Soft Comput (2015) 19:2655–2665
DOI 10.1007/s00500-014-1435-1

METHODOLOGIES AND APPLICATION

An approach to the script discrimination in the Slavic documents

Script discrimination

Darko Brodić · Zoran N. Milivojević ·
Čedomir A. Maluckov

Published online: 28 August 2014

Abstract The paper deals with the problem of the script discrimination in old Slavic printed documents. Therefore, an algorithm for script classification and identification is proposed. It creates coded text from initial document. Then, the coded text is subjected to statistical analysis. As a result, the texture feature extraction is carried out. Obtained texture features are used as criteria for script classification and identification. The proposed method is tested on the samples of old Slavic printed documents written in Glagolitic, Cyrillic and Latin script.

Keywords Coding · Script recognition · Optical character recognition · Statistical analysis

1 Introduction

Optical character recognition (OCR) is a computer-based system that recognizes printed characters by scanning the original text document image. Basically, it consists of the following stages: (1) preprocessing, (2) feature extraction with classification and (3) post-processing (Ghosh et al. 2010).

Script recognition represents the part of OCR included in the feature extraction with classification stage, a very important part of document image analysis (Ghosh et al. 2010). Different methods have been developed for the script recognition task. They are classified as global and local methods.

Global methods consider wider blocks in document images, which are subjected to the statistical and frequency-domain analysis (Joshi et al. 2007). To extend the effectiveness of the method, document image blocks have to be normalized. Furthermore, the image should be free of noise and high quality (Busch et al. 2006).

Local methods separate small pieces of text as connected components. Connected components contain characters, words or lines. After that, the analysis of different features, like for example the black pixel runs, is carried out (Pal and Chaudhury 2002). Local methods are suitable for low-quality and noisy documents, however, they are computationally expensive.

Textures are the image features that can be described according to their spatial, frequency and perceptual properties (Del Bimbo 2001; Tolambiya et al. 2010). An effective representation of textures can be based on statistical and structural properties of brightness patterns (Yang and Purves 2004). Texture can be measured by taking into account the spatial arrangement of gray-level primitives (Haralick 1979). Hence, the major statistical method used in texture analysis is based on the definition of the joint probability distributions of pairs of pixels (Valkealahti and Oja 1998). Texture analysis can be very helpful in cases where image objects are characterized more by their texture than by the intensity (Zhang and Tan 2002; Bharati et al. 2004; Eleyan and Demirel 2011).

This paper proposes a script recognition module. The object of research is the Slavic documents. These types of documents were chosen because they can be written in three different scripts: Latin, Glagolitic and Cyrillic. Consequently, their differentiation is a challenging task, and a new approach is introduced herein. Furthermore, our approach unites local and global methods. First, it treats the charac-

Communicated by V. Loia.

D. Brodić (✉) · Č. A. Maluckov
Technical Faculty in Bor, V.J. 12, University of Belgrade,
19210 Bor, Serbia
e-mail: dbrodic@tf.bor.ac.rs

Z. N. Milivojević
College of Applied Technical Sciences, Aleksandra Medvedeva 20,
18000 Niš, Serbia

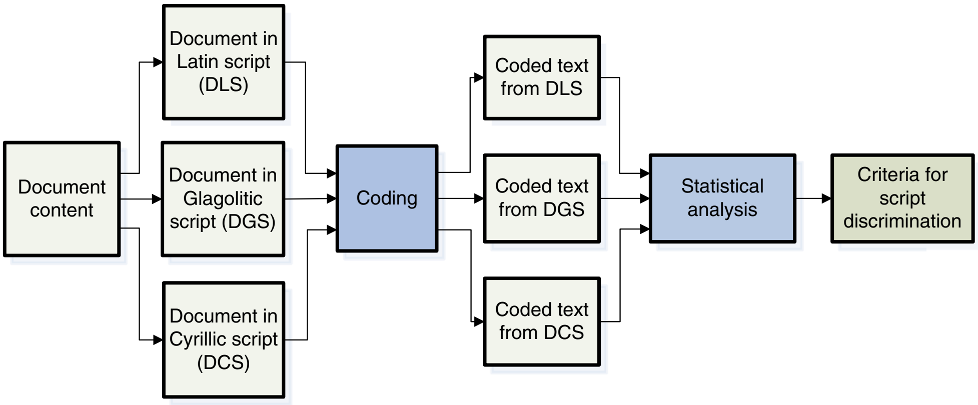

Document content
Document in Latin script (DLS)
Document in Glagolitic script (DGS)
Document in Cyrillic script (DCS)
Coding
Coded text from DLS
Coded text from DGS
Coded text from DCS
Statistical analysis
Criteria for script discrimination

Fig. 1 Structural diagram of the algorithm flow

ters in text, which is the manner of local methods. It maps each character into the corresponding script type according to its position in the text line. This way, the number of variables is significantly reduced. The result of this step is coded text (Brodić et al. 2013). Second, the script type distribution of coded text is analyzed. As a result, four script features are extracted. Then, the text is subjected to textural analysis, obtaining the gray-level co-occurrence matrix (GLCM), which is used to extract the texture features (Haralick et al. 1973) needed for classification. The textural analysis is a typical step of global methods. Finally, a discrimination function is established according to the comprehensive texture feature classification, representing criteria for script discrimination and identification.

The paper is organized as follows. Section 2 addresses all aspects concerning the proposed algorithm; it includes the coding, script distribution and co-occurrence analysis, feature extraction and establishment of criteria for script discrimination. Section 3 defines the custom oriented database of documents written in different scripts. Then, it explains the experiment that evaluates the proposed algorithm. Section 4 gives the results of the experiment and discusses them. Section 5 concludes the paper.

2 Proposed algorithm

The proposed algorithm consists of the following stages: (1) coding, (2) statistical analysis, and (3) determination of criteria for script discrimination. However, it can be divided into many sub-stages. Figure 1 illustrates the structural diagram of the algorithm flow.

