the Book: The Impact of Printing 1450–1800, Verso (1976)
17. Galloway, M.M.: Texture analysis using gray level run lengths. Comp. Graph. Im. Proc. **4**(2), 172–179 (1975)
18. Garrette, D., Alpert-Abrams, H.: An unsupervised model of orthographic variation for historical document transcription. In: The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, pp. 467–472 (2016)
19. Ivić, P.: Overview of History of the Serbian Language, Novi Sad (1998)
20. Lipovčan, S.: Discovering the Glagolitic Script of Croatia. Erasmus Publisher, Zagreb (2000)
21. Missale Romanum Glagolitice, Kosinje (1483)
22. Nosaka, R., Ohkawa, Y., Fukui, K.: Feature extraction based on co-occurrence of adjacent local binary patterns. In: Ho, Y.-S. (ed.) PSIVT 2011. LNCS, vol. 7088, pp. 82–91. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25346-1_8
23. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
24. Reffle, U., Ringlstetter, C.: Unsupervised profiling of OCRed historical documents. Pattern Recogn. **46**, 1346–1357 (2013)

25. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Prentice Hall, Egnlewood Cliffs (1995, 2003)
26. Stefanović Karadžić, V.: Građa za Srpsku Istoriju našega vremena. Štamparija Kraljevskog Univerziteta, Budim (1828)
27. Stojković, A.: Fisika. Štamparija Kraljevskog Univerziteta, Budim (1803)
28. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**(1), 141–188 (2010)
29. Zramdini, A., Ingold, R.: Optical font recognition using typographical features. IEEE Trans. Pattern Anal. Mach. Intell. **8**(20), 877–882 (1998)