Line-Level Script Identification for Six Handwritten Scripts Using Texture Based Features

Pawan Kumar Singh, Ram Sarkar and Mita Nasipuri

Abstract Script identification from a given document image has some important applicability in many computer applications such as automatic archiving of multilingual documents, searching online archives of document images and for the selection of script specific Optical Character Recognition (OCR) engine in any multilingual environment. In this paper, we propose a texture based approach for text line-level script identification of six handwritten scripts *namely*, *Bangla*, *Devnagari*, *Malayalam*, *Tamil*, *Telugu* and *Roman*. A set of 80 features based on Gray Level Co-occurrence Matrix (GLCM) has been designed for the present work. Multi Layer Perceptron (MLP) is found to be the best classifier among a set of popular multiple classifiers which is then extensively tested by tuning different parameters. Finally, an accuracy of 95.67 % has been achieved on a dataset of 600 text lines using 3-fold cross validation with epoch size 1,500 of MLP classifier.

Keywords Script identification · Handwritten documents · Texture based feature · Gray level Co-occurrence matrix · Multiple classifiers

1 Introduction

A very important task in the field of automated document analysis system is OCR, which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form. Till date, many algorithms have been presented in the literature to perform this task for a specific

P.K. Singh (✉) · R. Sarkar · M. Nasipuri
Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
e-mail: pawansingh.ju@gmail.com

R. Sarkar
e-mail: raamsarkar@gmail.com

M. Nasipuri
e-mail: mitanasipuri@gmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 340,
DOI 10.1007/978-81-322-2247-7_30

language/script. Almost all existing works on OCR make an implicit assumption that the script type of the document to be processed is known beforehand. In a multilingual environment, such document processing systems relying on OCR would clearly need human intervention to select the appropriate OCR package, which is certainly inefficient and undesirable. On the contrary, it is quite impossible to design a single OCR system which can recognize a reasonable number of scripts. Therefore, it is a general requirement to identify the script type before feeding the document image to their corresponding OCR engine.

Script identification from handwritten documents is a challenging process due to the following reasons: (1) difficulties addressed in pre-processing such as ruling lines, skewness, noise, etc., (2) complexity in feature selection due to large set of symbols, and (3) sensitivity of the scheme due to huge variations in handwriting styles. In general, script identification can be achieved at three levels *namely*, Page-level, Text line-level and Word-level. The ability to reliably identify the script type using least amount of textual data is essential when dealing with document pages that contain text words of different scripts. But identifying text words of different scripts with only a few numbers of characters may not always be feasible because at word-level, the number of characters present in a single word may not be always informative. On the contrary, identifying scripts at page-level can be sometimes too complicated and laborious. So, it would be better to perform the script identification at text line-level.

In the context of *Indic* scripts, most of the published methodologies [1–6] on text line-level script identification, have considered the printed text documents. A few number of research works [7, 8] are applied on handwritten text lines. Despite these research contributions, it can be noticed that most of works have been done for bilingual or trilingual scripts. But, this is a pure limitation in a multilingual country like India, where people residing at different locations use different script. So, in Indian context, a script recognition system should have the ability to identify more number of *Indic* scripts. This has motivated us to take the challenge of identifying script type of the handwritten text lines written in five *Indic* scripts *namely*, *Bangla*, *Devnagari*, *Malayalam*, *Tamil*, and *Telugu* along with *Roman* script. We have included *Roman* script in our work as this script is used in all official works in almost every state of India.

2 Proposed Work

The proposed work is based on a simple observation that every script/language consists of a finite set of characters, each having a distinct visual appearance, which serves as useful visual clues to recognize the script. That is why, we have been inspired to use texture based GLCM feature for identifying handwritten scripts.

2.1 Gray Level Co-occurrence Matrix (GLCM)

GLCM estimates the properties of an image related to second order statistics which considers the relationship among pixels or groups of pixels. Haralick [9] suggested the use of GLCM which has become one of the most well known and widely used texture features. This method is based on the joint probability distributions of pairs of gray levels. GLCM shows how often each gray level occurs at a pixel located at a fixed geometric position relative to other pixel, as a function of the gray level [10]. Mathematically, for a given image I of size $M \times N$ and for a displacement vector $d\left(d_x, d_y\right)$, the GLCM is defined to be a square matrix P of size $L \times L$ where, L is the number of gray level range (0, 1, …, L − 1) in the image.