2.1 Coding

Each text line can be split into three vertical zones: (1) upper, (2) middle and (3) lower (Zramdini and Ingold 1998; Chaudhuri et al. 2002). The letters in a certain script have different positions in the text line. Base letters (B), like the letter **a**, occupy a middle zone only; ascender letters (A), like the letter **h**, spread over the middle and upper zones; while descendent letters (D), like the letter **g**, include the middle and lower zones. Few letters like the capital letter **Lj** (in Serbian or Croatian Latin alphabet) comprise all three zones. They are classified as a full letter (F). Table 1 shows the script type classification.

Table 1 Script type classification

Script example	Script type (ST)	Identification (I)	Coding (C)
a, љ, ⰿ	Base	B	0
B, Љ, ⱅ	Ascender	A	1
g, y, ⱏ	Descender	D	2
Lj, ђ, ⱑ	Full	F	3

This way, the letters from Serbian, or Croatian Latin alphabet, Serbian Cyrillic alphabet and Croatian Glagolitic alphabet are mapped into the elements from the identification set I:

$$I = \{\text{B, A, D, F}\}. \qquad (1)$$

Furthermore, set I is coded to set C to effectively perform the statistical analysis (Brodić et al. 2013, 2014):

$$C = \{0, 1, 2, 3\}, \qquad (2)$$

where B → 0, A → 1, D → 2, and F → 3. Table 2 shows Latin, Glagolitic and Cyrillic letters as well as theirs codes according to Table 1 (slightly adapted to current Croatian or Serbian language) (Brodić et al. 2014).

The proposed algorithm replaces all letters from a certain script with the equivalent member of the set C by coding as in Table 2. This way, an initial text is converted into the coded text.

Table 2 Coding of Slavic alphabets

Glagolitic	Coding	Latin	Coding	Cyrillic	Coding
ⰎⰉ	2	Lj	3	Љ	1
ⰾⰹ	1	lj	3	љ	0
ⱀⰉ	2	Nj	3	Њ	1
ⱀⰹ	0	nj	3	њ	0
Ⰵ	2	E	1	Е	1
ⰵ	0	e	0	е	0
Ⱃ	2	R	1	Р	1
ⱃ	0	r	0	р	2
Ⱅ	2	T	1	Т	1
ⱅ	0	t	1	т	0
Ⰸ	2	Z	1	З	1
ⰸ	1	z	0	з	0
Ⱆ	2	U	1	У	1
ⱆ	0	u	0	у	2
Ⰹ	2	I	1	И	1
ⰹ	0	i	1	и	0
Ⱁ	2	O	1	О	1
ⱁ	0	o	0	о	0
Ⱂ	2	P	1	П	1
ⱂ	2	p	2	п	0
Ⱎ	2	Š	1	Ш	1
ⱎ	0	š	1	ш	0
Ⰼ	2	Đ	1	Ђ	1
ⰼ	0	đ	1	ђ	3
Ⰰ	2	A	1	А	1
ⰰ	1	a	0	а	0
Ⱄ	2	S	1	С	1
ⱄ	0	s	0	с	0
Ⰴ	2	D	1	Д	1
ⰴ	0	d	1	д	2
Ⱇ	2	F	1	Ф	1
ⱇ	2	f	1	ф	3
Ⰳ	2	G	1	Г	1
ⰳ	0	g	2	г	0
Ⱈ	2	H	1	Х	1
ⱈ	0	h	1	х	0
Ⰼ	2	J	1	J	1
ⰼ	0	j	3	j	3
Ⰽ	2	K	1	К	1
ⰽ	0	k	1	к	0
Ⰾ	2	L	1	Л	1
ⰾ	1	l	1	л	0
Ⱍ	2	Č	1	Ч	1
ⱍ	1	č	1	ч	0
Ⱋ	2	Ć	1	Ћ	1
ⱋ	1	ć	1	ћ	1
Ⰶ	2	Ž	1	Ж	1
ⰶ	1	ž	1	ж	0
ⰄⰆ	3	Dž	1	Џ	3
ⰴⰶ	1	dž	1	џ	2
Ⱌ	2	C	1	Ц	3
ⱌ	0	c	0	ц	2
Ⰲ	2	V	1	В	1
ⰲ	0	v	0	в	0
Ⰱ	2	B	1	Б	1
ⰱ	0	b	1	б	1
Ⱀ	2	N	1	Н	1
ⱀ	0	n	0	н	0
Ⰿ	2	M	1	М	1
ⰿ	0	m	0	м	0
Ⱑ	2	Ja, (I)je	-	Ja, (И)je	-
ⱑ	0	ja, (i)je	-	ja, (и)je	-

2.2 Statistical analysis

Statistical analysis is divided into two parts: script type distribution (Brodić et al. 2013) and co-occurrence (Brodić et al. 2013, 2014). In both analyses the input is coded text, which subjected to the statistical analysis. Figure 2 illustrates the same text written in different Slavic scripts along with their coding.

2.2.1 Script type distribution

First, the script type distribution of coded text is analyzed. As a result, four script features are extracted. Table 3 shows these features, which are obtained from the same text written in different Slavic scripts (see Fig. 2 for reference).

The script type distribution of Latin, Glagolitic and Cyrillic script is given in Fig. 3a–c, respectively.

Glagolitic script has the highest distribution of base script type, then follows Cyrillic script, while Latin script has the smallest distribution of base script type. Latin script has the highest distribution of ascending script type, Glagolitic script has slightly smaller distribution, while the Cyrillic script has a considerably lower distribution of ascending script type. Cyrillic script has the highest distribution of descending script type. Glagolitic and Latin scripts have a substantially lower distribution of descending script type. Latin and Cyrillic scripts have similar distributions of full script type, while the Glagolitic script has weak or even no distribution of full script type.

2.2.2 Co-occurrence analysis

Currently, the coded text is subjected to co-occurrence analysis (Haralick et al. 1973; Clausi 2002) to extract the texture features. This approach generates texture features of image according to calculated co-occurrence probabilities. These probabilities represent the conditional joint probabilities of all pair-wise combinations of gray levels in the spatial window of interest (WOI). WOI is determined by the inter-pixel distance (d) and orientation (θ) (Haralick et al. 1973; Clausi 2002). Figure 4 shows an illustration of WOI.

The following parameters are considered to describe the image with GLCM: (1) the number of gray levels, (2) the orientation angle (θ) and (3) the length of displacement (d). In our case, the codes are considered as the gray levels.

