An Image Texture Analysis Method for Minority Language Identification

Darko Brodić[1](✉), Alessia Amelio[2], and Zoran N. Milivojević[3]

[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

[3] College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia
zoran.milivojevic@vtsnis.edu.rs

Abstract. This paper introduces an image texture analysis method for minority language identification. In the first stage, each letter is associated with a given script type according to its energy status in the text-line area. Mapping is carried out by extracting unicode text and transforming it into coded text. There are four different script types, which correspond to four grey levels of an image. Then, the obtained image is subjected to a feature extraction process performed by the texture analysis. This way, the grey level co-occurrence matrix and its derivative features are calculated. Extracted features are compared and classified using the K-Nearest Neighbors and Naive Bayes methods to establish a difference that can identify a minority language such as Serbian language among other world languages in the text. Very good accuracy results prove the efficiency of the proposed approach, when compared to other state-of-the-art methods.

Keywords: Image processing · Natural language processing · Classification · Statistical analysis · Feature extraction

1 Introduction

As of July 2013, the United Nations website was available in six languages, while the official website of the European Union could be read in 24 languages. Furthermore, Google supported 90 languages, while Wikipedia supported 295 [25]. However, just a few of them have an above-average dispersion. As a consequence, the Web mainly employs texts written in the so-called world languages, such as English, French, German and Spanish (over 70% of all web content). In contrast, the vast majority, i.e., over 95% of the languages have already lost the capacity to ascend digitally [16]. A number of minority languages in Europe still exists despite the strong pressure from the majority languages such as English, French, German and Spanish. The later languages are clearly widespread and dominant languages in Europe and on the Web. As a consequence, most speakers of minority languages are bilingual or multilingual. Hence, minority languages

V.E. Brimkov and R.P. Barneva (Eds.): IWCIA 2017, LNCS 10256, pp. 280–293, 2017.
DOI: 10.1007/978-3-319-59108-7_22

are changing because the websites in these languages also have to include some of the widespread world languages, creating the concept of multilingual websites. Hence, the extraction of some text fragments in minority languages and their classification is a real challenge and worths investigation. If we take as an example of minority language the Serbian one, then it is used in around 0.1% of the websites [25].

Language identification is the process of language recognition in a certain text. Many methods have been proposed for language identification. They are classified into the following groups [13]: (i) Letter based approach, (ii) Word based approach, (iii) N-gram approach, and (iv) Language identification using a Markov model. Previous research has included the statistical analysis of the text content. In the letter based approach, the frequency distribution of certain letters or common letter combination is analyzed for making easier the language identification [23]. The problem is that the obtained results would be reliable if the number of words in the sentences was high (above 21). Word based approach uses only words up to a specific length or the most frequent word's appearance for establishing the language model [12,22]. The main limitation is the training phase needing a high number of documents to create the language model. Another technique generates a language n-gram model for each of the languages, extracting substrings of length n and computing their frequency in the text [4,8]. A problem can occur when the pieces of input text are composed of several languages. Unfortunately, this approach cannot solve such a case. The methods in the last group use Markov models in combination with Bayesian decision rules to produce models for each language [9]. It is worth noting that the training process is computationally intensive, needing at least 50-100 K words for successfully testing small parts of text of length above 100 characters [18].

In this paper, an image-based method for language identification is proposed to overcome the limitations of the previous approaches. In fact, it has the following advantages: (i) it does not require a large piece of text for training, (ii) it is not computer time intensive, (iii) it is able to identify a minority language among the most widespread languages on the Internet, (iv) it needs unicode or prior to Optical Character Recognition (OCR) input, and (v) it is not dependent on the certain alphabet extension with specific letters. The first stage of the method is based on coding. It maps the initial text into a coded text established according to the energy characteristic of each letter based on its position in the text line. A similar approach with six established elements, which is not converted to image elements, was proposed in [21]. Still, this method uses a typical n-gram method for language discrimination. Unlike that, in our method the coded text, which includes four different script types, corresponds to an image with four grey levels. Then, it is subjected to co-occurrence statistical analysis in order to extract texture features. This way, the feature extraction and further text classification are performed in the image processing area. To test the proposed approach, an experiment is conducted on a custom-oriented dataset containing text mainly from Web documents (unicode format) given in different world languages: German, Spanish, English and French and minority language such as Serbian. The classification is performed by employing K-Nearest Neighbors and Naive Bayes algorithms. The differences in the feature values establish an important aspect for classification that can identify a minority language such as Serbian among the world languages (in unicode, i.e. web, PDF, or in a scanned text document). This presents a new application of the method which has not been investigated in the previous literature [1,2]. Classification results are compared with the results obtained by the n-gram language model for identification of the minority language. This comparison confirms the superiority of the proposed method in language identification. It was also confirmed in a complex problem such as discrimination of evolving languages [3].

