Script Recognition by Statistical Analysis of the Image Texture

Invited Paper

Darko Brodić
Technical Faculty in Bor
University of Belgrade
Bor, Serbia
dbrodic@tf.bor.ac.rs

***Abstract*—The paper proposed an algorithm for the script identification using the statistical analysis of the texture obtained by script mapping. First, the algorithm models the script sign by the equivalent script type. The script type is determined by the position of the letter in the baseline area. Furthermore, the extraction of the features is performed. This step of the algorithm is based on the script type occurrence and co-occurrence pattern analysis. Then, the resultant features are compared. Their differences simplify the script feature classification. The algorithm is tested on the German and Slavic printed documents incorporating different scripts. The experiment gives the results that are promising.**

Keywords - Coding, Cultural heritage, Historical documents, Script recognition, Statistical analysis.

I. Introduction

Script recognition is a part of document image analysis [1]. Many techniques have been proposed for the script recognition. They are typically classified into global or local ones.

Global methods characterize the processing of the large image areas, which are subjected to the statistical or frequency-domain analysis [2]. However, the image area normalization is mandatory. [3]. On contrary, the local methods split up text into small pieces. They typically represents characters, words or lines. After that, the black pixel runs analysis is performed [4].

The proposed algorithm integrates the local and global approach. First, it extracts characters from the text. Then, it codes the characters according to their script type [5]. The coded text is obtained, which is an input to an occurrence (frequency analysis) and co-occurrence (statistical analysis) similarly as in global methods. As the results of aforementioned analysis, statistical measures of the gray-level co-occurrence matrix (GLCM) are extracted [6],[7]. To classify the results a linear discrimination function is proposed. It represents a key point in a decision-making process of the script discrimination.

The proposed approach incorporates the statistical analysis of the texture. Texture is suitable for extracting similarities and dissimilarities between images. However, the novelty of the proposed approach [8] is given by specific text modeling and 1-D texture analysis. During text modeling the number of variables is substantially reduced. Furthermore, the image is replaced with text. Hence, the image which represents a 2-D image is replaced with the text given by 1-D "image". All aforementioned contributes to the algorithm's speed.

The paper is organized as follows. Section 2 describes the proposed algorithm. Section 3 explains the experiment. Section 4 presents the results and gives the discussion. Section 5 makes conclusions.

II. The Methods

The proposed algorithm is a multi-stage method. It includes the following stages:

1. Coding,
2. Feature extraction,
3. Feature classification,
4. Script identification criteria.

Figure 1 illustrates the multi-stage method flow.

A. Coding

The first step of the algorithm represents a coding. It is established using into account the position of the letter in the text line. To explain it, the text line needs to be considered. It consists of three vertical zones [9]:

- Upper zone,
- Middle zone,
- Lower zone.

Figure 2 illustrates the vertical zones that belong to the text line.

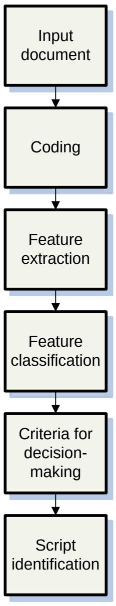


Figure 1. The multi-stage method flow

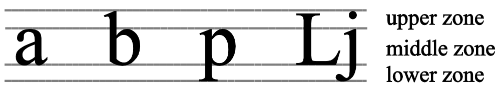


Figure 2. Vertical zones in text line

Using into account text line zones, one can distinguish four different script types [9]:

- Base letter (B),
- Ascender letter (A),
- Descendent letter (D),
- Full letter (F).

Base letters (B) like the letter **x**, occupy a middle zone only; ascender letters (A), like the letter **t**, spread over the middle and upper zones; while descendent letters (D), like the letter **p**, 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). This way, all letters are coded according to their script type classification. To organize data, the following mapping is made [5,10]:

$$B \rightarrow 1,\ A \rightarrow 2,\ D \rightarrow 3,\ F \rightarrow 4 \tag{1}$$

This way, the coded text is established (Appendix contains the alphabets and equivalent codes [10]).