The method starts in the top left corner and counts the occurrences of each reference pixel to neighbor pixel relationship. This way, each element (i, j) of GLCM represents the sum of the number of times the pixel with the value i is located at some distance d and angle θ from the pixel of intensity j. At the end of this process, the element (i, j) gives

Čuvaj uši svoje da slušaju samo svete i časne razgovore, a ne ružne i svjetovne, jer je napisano: Načini oko
svojih ušiju živicu od trnja i ne slušaj jezik pakostan.

(a)

```
1 0 0 0 3 0 1 1 0 0 0 3 0 1 0 0 1 0 1 0 3 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0
2 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 3 0 1 0 0 0 0 3 0 0 3 0 0 0 2 1 0 0 0 0 1 0 1
1 0 1 0 1 0 0 0 0 3 1 1 0 1 1 3 0 1 1 0 1 0 0 0 1 1 0 3 0 1 0 0 0 1 0 1 0 3 3 0
0 1 1 2 0 1 0 0 1 0 0
```

(b)

Ⱍⱆⰲⰰⰼ ⱆⱎⰻ ⱄⰲⱁⰼⰵ ⰴⰰ ⱄⰾⱆⱎⰰⰼⱆ ⱄⰰⰿⱁ ⱄⰲⰵⱅⰵ ⰻ ⱍⰰⱄⱀⰵ ⱃⰰⰸⰳⱁⰲⱁⱃⰵ, ⰰ ⱀⰵ
ⱃⱆⰶⱀⰵ ⰻ ⱄⰲⰼⰵⱅⱁⰲⱀⰵ, ⰼⰵⱃ ⰼⰵ ⱀⰰⱂⰻⱄⰰⱀⱁ: Ⱀⰰⱍⰻⱀⰻ ⱁⰽⱁ ⱄⰲⱁⰼⰻⱈ ⱆⱎⰻⰼⱆ ⰶⰻⰲⰻⱌⱆ
ⱁⰴ ⱅⱃⱀⰼⰰ ⰻ ⱀⰵ ⱄⰾⱆⱎⰰⰼ ⰼⰵⰸⰻⰽ ⱂⰰⰽⱁⱄⱅⰰⱀ.

(c)

```
2 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1
0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0 1 0 0 2 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0
1 0 0 2 1 0 0 0 0 1 0
```

(d)

Чувај уши своје да слушају само свете и часне разговоре, а не ружне и свјетовне, јер је написано:
Начини око својих ушију живицу од трња и не слушај језик пакостан.

(e)

```
1 2 0 0 3 2 0 0 0 0 0 3 0 2 0 0 0 2 0 0 3 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0
0 0 0 0 2 0 0 0 0 2 2 0 0 0 0 0 0 3 0 0 0 0 0 0 3 0 2 3 0 0 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 0 3 0 0 2 0 0 3 2 0 0 0 0 2 2 0 2 0 2 0 0 0 0 0 0 0 2 0 0 3 3 0
0 0 0 0 0 0 0 0 0 0 0
```

(f)

Fig. 2 Same text given in different scripts: **a** original text in Latin script, and **b** its coded counterpart; **c** original text in Glagolitic script, and **d** its coded counterpart; **e** original text in Cyrillic script, and **f** its coded counterpart

Table 3 Comparison of the script type distributions between scripts

Script type	Latin script	Glagolitic script	Cyrillic script
Base (B)	0.6336	0.8015	0.7786
Ascender (A)	0.2595	0.1679	0.0153
Descender (D)	0.0229	0.0305	0.1298
Full (F)	0.0840	0.0000	0.0763

the number of how many times the gray levels i and j appears as a sequence of two pixels located at a defined distance d along a chosen direction θ. The GLCM for an image **I** with M rows and N columns is parameterized by the offset (Δx, Δy) as (Eleyan and Demirel 2011):

$$P(i,j) = \sum_{x=1}^{M}\sum_{y=1}^{N}\begin{cases}1, \text{if } I(x,y)=i\ , & I(x+\Delta x, y+\Delta y)=j\\ 0, \text{otherwise}\end{cases} \tag{3}$$

The offset (Δx, Δy) represents the pixel displacement $d$and the orientation θ at which GLCM is calculated. In our example, the input represents the coded text given as a 1D image. Accordingly, the feasible values of the parameters d and θ are narrowed to $d = 1$ and $\theta = 0°$. Consequently, the number of gray levels G of a coded text is mapped to 4 (from 0 to 3) (Brodić et al. 2013, 2014).

The normalized probability version of the GLCM is given as:

$$C(i,j) = P(i,j) / \sum_{i,j}^{G} P(i,j). \tag{4}$$

To characterize different scripts, the same text written with different scripts (see Fig. 2 for reference) is subjected to the co-occurrence analysis. Figure 5 shows the normalized probability GLCM for each script.

Furthermore, the number of texture features can be extracted from the GLCM (Haralick et al. 1973; Clausi 2002). Unlike ref. Brodić et al. (2013, 2014), the eight texture fea-