$$P(i,j) = \sum_{x=1}^{M} \sum_{y=1}^{N} \begin{cases} 1 & \textit{if } I(x,y) = i \text{ and } I\left(x + d_x, y + d_y\right) = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here, i and j are the intensity values of the image I, x and y are the spatial pixel positions in the image I and the offset $\left(d_x, d_y\right)$ depends on the direction θ and the distance d for which the matrix is computed. Here, $p(i,j)$ is a count of the number of times $I(x,y) = i$ and $I\left(x + d_x, y + d_y\right) = 1$ occur in image I. Figure 1 illustrates the co-occurrence matrices along four directions $(0°, 45°, 90° \text{ and } 135°)$ considering a 2 × 2 image represented with two gray-tone values 0 and 1. For this purpose, we have considered two neighboring pixels ($d = 1$ and $d = 2$) along four possible directions.

A set of 10 features based on GLCM have been extracted which are described below in detail.

Energy Energy, also known as uniformity, measures the image homogeneity. It is the sum of squares of entries in the GLCM. Energy is high when image has very good homogeneity or when pixels are very similar. Energy [11] is calculated as

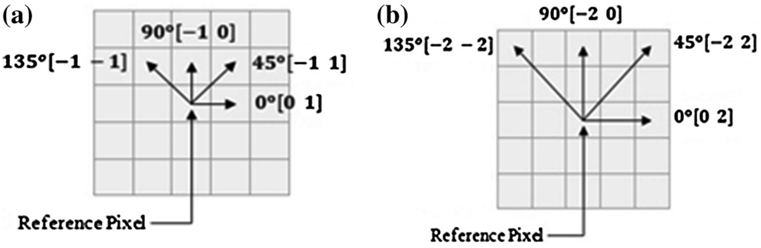

Fig. 1 Illustration of co-occurrence matrices along four directions (0°, 45°, 90° and 135°) for extracting texture features where **a** d = 1 and **b** d = 2

$$Energy = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} P^2(i,j) \tag{2}$$

Entropy Entropy shows the amount of information of the image that is needed for the image compression. Entropy measures the loss of information or message in a transmitted signal [9]. Entropy is calculated as

$$Entropy = -\sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} P(i,j) \cdot \ \log P(i,j) \tag{3}$$

Inertia Inertia [12] is the measure of the amount of the local variations. A large amount of variation gives a large inertia. It is defined as

$$Inertia = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} (i-j)^2 \cdot P(i,j) \tag{4}$$

Autocorrelation Autocorrelation [9] measures the linear spatial relationship between spatial sizes of texture primitives. Autocorrelation-based approach to texture analysis estimates the intensity value concentration on all or part of an image represented as a feature vector. It is defined as

$$Autocorrelation = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} i \cdot j \cdot P(i,j) \tag{5}$$

Covariance Covariance [12] provides a measure of the strength of the correlation between two or more distinct pairs of adjacent pixel distribution. If the change in one pair of pixel distribution corresponds with the change in the other one, i.e., the pixel distribution tends to show similar behavior, the covariance is positive otherwise it is negative. It can be defined as

$$Covariance = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} (i - M_x)\left(j - M_y\right) \cdot P(i,j) \tag{6}$$

$$M_x = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} i \cdot P(i,j) \tag{7}$$

$$M_y = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} j \cdot P(i,j) \tag{8}$$

Contrast The contrast [11] is a difference of moment of the GLCM. It is the measure of the amount of local variation present in the image. It can be thought of as a linear dependency of gray levels of neighboring pixels [9]. It is defined as

$$Contrast = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} P(i,j)\cdot|i-j|^{k}, \quad k\in\mathbb{Z} \tag{9}$$

In the present work, the value of k, which is chosen to be 4 proves to be optimal.

Local Homogeneity Local homogeneity [11] is the closeness of gray levels in the spatial distribution over the image. Homogeneous textured image comprises of a limited range of gray levels and hence, the GLCM image exhibits a few values with relatively high probability. It is defined as

$$Local\,Homogeneity = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} \frac{1}{1+(i-j)^2} P(i,j) \tag{10}$$

Cluster Shade Cluster shade [11] is the measure of skewness of the co-occurrence matrix, in other words, the lack of symmetry. It is defined as

$$Cluster\,Shade = \sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} \left(i-M_x+j-M_y\right)^3\cdot P(i,j) \tag{11}$$

Cluster Prominence Cluster prominence [11] is also the measure of skewness of the co-occurrence matrix. When cluster prominence is high, the image is not symmetric. In addition, when cluster prominence is low, there is a peak in the co-occurrence matrix around the mean values. Low cluster prominence implies little variation in gray-scales and can be defined as

$$Cluster\,Prominence = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} \left(i-M_x+j-M_y\right)^4\cdot P(i,j) \tag{12}$$