B. Feature Extraction

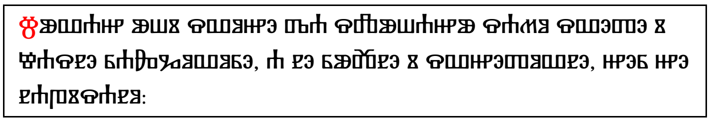
Feature extraction is based on statistical analysis of the coded text. Figure 3 illustrates the text written in different Slavic scripts and their coded text, while Figure 4 shows German text written in Fraktur and Latin with their coded text.

Čuvaj uši svoje da slušaju samo svete i časne razgovore, a ne ružne i svjetovne, jer je napisano:

(a)

2 1 1 1 4 1 2 2 1 1 1 4 1 2 1 1 2 1 2 1 4 1 1 1 1 1 1 1
1 2 1 2 2 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1
1 4 1 2 1 1 1 1 4 1 1 4 1 1 1 3 2 1 1 1 1

(b)



(c)

3 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 1 1 2 1 1 1 1
1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 3 1 1 2 1 1

(d)

Чувај уши своје да слушају само свете и часне разговоре, а не ружне и свјетовне, јер је написано:

(e)

2 3 1 1 4 3 1 1 1 1 1 4 1 3 1 1 1 3 1 1 4 3 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 3 1 1 1 1 3 3 1 1 1 1 1
1 4 1 1 1 1 1 1 4 1 3 4 1 1 1 1 1 1 1 1 1

(f)

Figure 3. Same text given in different Slavic 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.

Füllest wieder Busch und Tal Still mit Nebelglanz,
Lösest endlich auch einmal Meine Seele ganz;

(a)

2 2 2 2 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 2 2 1 2 2 2 1 2
2 1 1 2 2 1 2 1 2 3 2 1 1 1 2 2 1 1 1 2 1 1 2 2 1 1 2 1
1 1 2 1 1 1 1 1 2 2 1 1 1 1 2 1 1 2 1 3 1 1 1

(b)

Füllest wieder Busch und Tal Still mit Nebelglanz,
Lösest endlich auch einmal Meine Seele ganz;

(c)

4 2 2 2 1 4 2 1 1 1 2 1 1 2 1 4 1 4 1 1 2 2 1 2 2 2 1 2
2 1 2 2 2 1 2 1 2 3 2 1 1 3 2 2 4 1 4 2 1 1 2 2 1 1 4 1
1 1 4 1 1 1 1 1 2 2 1 1 1 1 2 1 1 2 1 3 1 1 3

(d)

Figure 4. Same text given in different German scripts: (a) Original text in Latin script, and (b) its coded counterpart; (c) Original text in Fraktur script, and (d) its coded counterpart

In the first step, the script type distribution of coded text is analyzed. As a result, four script features are extracted. After the script type distribution analysis, the coded text is subjected to the co-occurrence analysis [6,7]. This way, the texture features are calculated. They use the conditional joint probabilities of all pair wise combinations of grey levels in the window of interest (WOI). WOI is determined by the inter-pixel distance Δx and Δy shown in Figure 5.

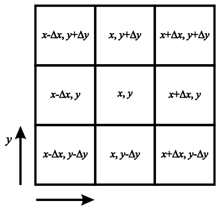


Figure 5. Window of interest (WOI).

The method starts from the top left corner and counts number of each reference pixel occurrences in respect to neighbour pixel relationship. At the end of this process, the element (i, j) gives the number of how many times the gray levels i and j appears as a sequence of two pixels located at Δx and Δy. This way, GLCM **P** for an image **I** with M rows and N columns is given as [6,7]:

$$P(i,j) = \sum_{x=1}^{M}\sum_{y=1}^{N} \begin{cases} 1, \text{if } I(x,y) = i \text{ , } I(x+\Delta x, y+\Delta y) = j \\ 0, \quad \text{otherwise} \end{cases} \tag{2}$$