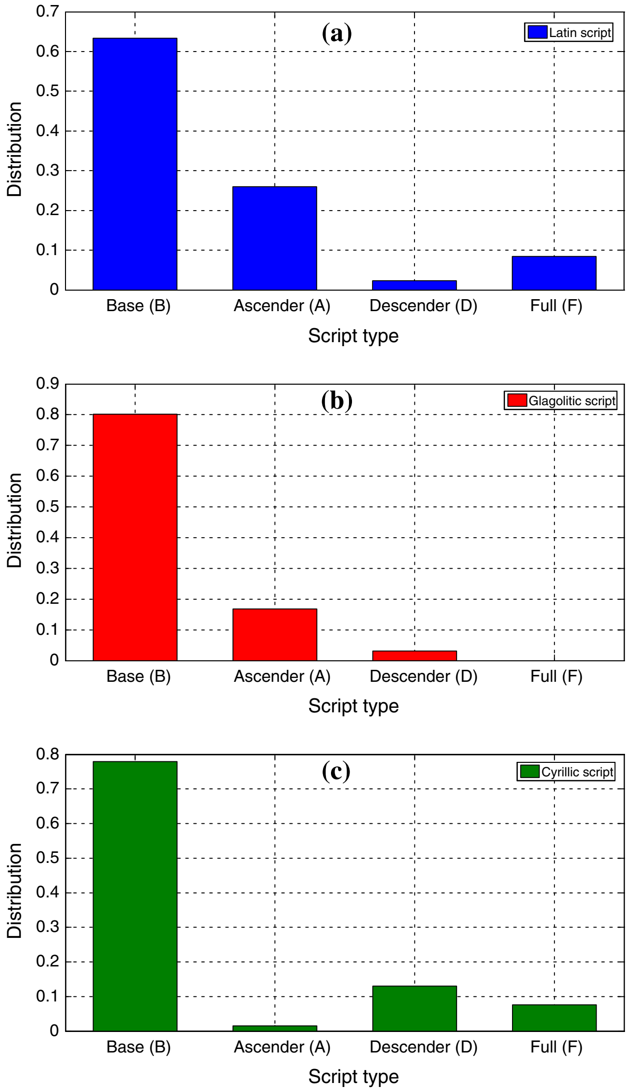


Fig. 3 Script type distribution: **a** Latin script, **b** Glagolitic script, and **c** Cyrillic script

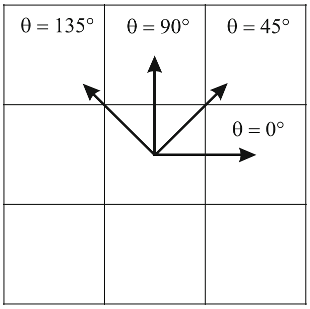


Fig. 4 WOI for the calculation of texture features, considering $d = 1$ and different directions

tures among the fourteen proposed in Haralick et al. (1973) is used. Their definition is given in Table 4.

Table 5 shows typical eight GLCM texture feature measures obtained from the same text written in different Slavic scripts (see Fig. 2 for reference).

2.3 Criteria for the script discrimination

Each script is characterized with its own set of specific features (mainly typographical). The statistical analysis of coded text is used to extract them. It starts with the script type distribution analysis followed by the co-occurrence analysis. The statistical analysis is enlarged compared to those given in Brodić et al. (2013, 2014) by including a bigger set of extracted texture features. Furthermore, it is not used for script characterization only (same document written by different scripts), but for script identification as well (different document written by different scripts). As an extension to the previous method (Brodić et al. 2014), the enlarged feature vector given by four script type distribution measures and eight GLCM texture measures is used. The proposed approach compared to previous ones (Brodić et al. 2013, 2014) contributes to increased validity in order to establish criteria for script discrimination based on thresholding decision making. Accordingly, the statistical analysis shows the clear difference between scripts.

3 Experiments

The experiment is determined to evaluate the quality of the proposed algorithm. The custom oriented database of documents similar to those given in http://www.croatianhistory.net/etf/juraj_slovinac_misli.html, http://www.croatianhistory.net/etf/badurina_parcic.html is created. It comprises the "training" and "test" set (Silva and Ribeiro 2007). Training set consists of total 130 documents, which includes at least 40 documents written in each script. Typical length of text is from approx. 300 to 3,000 characters. Test set consists of 10 documents written in each script, i.e., total of 30 documents. Typical length of text is from approx. 500 to 4,000 characters. Texts are extracted from the book "Le château de virginité" ("The Castle of Virginity") written in 1411 by George d'Esclavonie (Juraj Slovinac) (http://www.croatianhistory.net/etf/juraj_slovinac_misli.html). He was a Croatian Glagolitic priest and professor at Sorbonne in Paris around 1400. Figure 6 illustrates sample documents from the database written in different Slavic scripts.

4 Results and discussion

4.1 Results of the script type distribution

The script type distributions are used to extract four script features, which are used to characterize different scripts. To

0.3692	0.1769	0.0154	0.0692
0.1846	0.0615	0.0077	0.0077
0.0154	0.0077	0	0
0.0692	0.0077	0	0.0077

(a)

0.6538	0	0.1308	0.0154
0	0	0	0
0.1385	0	0.0231	0.0077
0.0154	0	0.0154	0

(b)

0.6231	0.0077	0.0846	0.0615
0.0077	0	0.0077	0
0.1077	0	0.0154	0.0077
0.0462	0	0.0231	0.0077

(c)

Fig. 5 GLCM for the coded text from Fig. 2: **a** Latin, **b** Glagolitic, and **c** Cyrillic

Table 4 GLCM texture feature definition

Feature	Definition
Energy	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)^2$
Entropy	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)\cdot\log C(i,j)$
Maximum	$\max\sum_{i}^{G}\sum_{j}^{G} C(i,j) \quad \forall i,j$
Dissimilarity	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)\cdot\lvert i-j\rvert$
Contrast	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)\cdot(i-j)^2$
Inverse different moment	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)/[1+(i-j)^2]$
Homogeneity	$\sum_{i}^{G}\sum_{j}^{G} C(i,j)/[1+(i-j)]$
Correlation	$\sum_{i}^{G}\sum_{j}^{G} (i-\mu_x)\cdot(j-\mu_y)\cdot C(i,j)/(\sigma_x\cdot\sigma_y)$

quantify the obtained results, we used the minimum and maximum values. Tables 6, 7 show the distributions, which are obtained from the training and test set.

Figure 7 shows the script type distributions for training set—a, c, e, g, and test set—b, d, f, h.

From the training set, we can establish the script discrimination relation:

```
IF ((B < 0.65) AND (A > 0.26))
   Writeln('Latin Text')
ELSEIF ((A < 0.16) AND (F > 0))
   Writeln('Cyrillic Text')
ELSE
  Writeln('Glagolitic Text')
END
```

Test set confirms the previously established script discrimination relation.