The remainder of the paper is organized as follows. Section 2 addresses all aspects concerning the proposed algorithm, including script coding, four grey level image definition, texture features extraction and classification. Section 3 discusses the experiment. Section 4 presents the classification results. Section 4 draws the conclusions and outlines future work directions.

2 Proposed Algorithm

The proposed algorithm is a multi-stage method that consists of the Following steps: (i) unicode text is mapped into the coded text according to the energy characteristics, (ii) creation of four grey level image that fully corresponds to the coded text, (iii) feature extraction by co-occurrence analysis, (iv) feature classification, and (v) identification of the language. Figure 1 shows the flow of the proposed algorithm.

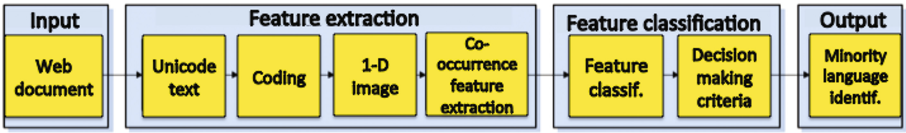
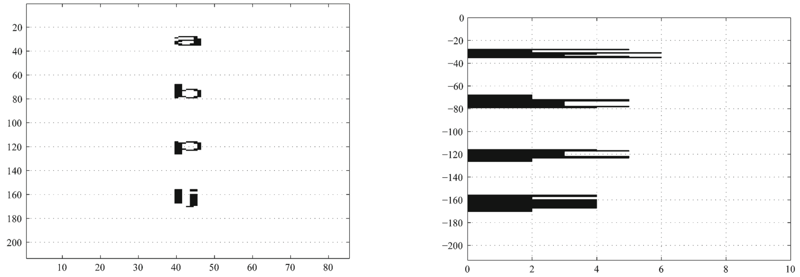


Fig. 1. The flow of the algorithm.

2.1 Unicode Text Mapping

Typically, the text in the documents is divided into text lines. Each character can be separated according to its energy characteristic, i.e. horizontal projections profile. Figure 2 shows the different characters and their corresponding energy characteristics. We can realize that the height of different characters corresponds to their energy. Obviously, all characters have different energy characteristic. The most diffused energy is given by the characters that outspread over the full text line height such as character lj in Serbian or Croatian language. Taking into account all aforementioned, we can draw virtual lines in each text line. They can be represented as: (i) top-line, (ii) upper-line, (iii) base-line, and (iv) bottom-line. Furthermore, they establish three vertical zones [26]: (i) upper zone, (ii) middle



(a) (b)

Fig. 2. Energy characteristic of different characters: (a) different characters in different text lines, (b) their corresponding energy.

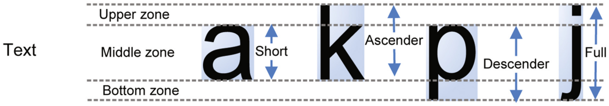


Fig. 3. Definition of the script characteristics according to baseline status of the text.

zone, and (iii) lower zone. All letters can be classified according to these vertical zones. The short letters (S) take the middle zone. The capital letters and letters with ascenders (A) take the middle and upper zones. The descendent ones (D) occupy the middle and lower zones. The full letters (F) outspread over the upper, middle and lower zones. Figure 3 shows the script characteristics according to the letter baseline position.

According to the vertical zone classification, all letters from the alphabet are substituted with the script types. Each letter is positioned into a given zone(s) in the text line. Consequently, it is mapped to a unique element of the set {S, A, D, F}.

2.2 Image Creation

In order to easily apply a statistical analysis, the script type should be injectively coded in the following way:

$$S \to 0, A \to 1, D \to 2, F \to 3. \tag{1}$$

These codes can be transformed into the grey level pixels of an image. Hence, it corresponds to an image with four grey levels. Figure 4 shows an example of the coding procedure for a text sample given in German language.

2.3 Feature Extraction

Currently, the initial text is transformed into an image through a variable reduction process. The obtained image is then subjected to the texture analysis. It

Erst jetzt entschied er, nicht zum Künstler, sondern zum Dichter geboren zu sein.

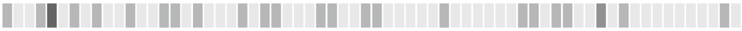
(a)

ASSAFSASASSASSAASASSSASAASSSAASSAASSSSSASSSSSSAASAASSDSASSSSSSSSAS

(b)

10013010100100110100010110001100110000010000001101100201000000000010

(c)



(d)

Fig. 4. An example of coding procedure for a German text: (a) initial text, (b) extraction of characters, (c) coding according to script types, (d) image creation.

includes the extraction of the co-occurrence probabilities, which provide second-order texture features [14]. These features are extracted from the image in two steps. At the first step, the pairwise spatial co-occurrences of pixels separated by a particular angle θ and distance d are tabulated using a Grey Level Co-occurrence Matrix (GLCM). At the second step, a set of texture measures is calculated from GLCM. The GLCM shows how often different combinations of grey level co-occur in a part or in the whole image [14].