Furthermore, matrix **P** is normalized giving a matrix **C**:

$$C(i,j) = P(i,j) / \sum_{i,j}^{G} P(i,j) \tag{3}$$

In our case, the coded text is given as 1-D image, which leads to following: $\Delta x = \pm 1$, $\Delta y = 0$ [10]. Furthermore, the number of texture features can be extracted from the GLCM:

$$Energy = \sum_{i}^{G}\sum_{j}^{G} C(i,j)^2 \tag{4}$$

$$Entropy = \sum_{i}^{G}\sum_{j}^{G} C(i,j) \cdot \log C(i,j) \tag{5}$$

$$\text{Maximum} = \max \sum_{i}^{G}\sum_{j}^{G} C(i,j) \quad \forall i, j \tag{6}$$

$$Dissimilarity = \sum_{i}^{G}\sum_{j}^{G} C(i,j) \cdot |i-j| \tag{7}$$

$$Contrast = \sum_{i}^{G}\sum_{j}^{G} C(i,j) \cdot (i-j)^2 \tag{8}$$

$$Inverse\ Different\ Moment = \sum_{i}^{G}\sum_{j}^{G} C(i,j) / [1 + (i-j)^2] \tag{9}$$

$$Homogeneity = \sum_{i}^{G}\sum_{j}^{G} C(i,j) / [1 + (i-j)] \tag{10}$$

$$Correlation = \sum_{i}^{G}\sum_{j}^{G} (i-\mu_x) \cdot (j-\mu_y) \cdot C(i,j) / (\sigma_x \cdot \sigma_y) \tag{11}$$

C. *Feature Classification*

According to the aforementioned script type distributions and GLCM features, the text examples from Figures 3 and 4 are analyzed.

Table I shows the script type distributions, which are obtained from the same text written in different Slavic scripts [11].

TABLE I. SCRIPT TYPE DISTRIBUTIONS BETWEEN SLAVIC SCRIPTS

Script Type	**Different Slavic Scripts**		
	Cyrillic	***Glagolitic***	***Latin***
Base	0.7786	0.8015	0.6336
Ascender	0.0153	0.1679	0.2595
Descendent	0.1298	0.0305	0.0229
Full	0.0763	0.0000	0.0840

Similarly, the same analysis is carried out with German text. The results are shown in Table II [12].

TABLE II. SCRIPT TYPE DISTRIBUTIONS BETWEEN GERMAN SCRIPTS

Script Type	**Different German Scripts**	
	Fraktur	***Latin***
Base	0.5324	0.6115
Ascender	0.3237	0.3669
Descendent	0.0360	0.0216
Full	0.1079	0.0000

Table III shows typical GLCM features extracted from the Slavic documents. Differences in their features characterize each script [11].

At the end, Table IV shows typical GLCM features extracted from the German documents [12].

TABLE III. GLCM FEATURES BETWEEN SLAVIC SCRIPTS

GLCM Feature	Different Slavic Scripts		
	Cyrillic	*Glagolitic*	*Latin*
Energy	0.4140	0.4651	0.2159
Entropy	−1.3957	−1.1347	−1.8432
Maximum	0.6231	0.6538	0.3692
Dissimilarity	0.7615	0.6540	0.8846
Contrast	1.7923	1.3769	1.8077
Inverse different moment	0.7223	0.7454	0.6500
Homogeneity	0.7641	0.7859	0.6769
Correlation	0.0791	0.0742	−0.1291

TABLE IV. GLCM FEATURES BETWEEN GERMAN SCRIPTS

GLCM Feature	Different German Scripts	
	Fraktur	*Latin*
Energy	0.1596	0.2604
Entropy	-2.0675	-1.4985
Maximum	0.2754	0.3768
Dissimilarity	0.9855	1.0000
Contrast	2.0725	2.0725
Inverse different moment	0.6159	0.6072
Homogeneity	0.6504	0.6715
Correlation	-0.1403	-0.0149

D. *Script Identification Criteria*

All criteria obtained from occurrence and co-occurrence analysis are used as input criteria for decision-making. However, the comprehensive criteria will be established after applying the algorithm to the database of Slavic and German text documents. As a result, the statistical analysis will show the clear difference between scripts in document with the same content.