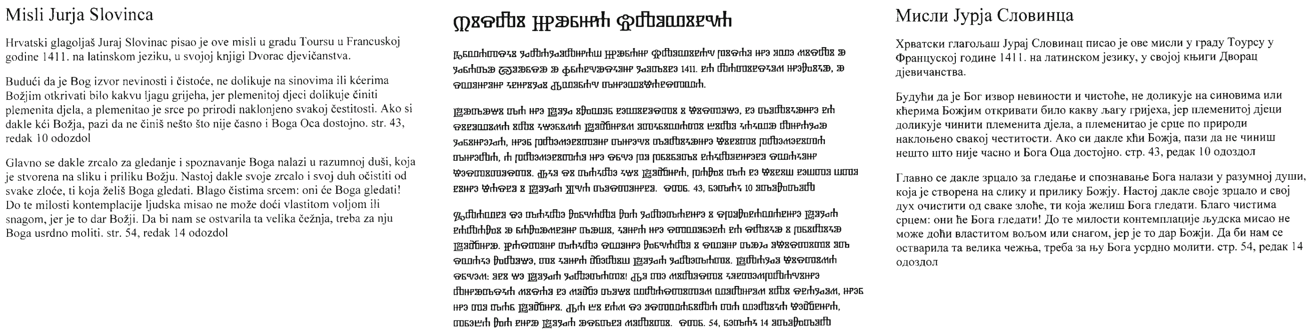
4.2 GLCM feature results

The extended set of eight GLCM texture features is used (compared to Brodić et al. 2013, 2014) as a basis to discriminate different scripts. To quantify the obtained results, we have used the minimum and maximum values. The texture features obtained from a statistical analysis of database texts written in Latin, Glagolitic and Cyrillic script for training and test set are shown in Tables 8, 9.

It should be noted that the values of entropy, inverse different moment and homogeneity are quite similar among scripts. Consequently, these features will be discarded from further discussion. In the further analysis, another five texture features are of the interest. Figure 8 shows the minimum and

Table 5 GLCM texture feature measures

Feature	Script					
	Latin	Glagolitic	Cyrillic	Latin/Glagolitic	Latin/Cyrillic	Glagolitic/Cyrillic
Energy	0.2159	0.4651	0.4140	<1	<1	≈
Entropy	−1.8432	−1.1347	−1.3957	>1	>1	<1
Maximum	0.3692	0.6538	0.6231	<1	<1	≈
Dissimilarity	0.8846	0.6540	0.7615	>1	>1	<1
Contrast	1.8077	1.3769	1.7923	>1	≈	<1
Inverse different moment	0.6500	0.7454	0.7223	<1	<1	≈
Homogeneity	0.6769	0.7859	0.7641	<1	<1	≈
Correlation	−0.1291	0.0742	0.0791	<0	<0	≈



Misli Jurja Slovinca

Hrvatski glagoljaš Juraj Slovinac pisao je ove misli u gradu Toursu u Francuskoj godine 1411. na latinskom jeziku, u svojoj knjigi Dvorac djevičanstva.

Budući da je Bog izvor nevinosti i čistoće, ne dolikuje na sinovima ili kćerima Božjim otkrivati bilo kakvu ljagu grijeha, jer plemenitoj djeci dolikuje činiti plemenita djela, a plemenitao je srce po prirodi naklonjeno svakoj čestitosti. Ako si dakle kći Božja, pazi da ne činiš nešto što nije časno i Boga Oca dostojno. str. 43, redak 10 odozdol

Glavno se dakle zrcalo za gledanje i spoznavanje Boga nalazi u razumnoj duši, koja je stvorena na sliku i priliku Božju. Nastoj dakle svoje zrcalo i svoj duh očistiti od svake zloće, ti koja želiš Boga gledati. Blago čistima srcem: oni će Boga gledati! Do te milosti kontemplacije ljudska misao ne može doći vlastitom voljom ili snagom, jer je to dar Božji. Da bi nam se ostvarila ta velika čežnja, treba za nju Boga usrdno moliti. str. 54, redak 14 odozdol

(a)

(b)

Мисли Јурја Словинца

Хрватски глагољаш Јурај Словинац писао је ове мисли у граду Тоурсу у Француској године 1411. на латинском језику, у својој књиги Дворац дјевичанства.

Будући да је Бог извор невиности и чистоће, не доликује на синовима или кћерима Божјим откривати било какву љагу гријеха, јер племенитој дјеци доликује чинити племенита дјела, а племенитао је срце по природи наклоњено свакојеститости. Ако си дакле кћи Божја, пази да не чиниш нешто што није часно и Бога Оца достојно. стр. 43, редак 10 одоздол

Главно се дакле зрцало за гледање и спознавање Бога налази у разумној души, која је створена на слику и прилику Божју. Настој дакле своје зрцало и свој дух очистити од сваке злоће, ти која желиш Бога гледати. Благо чистима срцем: они ће Бога гледати! До те милости контемплације људска мисао не може доћи властитом вољом или снагом, јер је то дар Божји. Да би нам се остварила та велика чежња, треба за њу Бога усрдно молити. стр. 54, редак 14 одоздол

(c)

Fig. 6 Sample documents from database: **a** Latin, **b** Glagolitic, and **c** Cyrillic

Table 6 Comparison of the script type distributions between scripts (training set)

Script type	Latin script		Glagolitic script		Cyrillic script	
	min	max	min	max	min	max
B	0.48	0.62	0.68	0.79	0.68	0.85
A	0.28	0.44	0.16	0.24	0.03	0.16
D	0.03	0.07	0.03	0.17	0.07	0.16
F	0.01	0.08	0.00	0.00	0.01	0.07

Table 7 Comparison of the script type distributions between scripts (test set)

Script type	Latin script		Glagolitic script		Cyrillic script	
	min	max	min	max	min	max
B	0.58	0.60	0.75	0.79	0.77	0.81
A	0.30	0.34	0.17	0.20	0.04	0.06
D	0.04	0.06	0.04	0.06	0.11	0.16
F	0.04	0.06	0.00	0.00	0.03	0.05

maximum values of the energy (Latin, Cyrillic and Glagolitic script).

The energy value of 0.3 differentiates the Latin from the other two scripts. Figure 9 shows the the minimum and maximum values of the GLCM maximum (Latin, Cyrillic and Glagolitic script).

Similarly, Latin (from the other two scripts) can be separated by the maximum value of 0.45. Figure 10 shows the minimum and maximum values of the dissimilarity (Latin, Cyrillic and Glagolitic script).

Currently, Cyrillic can be extracted from the other scripts by a dissimilarity value of 0.77 (test set only). Figure 11 shows the minimum and maximum values of the contrast of Latin, Cyrillic and Glagolitic script.