Let's suppose that our grey scale image is given as $\mathbf{I}$, featuring M rows, N columns, and T number of grey levels. GLCM represents the spatial relationship of grey levels in the image $\mathbf{I}$. It is a $T \times T$ square matrix. To compute GLCM $\mathbf{C}$, a central pixel $I(x, y)$ with a neighborhood defined by the Window Of Interest (WOI) is taken. WOI is defined by inter-pixel distance d and orientation θ. Hence, for the given image $\mathbf{I}$, GLCM $\mathbf{C}$ is defined as [10]:

$$C(i,j) = \sum_{x=1}^{T} \sum_{y=1}^{T} \begin{cases} 1 \text{ if } I(x,y) = i, \text{and} \\ \qquad I(x+\Delta x, y+\Delta y) = j, \\ 0 \qquad \text{otherwise} \end{cases} \tag{2}$$

where i and j are the intensity values of the image $\mathbf{I}$, x and y are the spatial positions in the image $\mathbf{I}$, the offset $(\Delta x, \Delta y)$ is the distance between the pixel-of-interest and its neighbor. It should be noted that the offset depends on the direction θ that is used and the distance d at which the matrix is computed.

In our case, the neighborhood is given as 2-connected only, due to the nature of the text. Accordingly, θ is 0°, while d is typically used as first neighborhood, i.e. $d = 1$. Then, the normalized matrix $\mathbf{P}$ of GLCM $\mathbf{C}$ is calculated as [5]:

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

Table 1. Twelve co-occurrence elements.

μ_x	$\sum_{i=1}^{T} i \sum_{j=1}^{T} P(i,j)$,
μ_y	$\sum_{j=1}^{T} j \sum_{i=1}^{T} P(i,j)$,
σ_x	$\sqrt{\sum_{i=1}^{T}(i-\mu_x)^2 \sum_{j=1}^{T} P(i,j)}$,
σ_y	$\sqrt{\sum_{j=1}^{T}(j-\mu_y)^2 \sum_{i=1}^{T} P(i,j)}$,
Correlation	$\sum_{i=1}^{T}\sum_{j=1}^{T} \frac{(i\cdot j)\cdot P(i,j)-(\mu_x\cdot\mu_y)}{\sigma_x\cdot\sigma_y}$,
Energy	$\sum_{i=1}^{T}\sum_{j=1}^{T} P(i,j)^2$,
Entropy	$-\sum_{i=1}^{T}\sum_{j=1}^{T} P(i,j)\cdot logP(i,j)$,
Maximum	$\max\{P(i,j)\}$,
Dissimilarity	$\sum_{i=1}^{T}\sum_{j=1}^{T} P(i,j)\cdot \lvert i-j\rvert$,
Contrast	$\sum_{i=1}^{T}\sum_{j=1}^{T} P(i,j)\cdot (i-j)^2$,
Invdmoment	$\sum_{i=1}^{T}\sum_{j=1}^{T} \frac{1}{1+(i-j)^2} P(i,j)$,
Homogeneity	$\sum_{i=1}^{T}\sum_{j=1}^{T} \frac{P(i,j)}{1+\lvert i-j\rvert}$

Still, GLCM provides only a quantitative description of the spatial patterns. Hence, it is not used for practical image analysis. Ref. [14] proposed a set of texture measures, which summarize the information from GLCM. Although a total of 14 quantities, i.e. features was originally proposed, only subsets of them are used [17]. These are the following twelve GLCM texture measures: (i) mean value μ_x, (ii) mean value μ_y, (iii) standard deviation σ_x, (iv) standard deviation σ_y, (v) correlation, (vi) energy, (vii) entropy, (viii) maximum, (ix) dissimilarity, (x) contrast, (xi) inverse difference moment and (xii) homogeneity. The twelve co-occurrence elements are shown in Table 1. Hence, after this phase, the four grey level image is represented by a 12-dimensional feature vector.

2.4 Feature Classification

In order to classify the obtained feature vector, the K-Nearest Neighbors and Naive Bayes methods have been used, which are well-known algorithms for data classification.

K-Nearest Neighbors. K-Nearest Neighbors (K-NN) is a very easy approach to classify feature vectors [7,11]. Let Tr be the training set composed of n feature vectors with associated class labels and x_t be a test feature vector to classify. Classification of x_t is performed by computing the distance between x_t and each training vector in Tr from 1 to n. The K training vectors Tr' which are the nearest to x_t are finally considered. K is a fixed parameter of the algorithm, determining the amplitude of the neighborhood. The predicted class label for x_t is the one occurring most frequently in Tr'. Because even values of K can determine class labels with the same frequency in Tr' [19], the value of the K parameter is usually fixed to a small odd integer. When the instances are fixed-length vectors of real-value features, the distance function often adopted for

K-NN is the Euclidean one. However, different distance functions can determine variations in the similarity evaluation [19]. In fact, based on the chosen distance function, two vectors x_i and x_j can be considered more or less similar to each other. Consequently, other possible functions are selected for K-NN, such as the Manhattan distance or the Chebyshev distance.