III. EXPERIMENTS

The algorithm is subjected to the experiment in order to investigate its efficiency and correctness. To perform the experiment, a custom-oriented database is created. First part of database includes 100 Slavic documents written in Glagolitic, Latina and Cyrillic script. 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) [13]. Second part of database includes 100 German documents with the poems written by J. W. von Goethe written in Latin and Fraktur script. Typical length of text is from approx. 200 to 1,000 characters. The result of the experiment gives the percentage of the correct script recognition.

IV. RESULTS AND DISCUSSION

The results of experiment are given in the form of the script type distributions and the extended set of eight GLCM texture features. The script type distributions are used to extract four script features, which are used to characterize different scripts. To quantify the obtained results, we used the minimum and maximum values. Furthermore, the extended set of eight GLCM texture features is used 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 in the first place, and texts written in Latin and Fraktur in the second place.

It is very important to use only the measures with distinct difference in values for the different scripts. Establishing the ratio between these measures for different scripts gives their relation that can be utilized as a part of the identification criteria. These measures create the criteria for script discrimination.

The experiment with Slavic documents shows the results given in Tables V-VI [11,12].

TABLE V. SCRIPT TYPE DISTRIBUTIONS OF THE SLAVIC SCRIPTS

Script Type	Different Slavic Scripts					
	Cyrillic		*Glagolitic*		*Latin*	
	min	max	min	max	min	max
Base	0.48	0.62	0.68	0.79	0.68	0.85
Ascender	0.28	0.44	0.16	0.24	0.03	0.16
Descendent	0.03	0.07	0.03	0.17	0.07	0.16
Full	0.01	0.08	0.00	0.00	0.01	0.07

TABLE VI. GLCM FEATURES OF THE SLAVIC SCRIPTS

GLCM Feature	Different Slavic Scripts					
	Cyrillic		*Glagolitic*		*Latin*	
	min	max	min	max	min	max
Energy	0.167	0.214	0.309	0.432	0.325	0.507
Entropy	-2.026	-1.757	-1.499	-1.211	-1.647	-1.173
Maximum	0.237	0.355	0.497	0.626	0.539	0.701
Dissimilarity	0.757	0.986	0.684	0.981	0.595	0.871
Contrast	1.076	1.978	1.520	2.243	1.227	2.021
ID moment	0.604	0.679	0.636	0.742	0.679	0.780
Homogeneity	0.636	0.700	0.694	0.784	0.727	0.810
Correlation	-0.243	-0.159	-0.131	0.480	-0.118	0.075

Hence, the combination of the aforementioned results creates the final criteria for the script differentiation in the Slavic documents. It is given 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
```

The experiment with German documents shows the results given in Tables VII-VIII [11,12].

TABLE VII. SCRIPT TYPE DISTRIBUTIONS OF THE GERMAN SCRIPTS

Script Type	Different German Scripts			
	Fraktur		***Latin***	
	min	max	min	max
Base	0.49	0.52	0.55	0.57
Ascender	0.37	0.39	0.41	0.43
Descendent	0.01	0.04	0.01	0.03
Full	0.08	0.10	0.00	0.01

TABLE VIII. GLCM FEATURES OF THE GERMAN SCRIPTS

GLCM Feature	Different German Scripts			
	Fraktur		***Latin***	
	min	max	min	max
Energy	0.152	0.175	0.236	0.248
Entropy	-2.108	-1.952	-1.572	-1.495
Maximum	0.219	0.241	0.2709	0.2871
Dissimilarity	0.921	1.037	0.596	1.199
Contrast	1.726	2.029	0.661	2.455
Inverse different moment	0.581	0.620	0.526	0.708
Homogeneity	0.611	0.641	0.604	0.711
Correlation	-0.186	-0.119	-0.181	-0.130

Accordingly, the combination of the aforementioned results creates the final criteria for the script differentiation in the German documents. It is given by the following pseudo code:

```
IF [(B < 0.53) AND (A < 0.4) AND (F > 0.05)
    AND (Energy < 0.2) AND (Entropy < -1.85)
    AND (Maximum < 0.25)]
            Writeln('Fraktur Text')
ELSE
            Writeln('Latin Text')
END
```

At the end, the speed testing of the proposed method shows that it is a computationally non-intensive. Its processing time is as low as 0.1 sec. per text that includes 2K characters.

V. CONCLUSIONS

The manuscript proposed the algorithm for the script characterization and identification. The algorithm includes the comprehensive statistical analysis of coded document, which is obtained by mapping the initial text document according to the script types of each character. Because the characteristics of both scripts are different, the statistical analysis shows significant diversity between them. Hence, the successful script identification can be conducted by creating joint criteria which is based on the obtained statistical features. The proposed technique is tested on the example of Slavic and German printed documents . The experiments gave encouraging results.

ACKNOWLEDGMENT

This work was supported by the Grant of the Ministry of Republic Serbia, as a part of the project TR33037.

REFERENCES

[1] D. Ghosh, T. Dube, A. P.Shivaprasad, “Script Recognition - A Review,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.32, no.12, 2010, pp. 2142–2161.

[2] G. D. Joshi, S. Garg, J. Sivaswamy, “A Generalised Framework for Script Identification,” International Journal of Document Analysis and Recognition, vol.10, no.2, 2007, pp.55–68.

[3] A. Busch, W. W. Boles, S. Sridharan, “Texture for Script Identification. IEEE Transaction on Pattern Analysis and Machine Intelligence,” vol.27, no.11, 2006, pp.1720–1732.

[4] U. Pal, B. B. Chaudhury, “Identification of Different Script Lines from Multi-Script Documents,” Image and Vision Computing, vol.20, no.13-14, 2002, pp.945–954.

[5] D. Brodić, Z. N. Milivojević, Č. A. Maluckov, “Recognition of the Script in Serbian Documents using Frequency Occurrence and Co-occurrence Analysis,” The Scientific World Journal, vol.2014, art.896328, 2013, pp.1–14. doi:10.1155/2013/896328

[6] R. Haralick, K. Shanmugam, I. Dinstein, “Textural Features for Image Classification,” IEEE Transactions on Systems, Man, and Cybernetics vol.3, no.6, 1973, pp.610–621.

[7] D. A. Clausi, “An Analysis of Co-occurrence Texture Statistics as a Function of Grey Level Quantization,” Canadian Journal of Remote Sensing, vol28, no.1, 2002, pp.45–62.