Information Measure of Correlation Correlation measures the linear dependency of gray levels of neighboring pixels. Digital Image Correlation is an optical method that employs tracking and image registration techniques for accurate 2-D and 3-D measurements of changes in images. This is often used to measure deformation, displacement, strain and optical flow [10] and can be defined as

$$Information\,Measure\,of\,Correlation = \frac{-\sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} P(i,j)\log P(i,j) - H_{xy}}{\max\left(H_x, H_y\right)} \tag{13}$$

where,

$$H_{xy} = -\sum_{i=0}^{\mathrm{L}-1}\sum_{j=0}^{\mathrm{L}-1} P(i,j)\cdot\log\left(\sum_{j=0}^{\mathrm{L}-1} P(i,j)\cdot\sum_{i=0}^{\mathrm{L}-1} P(i,j)\right) \tag{14}$$

$$H_x = -\sum_{i=0}^{\mathrm{L}-1}\left\{\sum_{j=0}^{\mathrm{L}-1} P(i,j)\cdot\log\sum_{j=0}^{L-1} P(i,j)\right\} \tag{15}$$

$$H_y = -\sum_{j=0}^{\mathrm{L}-1}\left\{\sum_{i=0}^{\mathrm{L}-1} P(i,j)\cdot\log\sum_{i=0}^{L-1} P(i,j)\right\} \tag{16}$$

Due to the binary nature of the document images from which the features are estimated, the extraction of such features is unnecessary and indeed counterproductive. Since, there are only two gray levels, the matrices will be of size 2 × 2, i.e., it is possible to fully describe each matrix with only three unique parameters due to the diagonal symmetry property [11]. For each of the 10 measurements defined above, the values of $d = \{1, 2\}$ and $\theta \in 0^\circ, 45^\circ, 90^\circ$ and 135° lead to a total of 80 (10 × 8) features using GLCM.

3 Experimental Results and Discussion

The performance of the proposed approach has been tested using a dataset of 600 text lines written in six handwritten scripts *namely*, *Bangla*, *Devnagari*, *Malayalam*, *Tamil*, *Telugu* and *Roman.* Here, each script consists of an equal number of text lines which are extracted from the document images by using piecewise water flow technique as described in [13]. The extracted text lines are stored as gray scale images. Noise is removed by using Gaussian filter [12]. Text lines are then binarized using well known Otsu's global thresholding approach [14]. The proposed approach is then applied on the preprocessed text lines and evaluated using seven well-known classifiers (with the help of a software tool known Weka [15]) *namely*, Naïve Bayes, Bayes Net, MLP, Support Vector Machine (SVM), Random Forest, Bagging and MultiClass Classifier. The recognition performances and their corresponding scores achieved at 95 % confidence level are shown in Table 1.

The accuracy achieved by MLP classifier (as evident form Table 1) shows that it has the ability to perform better if executed comprehensively using different tuning parameters. For this purpose, we have used 3-fold, 5-fold and 7-fold cross validation schemes with varied epoch sizes of MLP classifiers (see Table 2). From the table, it is observed that for 3-fold cross validation with epoch size 1,500 of MLP, the best identification accuracy achieved is 95.67 %.

Table 1 Recognition performances of the proposed script identification technique using seven well-known classifiers

	Classifiers						
	Naïve Bayes	Bayes Net	MLP	SVM	Random forest	Bagging	MultiClass classifier
Success rate (%)	88.54	88.92	**93.8**	91.45	90.62	89.1	90.51
95 % confidence score (%)	93.26	94.39	**97.78**	96.02	95.37	94.84	96.55

Best cases are shaded in gray as well as styled in bold

Table 2 Recognition accuracies of script identification technique for different folds and different number of epochs of MLP classifier

	Success rate of MLP classifier (%)		
	#-Fold		
Epoch size	3-fold	5-fold	7-fold
500	93.18	94.25	92.49
1,000	94.3	93.82	93.9
1,500	**95.67**	95.16	94.73

The best performance is shaded in gray as well as styled in bold

Though Table 2 shows encouraging results but some misclassifications occur during the experimentation. The main reasons are presence of noise and poor segmentation of text lines due to skewness, punctuation symbols, etc. in the document images. The structural resemblance in the character set of the scripts like *Bangla* and *Devnagari* causes similarity in the adjacent pixel distribution which in turn misclassifies them among each other. Due to similar reason, *Telugu* and *Malayalam* scripts have been misclassified among each other. Figure 2 shows some instances of misclassification of the present technique.