Naive Bayes. Naive Bayes (NB) classifier is a probabilistic learning method based on the assumption that all variables are mutually independent, given the class variable [20]. The classifier is defined as: $f_{nb}(x_i) = \frac{p(Y=1)}{p(Y=0)} \prod_{k=1}^{h} \frac{p(x_i^k|Y=1)}{p(x_i^k|Y=0)}$, where $x_i = \{x_i^1, ..., x_i^h\}$ represents a vector of h features and class variable Y. In order to classify the test feature vector x_i in class 1 or class 0, the probability of each of its features conditioned to class 1 or 0 and the probability of occurrence of class 1 and 0 in the training set are computed. x_i is predicted to be in class 1 if and only if $f_{nb}(x_i) \geq 1$. Otherwise, it is predicted to be in class 0.

For numerical features, the normal distribution is considered for computing the probability values:

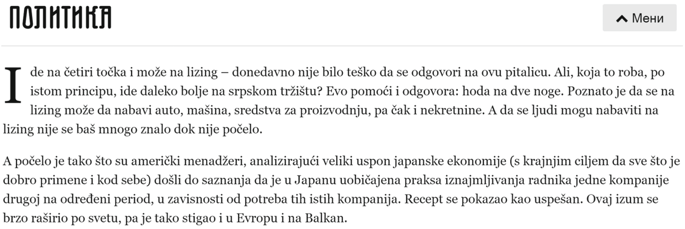
$$f(w, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(w-\mu)^2}{\sigma^2}}. \tag{4}$$

Accordingly, $p(x_i^k|Y = 1) = f(x_i^k, \mu_{y_i}, \sigma_{y_i})$, where $\mu_{Y=1}$ and $\sigma_{Y=1}$ are respectively the mean and standard deviation of the values of k-th feature with class 1.

3 Experiment

A test is conducted to evaluate the quality of the proposed method in correctly identifying a minority language among a set of world languages. Accordingly, a custom-oriented dataset extracted from the web text given in unicode format of different languages is employed. It represents a set of text excerpts in German, Spanish, English, French and Serbian languages. It consists of a total of 150 texts, divided into two classes respectively of 25 Serbian and 125 world language texts (German, Spanish, English and French). The class label for each text in the dataset corresponds to Serbian or world language. All texts have different contents and size. The size of the text excerpts is between 378 and 2822 characters. The length of the texts is chosen according to the size standard in factor analysis, which means that the total number of analyzed elements would be higher than 300 [24]. It means that our text samples contain approx. more than 300 characters. Figure 5 illustrates a web page in Serbian language from which the unicode text is extracted.

The proposed features are extracted from each text in the dataset, in order to create 150 feature vectors. Then, K-NN and NB algorithms are adopted on the feature representation of the dataset for identification of the minority language. Finally, classification results are compared and discussed.



Ide na četiri točka i može na lizing – donedavno nije bilo teško da se odgovori na ovu pitalicu. Ali, koja to roba, po istom principu, ide daleko bolje na srpskom tržištu? Evo pomoći i odgovora: hoda na dve noge. Poznato je da se na lizing može da nabavi auto, mašina, sredstva za proizvodnju, pa čak i nekretnine. A da se ljudi mogu nabaviti na lizing nije se baš mnogo znalo dok nije počelo.

A počelo je tako što su američki menadžeri, analizirajući veliki uspon japanske ekonomije (s krajnjim ciljem da sve što je dobro primene i kod sebe) došli do saznanja da je u Japanu uobičajena praksa iznajmljivanja radnika jedne kompanije drugoj na određeni period, u zavisnosti od potreba tih istih kompanija. Recept se pokazao kao uspešan. Ovaj izum se brzo raširio po svetu, pa je tako stigao i u Evropu i na Balkan.