Glagolitic can be distinguished from the other scripts by contrast value of 1.7 (test set only). Figure 12 shows the minimum and maximum values of the correlation (Latin, Cyrillic and Glagolitic script).

Latin can be differentiated from the other scripts by setting the correlation value to −0.15.

Taking into account all aforementioned features, i.e., energy, maximum, correlation, dissimilarity and contrast, we can establish the discrimination criteria that can be used for script recognition in the Slavic documents (test set only). The criteria are given by the following pseudo-code:

```
IF ((Energy < 0.3) AND (Entropy < -1.65) AND (Maximum < 0.45) AND
   (Correlation < -0.15))
        Writeln('Latin Text')
ELSEIF ((Dissimilarity < 0.7) AND (Contrast < 1.7))
        Writeln('Cyrillic Text')
ELSE
        Writeln(Glagolitic Text')
END
```

Although, the aforementioned features show significant variation among scripts, they are valid for the test set only. To establish generalized criteria for script discrimination, we should use the broader information, i.e., those obtained from the training set. In this way, the results from the script type distribution have to be included as well. The extended criteria can be expressed by the following pseudo-code:

```
IF ((Energy < 0.25) AND (Entropy < -1.7) AND (Maximum < 0.45) AND
   (Correlation < -0.15) AND (B < 0.65) AND (A > 0.26))
        Writeln('Latin Text')
ELSEIF ((A < 0.16) AND (F > 0))
        Writeln('Cyrillic Text')
ELSE
        Writeln(Glagolitic Text')
END
```

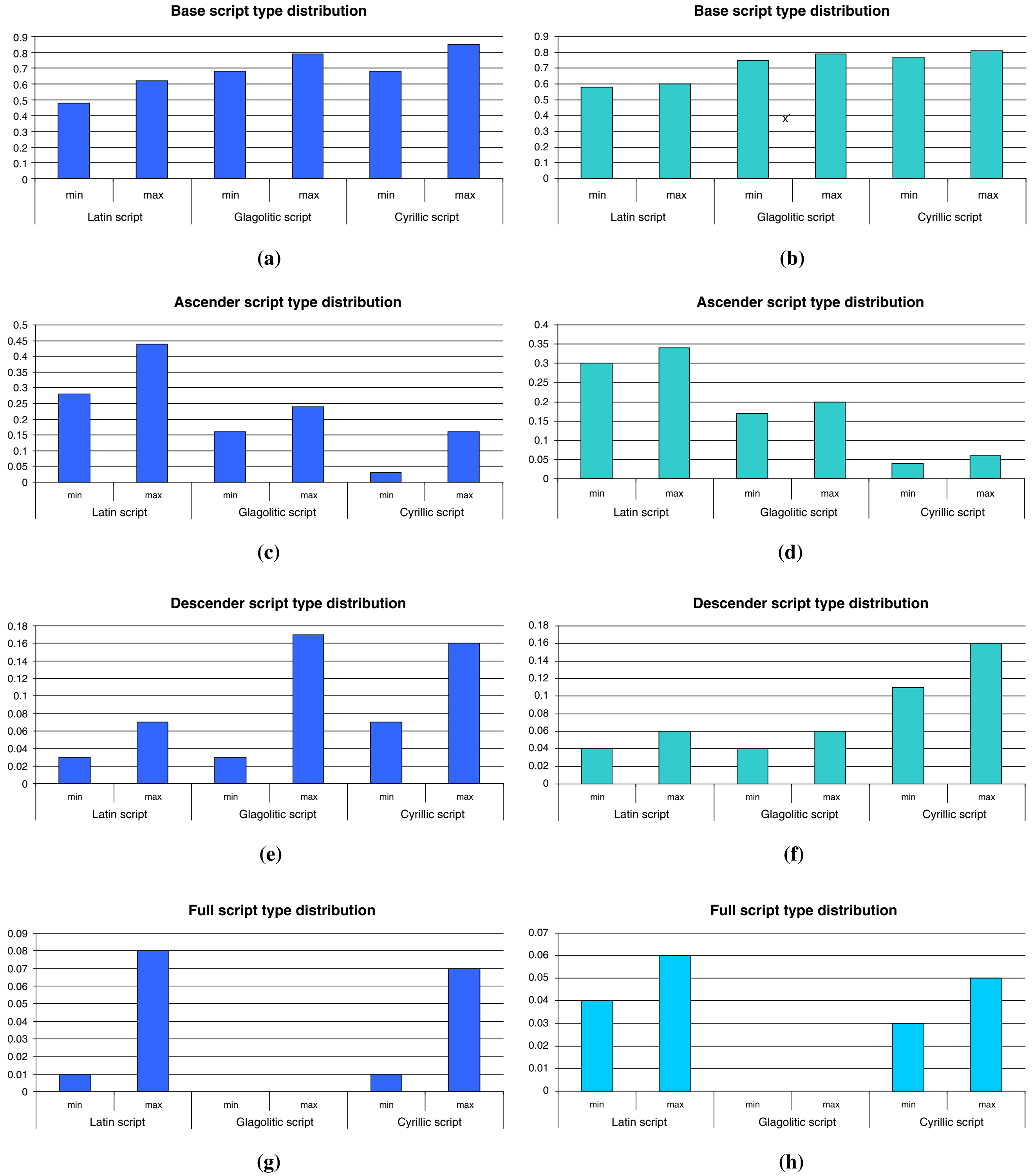


Fig. 7 Script type distributions: **a** base for training set, **b** base for test set, **c** ascender for training set, **d** ascender for test set, **e** descendent for training set, **f** descendent for test set, **g** full for training set, **h** full for test set

The above criteria can be used to effectively discriminate certain script, i.e., Latin, Cyrillic and/or Glagolitic script. The presented concept recognizes the scripts in the document without errors. However, it can be noted that the established concept is based on the ideal conditions. To prove their validity in real circumstances, their effectiveness should be evaluated by incorporating in an OCR system.