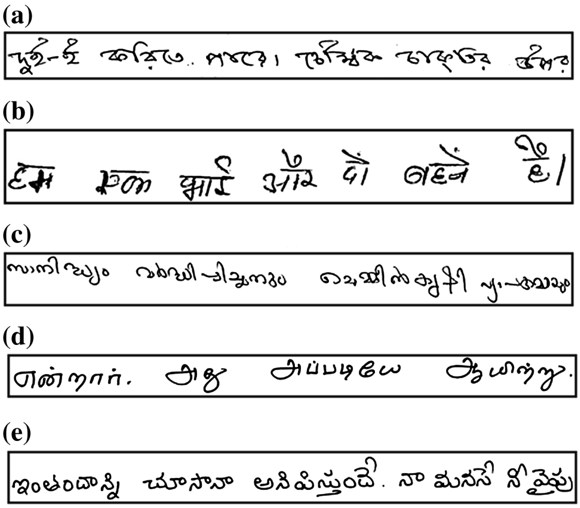

Fig. 2 Sample text lines written in **a** *Bangla*, **b** *Devnagari*, **c** *Malayalam*, **d** *Tamil*, and **e** *Telugu* scripts misclassified by the present technique as *Devnagari*, *Bangla*, *Telugu*, *Roman*, *Malayalam* scripts respectively

4 Conclusion

Based on the observation of human ability to classify dissimilar scripts, we have inspected the possibility of using only global analysis of scripts for identifying them. In this paper, a texture based approach for text line-level script identification from handwritten documents is presented. An 80-element feature set based on GLCM is applied in the present work to distinguish six popular scripts used in India. Experiments are performed by considering only the text lines from the document images and an overall classification rate of 95.67 % is achieved. Although the present technique is evaluated on a limited dataset, we have achieved encouraging results. The work, presented here, is a footstep towards building a general multi-script OCR system pertinent to Indian subcontinent that can work for all official *Indic* scripts.

References

1. Pal, U., Chaudhuri, B.B.: Script line separation from indian multi-script documents. In: Proceedings of 5th International Conference on Document Analysis and Recognition (ICDAR), pp. 406–409. (1999)
2. Pal, U., Chaudhuri, B.B.: Identification of different script lines from multi-script documents. Image Vis. Comput. **20**(13–14), 945–954 (2002)
3. Pal, U., Sinha, S., Chaudhuri, B.B.: Multi-script line identification from indian documents. In: Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 880–884. (2003)
4. Joshi, G.D., Garg, S., Sivaswamy, J.: Script identification from Indian documents. In: International Workshop Document Analysis Systems, Nelson. Lecture Notes in Computer Science, vol. 3872, pp. 255–267. (2006)
5. Padma, M.C., Vijaya, P.A.: Identification of Telugu, Devnagari and English scripts using discriminating features. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **1**(2), 64–78 (2009)
6. Gopakumar, R., SubbaReddy, N.V., Makkithaya, K., Dinesh Acharya, U.: Script identification from multilingual indian documents using structural features. J. Comput. **2**(7), 106–111 (2010)
7. Chaudhuri, B.B., Bera, S.: Handwritten text line identification in Indian scripts. In: Proceedings of 10th International Conference on Document Analysis and Recognition, pp. 636–640. (2009)
8. Hangarge, M., Dhandra, B.V.: Offline handwritten script identification in document images. Int. J. Comput. Appl. (IJCA) **4**(6), 6–10 (2010)
9. Haralick, R.M., Shanmungam, K., Dinstein, I.: Textural features of image classification. IEEE Trans. Syst. Man, Cybern. **3**, 610–621 (1973)
10. Haralick, R.M., Watson, L.: A facet model for image data. Comput. Vision Graph. Image Process. **15**, 113–129 (1981)
11. Busch, A., Boles, W.W., Sridharan, S.: Texture for script identification. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1720–1732 (2005)
12. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, vol. I. PHI, New Delhi (1992)

13. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M.: Extraction of text lines from handwritten documents using piecewise water flow technique. J. Intell. Syst. **23**(3), 245–260 (2014)
14. Ostu, N.: A thresholding selection method from gray-level histogram. IEEE Trans. Syst. Man Cybern. **SMC-8**, 62–66 (1978)
15. www.cs.waikato.ac.nz/ml/weka/documentation.html