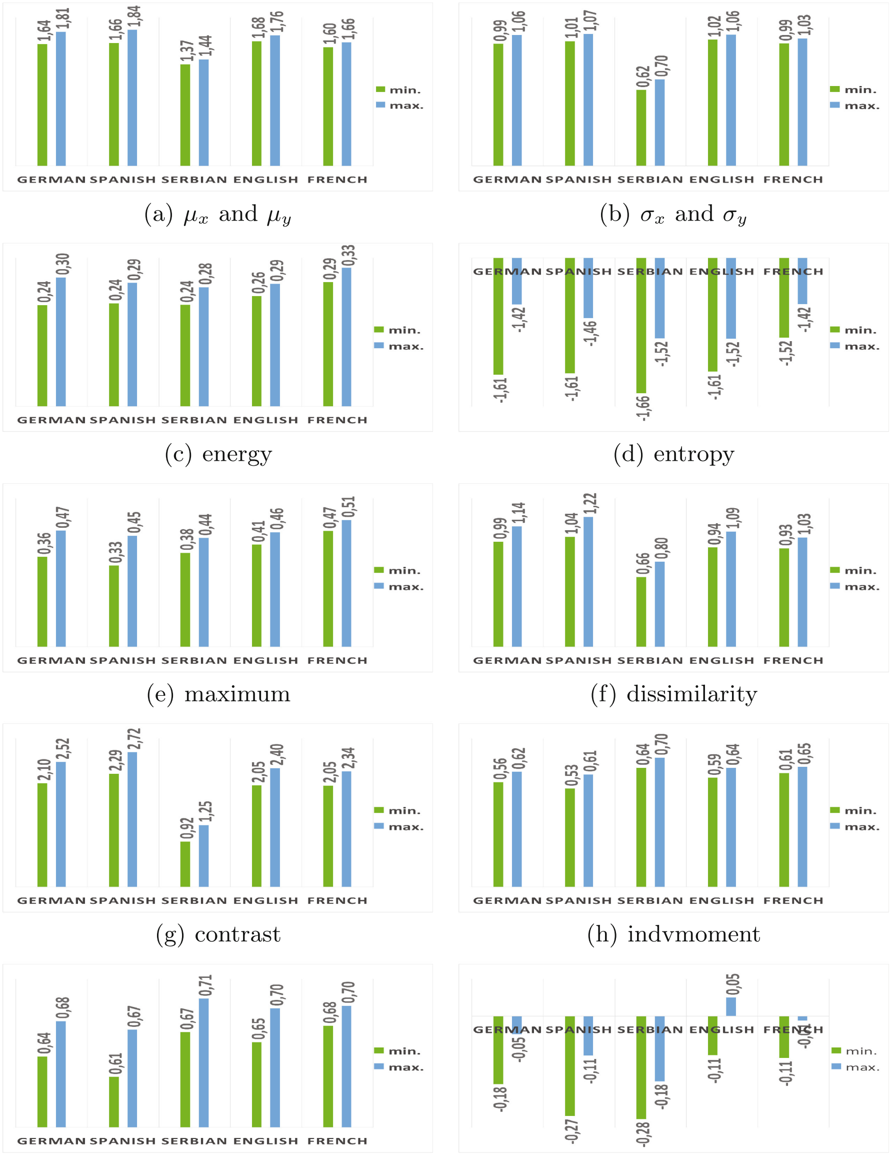
Fig. 5. Web page sample in Serbian language.

4 Results and Discussion

Figure 6 and Table 2 show the GLCM features obtained by the dataset for the world and Serbian languages in a min-max manner. μ_x and μ_y are the same as well as σ_x and σ_y count on two decimal places. Hence, only one graph representing both is enough. It is worth noting that Serbian language can mostly be discriminated from the world languages by the GLCM dissimilarity and contrast. In fact, the overall characteristics of Serbian text have much smaller values of dissimilarity and contrast.

Because the classification accuracy depends on the choice of training and test sets, the dataset is processed by a k-fold cross-validation strategy [15], for obtaining different results on multiple training and test sets. The dataset is randomly divided into k folds. Then, each fold is considered as the test set and the remaining $k-1$ folds are considered as the training set. Each fold has roughly equal dimension and roughly the same language class proportions as in the dataset. Consequently, the test set is composed of a small number of texts in Serbian and a number of texts in world languages. Model learning by K-NN and NB is performed each time by using the current training set, then classification evaluation is established on the current test set. The K value of the K-Nearest Neighbor has been fixed to small odd values (see Sect. 2.4), i.e. 1, 3 and 5. Furthermore, the K-NN algorithm has been executed with the traditional Euclidean distance, which revealed to be particularly reliable in this context with respect to other distance measures. Because the feature vectors have numerical values, probabilities of the NB have been computed by using (4).

Our task is in the domain of binary classification, because we need to classify a model and correctly predict the classes that represent texts written in the world languages (German, Spanish, English and French) and in minority Serbian language. The problem of language classification and identification is similar to the information retrieval one. Hence, precision, recall and f-measure are preferred metrics for the evaluation of the proposed algorithm [6]. They are calculated from the confusion matrix between the classification results obtained by the test set and the ground truth partitioning of the test set in Serbian and world languages. Performance measures have been computed for each selection of the test and training folds and the average values together with the standard deviation have been reported for each of the measures. Furthermore, k-fold cross-validation has



(a) μ_x and μ_y

(b) σ_x and σ_y

(c) energy

(d) entropy

(e) maximum

(f) dissimilarity

(g) contrast

(h) indvmoment

(i) homogeneity

(j) correlation

Fig. 6. The twelve GLCM features obtained by the dataset for the world and Serbian languages in a min-max manner

Table 2. GLCM feature values from Fig. 6 in the min-max manner.