Table 8 GLCM texture feature measures (training set)

Feature	Script					
	Latin		Glagolitic		Cyrillic	
	min	max	min	max	min	max
Energy	0.1667	0.2141	0.3086	0.4318	0.3247	0.5067
Entropy	−2.0266	−1.7570	−1.4991	−1.2105	−1.6472	−1.1726
Maximum	0.2374	0.3546	0.4973	0.6261	0.5389	0.7014
Dissimilarity	0.7569	0.9864	0.6836	0.9813	0.5953	0.8712
Contrast	1.0761	1.9783	1.5201	2.2433	1.2273	2.0215
Inverse different moment	0.6044	0.6793	0.6356	0.7418	0.6794	0.7798
Homogeneity	0.6365	0.7004	0.6943	0.7835	0.7271	0.8098
Correlation	−0.2430	−0.1590	−0.1309	0.4804	−0.1183	0.0754

Table 9 GLCM texture feature measures (test set)

Feature	Script					
	Latin		Glagolitic		Cyrillic	
	min	max	min	max	min	max
Energy	0.1863	0.2036	0.3684	0.4211	0.3786	0.4400
Entropy	−1.9375	−1.8527	−1.3731	−1.2487	−1.4686	−1.3479
Maximum	0.2985	0.3173	0.5664	0.6174	0.5861	0.6467
Dissimilarity	0.7838	0.8938	0.7659	0.8633	0.6774	0.7714
Contrast	1.3762	1.6756	1.7217	1.9461	1.4486	1.6597
Inverse different moment	0.6253	0.6673	0.6766	0.7126	0.7032	0.7436
Homogeneity	0.6509	0.6889	0.7281	0.7591	0.7459	0.7795
Correlation	−0.2032	−0.1597	−0.0418	0.0044	−0.0844	−0.0118

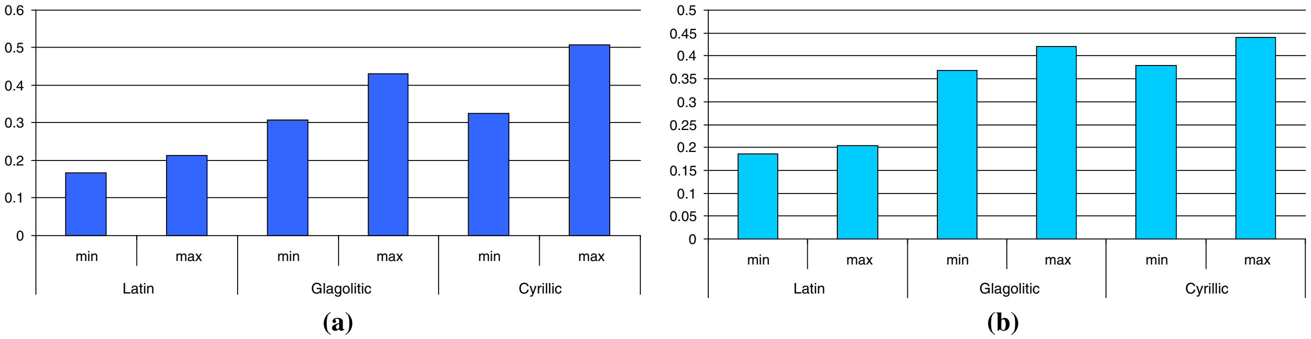


Fig. 8 The energy of Latin, Cyrillic and Glagolitic script: **a** training set, **b** test set

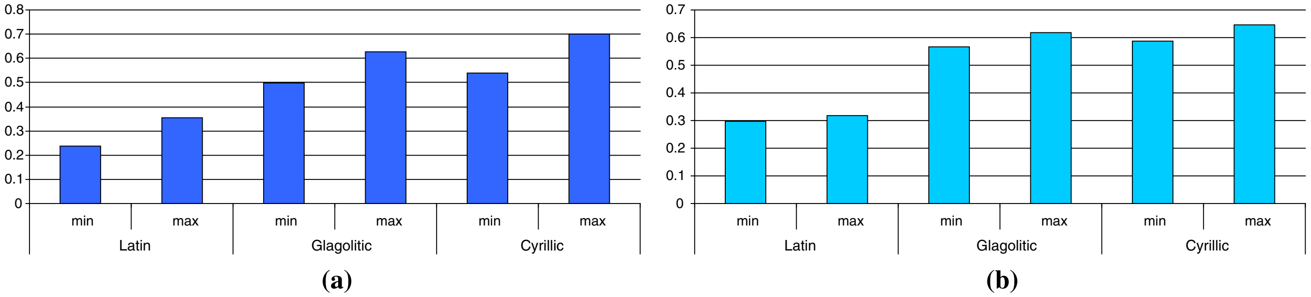


Fig. 9 The maximum of Latin, Cyrillic and Glagolitic script: **a** training set, **b** test set

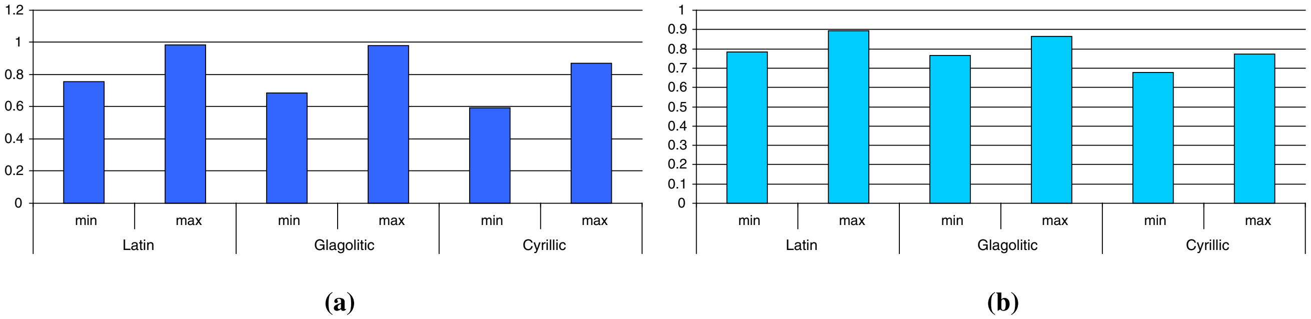


Fig. 10 The dissimilarity of Latin, Cyrillic and Glagolitic script: **a** training set, **b** test set