[8] S. Bourennane, J. Marot, C. Fossati, A. Bouridane, K. Spinnler, “Multidimensional Signal Processing and Applications,” The Scientific World Journal, vol.2014, art.365126, 2014, pp. 1–2.

[9] A. W. Zramdini, R. Ingold, “Optical font recognition using typographical features,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.20, no.8, 1998, pp.877–882.

[10] D. Brodić, Z. N. Milivojević, Č. A. Maluckov, “Script Characterization in the Old Slavic Documents,” Image and Signal Processing (ICISP 2014), LNCS 8509 (Eds. A. Elmoataz et al.), 2014, pp.230–238.

[11] D. Brodić, Z. N. Milivojević, Č. A. Maluckov, “An approach to the script discrimination in the Slavic documents,” Soft Computing, 2014, pp.1–11. doi: 10.1007/s00500-014-1435-1

[12] D. Brodić, Z. N. Milivojević, Č. A. Maluckov, “Identification of the Fraktur and Latin Script in German Printed Documents by Texture Analysis,” In review.

[13] http://www.croatianhistory.net/etf/juraj_slovinac_misli.html

APPENDIX

TABLE I. CODING OF SLAVIC ALPHABETS

Glagolitic	Coding	Latin	Coding	Cyrillic	Coding	Glagolitic	Coding	Latin	Coding	Cyrillic	Coding
ⰎⰉ	3	Lj	4	Љ	2	ⰾⰹ	2	lj	4	љ	1
ⰐⰉ	3	Nj	4	Њ	2	ⱀⰹ	1	nj	4	њ	1
Ⰵ	3	E	2	Е	2	ⰵ	1	e	1	е	1
Ⱃ	3	R	2	Р	2	ⱃ	1	r	1	р	3
Ⱅ	C	T	2	Т	2	ⱅ	1	t	2	т	1
Ⰸ	3	Z	2	З	2	ⰸ	2	z	1	з	1
Ⱆ	3	U	2	У	2	ⱆ	1	u	1	у	3
Ⰻ	3	I	2	И	2	ⰻ	1	i	2	и	1
Ⱁ	3	O	2	О	2	ⱁ	1	o	1	о	1
Ⱂ	3	P	2	П	2	ⱂ	3	p	3	п	1
Ⱎ	3	Š	2	Ш	2	ⱎ	1	š	2	ш	1
Ⰼ	3	Đ	2	Ђ	2	ⰼ	1	đ	2	ђ	4
Ⰰ	3	A	2	А	2	ⰰ	2	a	1	а	1
Ⱄ	3	S	2	С	2	ⱄ	1	s	1	с	1
Ⰴ	3	D	2	Д	2	ⰴ	1	d	2	д	3
Ⱇ	3	F	2	Ф	2	ⱇ	3	f	2	ф	4
Ⰳ	3	G	2	Г	2	ⰳ	1	g	3	г	1
Ⱈ	3	H	2	Х	2	ⱈ	1	h	2	х	1
Ⰼ	3	J	2	Ј	2	ⰼ	1	j	4	ј	4
Ⰽ	3	K	2	К	2	ⰽ	1	k	2	к	1
Ⰾ	3	L	2	Л	2	ⰾ	2	l	2	л	1
Ⱍ	3	Č	2	Ч	2	ⱍ	2	č	2	ч	1
Ⱋ	3	Ć	2	Ћ	2	ⱋ	2	ć	2	ћ	2
Ⰶ	3	Ž	2	Ж	2	ⰶ	2	ž	2	ж	1
ⰄⰆ	4	Dž	2	Џ	4	ⰴⰶ	2	dž	2	џ	3
Ⱌ	3	C	2	Ц	4	ⱌ	1	c	1	ц	3
Ⰲ	3	V	2	В	2	ⰲ	1	v	1	в	1
Ⰱ	3	B	2	Б	2	ⰱ	1	b	2	б	2
Ⱀ	3	N	2	Н	2	ⱀ	1	n	1	н	1
Ⰿ	3	M	2	М	2	ⰿ	1	m	1	м	1
Ⱑ	3	Ja, (I)je	-	Ja, (И)je	-	ⱑ	1	ja, (i)je	-	ja, (и)je	-

TABLE II. CODING OF GERMAN DIACRITICS

Latin	Coding	Fraktur	Coding	Latin	Coding	Fraktur	Coding
Ä	2	𝔄̈	2	ä	2	𝔞̈	2
Ö	2	𝔒̈	2	ö	2	𝔬̈	2
Ü	2	𝔘̈	2	ü	2	𝔲̈	2
-		-		ß	2	ß	4

TABLE III. CODING OF GERMAN ALPHABETS

Latin	Coding	Fraktur	Coding	Latin	Coding	Fraktur	Coding
A	2	𝔄	2	a	1	𝔞	1
B	2	𝔅	2	b	2	𝔟	2
C	2	ℭ	2	c	1	𝔠	1
D	2	𝔇	2	d	2	𝔡	2
E	2	𝔈	2	e	1	𝔢	1
F	2	𝔉	4	f	2	𝔣	4
G	2	𝔊	2	g	3	𝔤	3
H	2	ℌ	4	h	2	𝔥	4
I	2	ℑ	2	i	2	𝔦	2
J	2	𝔍	4	j	4	𝔧	4
K	2	𝔎	2	k	2	𝔨	2
L	2	𝔏	2	l	2	𝔩	2
M	2	𝔐	2	m	1	𝔪	1
N	2	𝔑	2	n	1	𝔫	2
O	2	𝔒	2	o	1	𝔬	2
P	2	𝔓	4	p	3	𝔭	3
Q	2	𝔔	2	q	3	𝔮	3
R	2	ℜ	2	r	1	𝔯	1
S	2	𝔖	2	s	1	𝔰	1
T	2	𝔗	2	t	2	𝔱	1
U	2	𝔘	2	u	1	𝔲	1
V	2	𝔙	2	v	1	𝔳	1
W	2	𝔚	2	w	1	𝔴	1
X	2	𝔛	2	x	1	𝔵	3
Y	2	𝔜	4	y	3	𝔶	3
Z	2	ℨ	4	z	1	𝔷	3