		German	Spanish	Serbian	English	French
μ_x	min.	1.6418	1.6557	1.3700	1.6809	1.6007
	max.	1.8090	1.8353	1.4382	1.7610	1.6634
μ_y	min.	1.6399	1.6526	1.3691	1.6813	1.5999
	max	1.8081	1.8329	1.4375	1.7612	1.6618
σ_x	min.	0.9865	1.0088	0.6153	1.0209	0.9904
	max.	1.0586	1.0665	0.7003	1.0611	1.0303
σ_y	min.	0.9857	1.0077	0.6151	1.0204	0.9901
	max.	1.0585	1.0655	0.7002	1.0606	1.0298
Energy	min.	0.2388	0.2431	0.2395	0.2602	0.2928
	max.	0.3041	0.2911	0.2811	0.2890	0.3270
Entropy	min.	−1.6149	−1.6106	−1.6641	−1.6062	−1.5151
	max.	−1.4249	−1.4624	−1.5171	−1.5180	−1.4245
Maximum	min.	0.3639	0.3283	0.3805	0.4136	0.4684
	max.	0.4695	0.4496	0.4401	0.4636	0.5127
Dissimilarity	min.	0.9913	1.0364	0.6593	0.9385	0.9300
	max.	1.1361	1.2244	0.8037	1.0867	1.0324
Contrast	min.	2.0975	2.2870	0.9201	2.0532	2.0479
	max	2.5232	2.7229	1.2532	2.4019	2.3350
Invdmoment	min.	0.5645	0.5309	0.6431	0.5882	0.6141
	max.	0.6211	0.6068	0.6965	0.6422	0.6468
Homogeneity	min.	0.6356	0.6081	0.6685	0.6549	0.6768
	max.	0.6832	0.6717	0.7134	0.7005	0.7038
Correlation	min.	−0.1833	−0.2696	−0.2779	−0.1055	−0.1130
	max.	−0.0488	−0.1068	−0.1758	0.0498	−0.0122

been repeated separately for three different values of k equal to 2, 5 and 10 [15]. Finally, for avoiding the dependence of the classification results from the particular division in folds, k-fold cross-validation has been executed 50 times for each value of k.

The classification results obtained by the proposed method using K-NN or NB classifier are very positive. Evaluation reveals a perfect identification of the Serbian texts in the test set. In fact, precision, recall and f-measure obtain a value of 1 in all the cases, when the k value of fold cross-validation is equal to 2, 5 and 10 for all the 50 runs and when the K value of the Nearest Neighbor classifier is fixed to 1, 3 and 5. The classification results of the proposed method are compared with those obtained by the n-gram language model. In particular, each text in the dataset is represented by the normalized frequency values of the extracted bi-grams. The same classification experiment is performed with the bi-gram feature vectors by adopting K-NN and NB algorithms and k-fold cross-validation. Also, the same parameter values for the classifiers are selected in the experiment with bi-grams.

Table 3 reports the classification results of bi-grams with the K-NN classifier and Euclidean distance at the different values of $k = 2, 5, 10$ folds and with $K = 1, 3, 5$.

Table 3. Average results in terms of precision, recall and f-measure, together with the standard deviation (in parenthesis), obtained by bi-grams and K-NN classifier using k-fold cross-validation.

		2-fold		5-fold		10-fold	
		World lang	Serbian	World lang	Serbian	World lang	Serbian
$K = 1$	*Precision*	0.9994	1.0000	0.9997	1.0000	1.0000	1.0000
		(0.0001)	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0000)
	Recall	1.0000	0.9962	1.0000	0.9984	1.0000	1.0000
		(0.0000)	(0.0054)	(0.0000)	(0.0036)	(0.0000)	(0.0000)
	F-Measure	0.9997	0.9979	0.9998	0.9990	1.0000	1.0000
		(0.0004)	(0.0030)	(0.0003)	(0.0022)	(0.0000)	(0.0000)
$K = 3$	*Precision*	0.9935	1.0000	0.9988	1.0000	1.0000	1.0000
		(0.0091)	(0.0000)	(0.0026)	(0.0000)	(0.0000)	(0.0000)
	Recall	1.0000	0.9587	1.0000	0.9924	1.0000	1.0000
		(0.0000)	(0.0584)	(0.0000)	(0.0170)	(0.0000)	(0.0000)
	F-Measure	0.9967	0.9771	0.9994	0.9950	1.0000	1.0000
		(0.0046)	(0.0324)	(0.0013)	(0.0112)	(0.0000)	(0.0000)
$K = 5$	*Precision*	0.9859	1.0000	0.9954	1.0000	0.9976	1.0000
		(0.0134)	(0.0000)	(0.0103)	(0.0000)	(0.0077)	(0.0000)
	Recall	1.0000	0.9120	1.0000	0.9710	1.0000	0.9843
		(0.0000)	(0.0854)	(0.0000)	(0.0648)	(0.0000)	(0.0495)
	F-Measure	0.9929	0.9519	0.9976	0.9831	0.9987	0.9893
		(0.0068)	(0.0475)	(0.0053)	(0.0379)	(0.0040)	(0.0337)