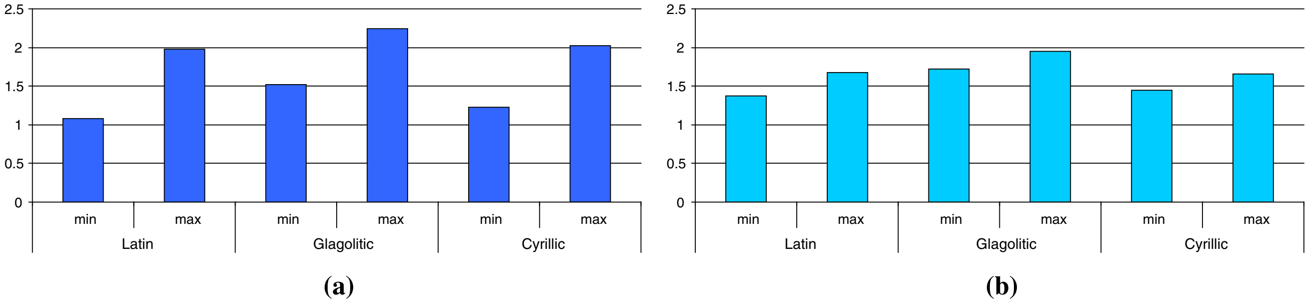


Fig. 11 The contrast of Latin, Cyrillic and Glagolitic script: **a** training set, **b** test set

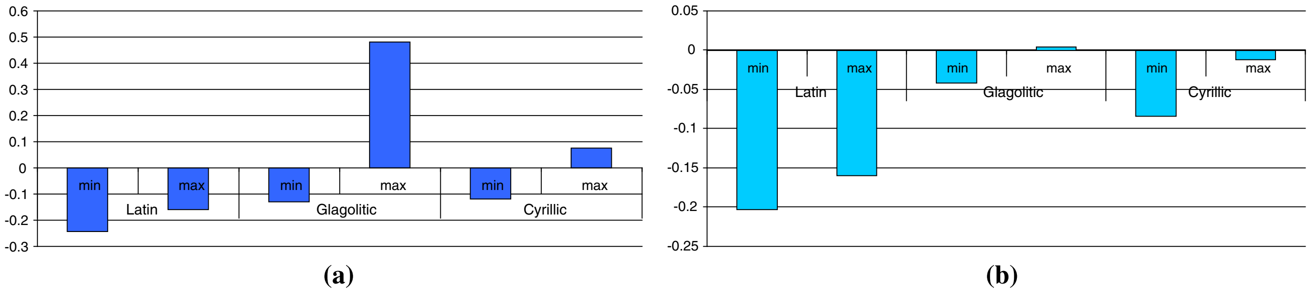


Fig. 12 The Correlation of Latin, Cyrillic and Glagolitic script: **a** training set, **b** test set

5 Conclusions

This manuscript proposed an algorithm for the script characterization and identification in Slavic documents written in Latin, Glagolitic and Cyrillic scripts. The algorithm accompanies the statistical analysis of the coded text. It is obtained by coding text from document according to the baseline status of each letter. The statistical analysis is performed with the script type distribution and co-occurrence analysis of the coded text. As the results, four script type features and eight GLCM texture features are obtained from a statistical analysis. Due to the difference in the script characteristics, the results of the statistical analysis show significant diversity among scripts. This represents the key point for decision-making process of script identification. The proposed method is tested on documents from custom oriented database. The experiments gave encouraging results.

Acknowledgments This work was partially supported by the Grant of the Ministry of Education, Science and Technological Development of the Republic Serbia, as a part of the project TR33037 and III43011.

References

Bharati MH, Liu JJ, MacGregor JF (2004) Image texture analysis: methods and comparisons. Chemom Intell Lab Systems 72(1):57–71

Brodić D, Milivojević ZN, Maluckov Č (2013) Recognition of the script in Serbian documents using frequency occurrence and co-occurrence analysis. Sci World J 2013(896328):1–14

Brodić D, Milivojević Z, Maluckov Č A (2014) Script characterization in the old Slavic documents. In: Elmoataz A, Lezoray O, Nouboud F, Mammass D (eds) Image and Signal Processing, LNCS 8509, pp 230–238. Springer, Berlin

Busch A, Boles WW, Sridharan S (2006) Texture for script identification. IEEE Trans Pattern Anal Mach Intell 27(11):1720–1732

Chaudhuri BB, Pal U, Mitra M (2002) Automatic recognition of printed Oriya script. Sadhana 27(1):23–34

Clausi DA (2002) An analysis of co-occurrence texture statistics as a function of grey level quantization. Can J Remote Sens 28(1):45–62

Del Bimbo A (2001) Visual information retrieval. Morgan Kaufmann Publishers Inc, San Francisco

Eleyan A, Demirel H (2011) Co-occurrence matrix and its statistical features as a new approach for face recognition. Turkish J Electrical Eng Comput Sci 19(1):98–107

Ghosh D, Dube T, Shivaprasad AP (2010) Script recognition—a review. IEEE Trans Pattern Anal Mach Intell 32(12):2142–2161

Haralick R, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Systems Man Cybern 3(6):610–621

Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67(5):786–804

Joshi GD, Garg S, Sivaswamy J (2007) A generalised framework for script identification. Int J Document Anal Recogn (IJDAR) 10(2):55–68

Pal U, Chaudhury BB (2002) Identification of different script lines from multi-script documents. Image Vis Comput 20(13–14):945–954

Silva C, Ribeiro B (2007) On text-based mining with active learning and background knowledge using SVM. Soft Comput 11(6):519–530

Tolambiya A, Venkatraman S, Kalra PK (2010) Content-based image classification with wavelet relevance vector machines. Soft Comput 14(2):129–136

Valkealahti K, Oja E (1998) Reduced multidimensional co-occurrence histograms in texture classification. IEEE Trans Pattern Anal Mach Intell 20(1):90–94

Yang Z, Purves D (2004) The statistical structure of natural light patterns determines perceived light intensity. In: Proceedings of the National Academy of sciences of the United States of America 101(23):8745–8750

Zhang J, Tan T (2002) Brief review of invariant texture analysis methods. Pattern Recogn 35(3):735–747

Zramdini AW, Ingold R (1998) Optical font recognition using typographical features. IEEE Trans Pattern Anal Mach Intell 20(8):877–882