Although precision, recall, and f-measure are in the range of 0.91–1.00, it is worth noting that bi-grams are not able to obtain the perfect identification of the minority language among the world languages in all the cases. In fact, in 10-fold the f-measure value for Serbian class is in the range 0.98–1.00. However, in 5-fold it varies in the range 0.98–0.99. Finally, in 2-fold the f-measure value is in the range 0.95–0.99, depending on the K value.

Table 4 shows the classification results of bi-grams with the NB classifier at the different values of $k = 2, 5, 10$ folds. We may observe a poor classification result w.r.t that obtained by our method. In particular, the method is mostly poor in recognition of the minority Serbian language, with the highest f-measure

Table 4. Average results in terms of precision, recall and f-measure, together with the standard deviation (in parenthesis), obtained by bi-grams and NB classifier using k-fold cross-validation.

	2-fold		5-fold		10-fold	
	World lang	Serbian	World lang	Serbian	World lang	Serbian
Precision	0.9141	0.6409	0.8837	0.2000	0.8701	0.0867
	(0.0754)	(0.4821)	(0.0495)	(0.3332)	(0.0433)	(0.2574)
Recall	0.9984	0.4045	1.0000	0.1690	0.9992	0.0850
	(0.0022)	(0.5310)	(0.0000)	(0.3187)	(0.0024)	(0.2477)
F-Measure	0.9532	0.4356	0.9375	0.1747	0.9296	0.0847
	(0.0397)	(0.5410)	(0.0269)	(0.3180)	(0.0234)	(0.2491)

value of 0.44 in the 2-fold. Also, classification of the world languages is not perfect, reaching a peak of 0.95 in the 2-fold.

Figure 7 illustrates the f-measure values obtained by our method and by bi-grams using K-NN and NB classifiers for the different k-folds. Bars represent the f-measure values obtained by bi-grams. The black dashed lines represent the value of f-measure equal to 1 obtained by our method in all cases. This graphical comparison confirms that our method outperforms the competitor in most cases.

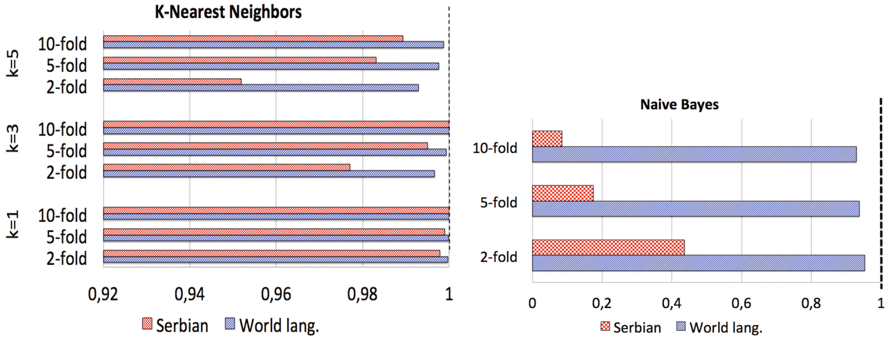


Fig. 7. F-measure values obtained by our method and by the bi-grams feature representation. Bars represent the f-measure values obtained by bi-grams. The black dashed lines represent the value of f-measure equal to 1 obtained by our method in all cases

These results indicate that the bi-grams are not totally robust and reliable in the identification of a minority language such as Serbian among the world languages. Also, the tri-grams did not obtain a meaningful improvement w.r.t. bi-grams, similarly as in [3]. Consequently, their results will be omitted. On the contrary, the proposed feature representation demonstrated to be robust and effective in solving the same task on the same dataset and in the same operating conditions when two different machine learning algorithms are employed. Hence, it is a very promising approach in language identification.

The experiment has been performed in MATLAB R2015a, on a Desktop computer quad-core 2.3 GHz CPU with 8 GB RAM and operating system Windows 7. Based on these specifications, the CPU time for the feature extraction procedure is below 0.1 s. Again, the CPU time for feature classification by K-NN and NB is around 0.003 s. Obviously, the method has proven to be computationally non-intensive.

5 Conclusion

This paper proposed an image-based method for minority language identification by considering statistical analysis of the text based on the text-line status of each script element. The statistical analysis was performed by the grey level co-occurrence matrix. Due to the difference in the language characteristics, the results of the statistical analysis in terms of features showed significant dissimilarity. Classification of the introduced features was performed by the well-known supervised learning algorithms K-Nearest Neighbors and Naive Bayes. The proposed method was tested on text excerpts from a custom oriented dataset. It incorporated texts given in German, Spanish, English, French (world languages) and Serbian (minority language). The experiments showed encouraging results: minority language was perfectly classified by the proposed method. Furthermore, a comparison with the *n*-gram language model demonstrated the superiority of the proposed feature representation in minority language identification. The research presented in this manuscript can be used for language identification on the Web, in preprocessing steps of OCR, and for video text identification.

Future work will enlarge the dataset with more complex samples and will employ other well-known classification algorithms for the experiment, such as deep learning based approaches.

Acknowledgments. This work was partially supported by the Grant of the Ministry of Education, Science and Technological development of the Republic Serbia within the project TR33037.

References

1. 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., 1–19 (2016, in press). doi:10.1080/0952813X.2016.1264090
2. Brodić, D., Amelio, A., Milivojević, Z.N.: Language discrimination by texture analysis of the image corresponding to the text. Neural Comput. Appl., 1–22 (2016, in press). doi:10.1007/s00521-016-2527-x
3. Brodić, D., Amelio, A., Milivojević, Z.N.: Clustering documents in evolving languages by image texture analysis. Appl. Intell. **46**(4), 916–933 (2017)
4. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Document Analysis and Information Retrieval, Las Vegas, USA, pp. 161–175 (1994)

5. Clausi, D.A.: An analysis of co-occurrence texture statistics as a function of grey level quantization. Can. J. Remote Sens. **28**(1), 45–62 (2002)
6. Confusion Matrix. http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html
7. Dasarathy, B.V.: Nearest Neighbor: Pattern Classification Techniques (Nn Norms: Nn Pattern Classification Techniques). IEEE Computer Society Press, Los Alamitos (1990)
8. Dunning, T.: Statistical Identification of Language. Technical report MCCS 94–273, New Mexico State University (1994)
9. Dunning, T.: Statistical Identification of Language. Technical report CRLMCCS-94-273, Computing Research Lab, New Mexico State University (1994)
10. Eleyan, A., Demirel, H.: Co-occurrence matrix and its statistical features as a new approach for face recognition. Turkish J. Electr. Eng. Comput. Sci. **19**(1), 97–107 (2011)
11. Elkan, C.: Nearest Neighbor Classification (2011). http://cseweb.ucsd.edu/~elkan/250Bwinter2010/nearestn.pdf
12. Grefenstette, G.: Comparing two language identification schemes. In: Statistical Analysis of Textual Data, Rome, Italy, pp. 1–6 (1995)
13. Grothe, L., De Luca, E.W., Nurnberger, A.: A comparative study on language identification methods. In: Language Resources and Evaluation, Marrakech, Morocco, pp. 980–985 (2008)
14. Haralick, R.M., Shanmugan, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1978)
15. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2009)
16. Kornai, A.: Digital language death. PLoS ONE **8**(10), 1–11 (2013)
17. Newsam, S., Kamath, C.: Comparing shape and texture features for pattern recognition in simulation data. In: Image Processing: Algorithms and Systems IV, San Jose, USA, pp. 1–14 (2005)
18. Padro, M., Padro, L.: Comparing methods for language identification. In: XXCongreso de la Sociedad Espanola para el Procesamiento del Lenguage Natural, Barcelona, Spain, pp. 155–161 (2004)
19. Proietti, A., Panella, M., Leccese, F., Svezia, E.: Dust detection and analysis in museum environment based on pattern recognition. Measurement **66**, 62–72 (2015)
20. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd ed. Prentice Hall (2003). [1995]
21. Sibun, P., Spitz, A.L.: Language determination: natural language processing from scanned document images. In: 4th Conference on Applied Natural Language Processing, Stuttgart, Germany, pp. 15–21 (1994)
22. Souter, C., Churcher, G., Hayes, J., Hughes, J., Johnson, S.: Natural language identification using corpus-based models. Hermes J. Linguist. **13**, 183–203 (1994)
23. Takcı, H., Soğukpınar, İ.: Letter based text scoring method for language identification. In: Yakhno, T. (ed.) ADVIS 2004. LNCS, vol. 3261, pp. 283–290. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30198-1_29
24. Wackerly, D.D., Mendenhall, W., Scheaffer, R.L.: Mathematical Statistics with Applications. Duxbury Press, Belmont (1996)
25. Web 2014. http://w3techs.com/technologies/overview/content_language/all
26. Zramdini, A.W., Ingold, R.: Optical font recognition using typographical features. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 877–882 (1998)