Statistical Textural Features for Text-Line Level Handwritten *Indic* Script Identification

Pawan Kumar Singh, Ram Sarkar and Mita Nasipuri

Abstract As India is a multilingual country, hence, a variety of scripts are used here to write different languages. However, it becomes essential to recognize a particular script before the selection of an appropriate Optical Character Recognition (OCR) system. The research in this field is comparatively less explored and further research is required, particularly in the field of handwritten documents. This paper presents a robust script identification technique for 11 official handwritten *Indic* scripts *namely, Bangla, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Manipuri, Oriya, Tamil, Telugu, Urdu* along with *Roman* script. The recognition is performed at text-line level by using statistical textural features called Neighborhood Gray-Tone Difference Matrix along with Gray-level Run Length Matrix. The proposed method is experimented on a total dataset of 2400 handwritten text-lines of various scripts and yielded an identification rate of 97.69% using Multi Layer Perceptron (MLP) classifier.

Keywords Handwritten script identification · *Indic* scripts · Statistical textural features · Neighborhood gray-tone difference matrix · Gray-level run length matrix · Multiple classifiers

P.K. Singh (✉) · R. Sarkar · M. Nasipuri
Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, West Bengal, India
e-mail: pawansingh.ju@gmail.com

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

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

R. Chaki et al. (eds.), *Advanced Computing and Systems for Security*,
Advances in Intelligent Systems and Computing 568,
DOI 10.1007/978-981-10-3391-9_9

1 Introduction

Script is considered as a graphic form which is used in any writing system. The languages used in the human society are typeset with the different scripts. A script can be used by only one language or it can be shared by several languages, with or without any variations [1]. India has 23 languages [2], recognized by constitution, *viz., Assamese*, *Bengali*, *Bodo*, *Dogari*, *Kannada*, *Hindi*, *Sindhi*, *Nepali*, *Urdu*, *Punjabi*, *Marathi*, *Gujarati*, *Oriya*, *Sanskrit*, *Tamil*, *Telugu*, *Malayalam*, *Kashmiri*, *Manipuri*, *Konkani*, *Maithali*, *Santhali* and *English*. The 12 major scripts used to write these languages are: *Bangla*, *Devanagari*, *Gujarati*, *Gurumukhi*, *Manipuri*, *Malayalam*, *Oriya*, *Tamil*, *Telugu*, *Kannada*, *Roman* and *Urdu*. Among these, only *Urdu* is written from right to left whereas the rest of the scripts are written from left to right. The first 10 scripts, originated from the ancient *Brahmi* script, are also known as *Indic* scripts.

In general, any OCR system is used to recognize only a script of particular type, and for the same reason, it is not viable to model a single OCR system for recognizing variety of scripts/languages. Hence, one can think of making a pool of OCR engines which correspond to different scripts in a multi-lingual environment. However, for this, it is necessary to have the knowledge about the script used to write the document. Hence, this necessity could be fulfilled if the researchers can be able to design an automatic script recognition module for the multi-script scenario.

An automatic script identification module would be used to sort or search the relevant information when the domain is multilingual/multi-script or even it helps to index/categorize the documents images on the basis of its script type. When a script is used to write only one language, then the script recognition technique can also be considered as language recognition technique. Otherwise, script recognition is the first step of classification followed by language identification among the languages which share a common script.

Different methodologies have been reported in the literature for accomplishing this task, sometimes with high degree of accuracy. A comprehensive survey based on script recognition techniques had been prepared by Singh et al. [1], with emphasis on script identification in both printed and handwritten *Indic* scripts scenario. Script identification for printed documents at page level [3–5], text-line level [6–10], and word level [11–16] have been found in the literature. On the contrary, only few works have been done for handwritten script identification at page level [17], text-line level [18, 19], and word level [20–22]. Singh et al. [17] developed a page-level script identification technique for handwritten document pages using Gray Level Co-occurrence Matrix (GLCM). The proposed technique had been experimented on four scripts *namely*, *Bangla, Devanagari, Telugu,* and *Roman* and the system was found to identify 91.48% scripts successfully using MLP classifier. Hangarge et al. [18] described a set of 13 spatial spread features of the three scripts *namely*, *English*, *Devanagari* and *Urdu* which were extracted using morphological filters. Experiments were carried out with k-NN classifier by varying the number of neighbors ($k = 3, 5, 7, 9, 11, 13, 15$) and the performance of the technique was optimal when the value of k is set to 3. The overall recognition accuracies of the proposed system were found to be 88.67% and 99.2% for tri-script and bi-script cases respectively. Singh et al. [19] proposed a texture based concept for script identification at text-line level for six handwritten scripts *namely*, *Bangla*, *Devanagari*, *Malayalam*, *Tamil*, *Telugu* and *Roman*. An accuracy of 95.67% had been achieved using 3-fold cross validation of MLP classifier. Roy et al. [20] described a scheme for word-wise identification of handwritten *Roman* and *Oriya* scripts for Indian postal automation using water-reservoir and topological features. The overall accuracy rate achieved on the test dataset was found to be 99.6% and 97.69% respectively. Sarkar et al. [21] presented a system, which identified the scripts of the handwritten words from the document images, written in *Bangla* or *Devanagari* mixed with *Roman* scripts with the help of eight holistic features. The recognition performances of 99.29% and 98.43% had been achieved on the test sets of *Bangla-English* words and *Devanagari-English* words respectively. P.K. Singh et al. reported a technique [22] which recognized the scripts of handwritten words from a document page, written in *Devanagari* script mixed with *Roman* script. A set of 39 distinctive features using topological along with convex hull based features were designed for the recognition purpose and the overall script identification accuracy of 99.54% was achieved. However, a major limitation of the above works is that researchers have considered only a few *Indic* scripts. This has been a major point of motivation behind developing a robust handwritten script identification technique including all the official *Indic* scripts along with *Roman* script.

2 Challenges Related to Handwritten Script Identification

There are some unique challenges that must be addressed in the domain of handwritten script recognition system. Among many, two basic problems are: *inter-writer* variability and *intra-writer* variability. *Inter-writer* variability encompasses the variations seen among different writers i.e., different writers will invariably have different writing styles. In contrast, *intra-writer* variability takes into consideration that the same writer tends to write the same textual content in a different manner depending upon his/her frame of mind. The challenge in this regard is to create a writer-independent script identification system that has the ability to adapt these variations like humans. Another major challenge that the system has to address is the problem of constrained versus unconstrained handwriting. Constrained handwriting refers to handwritten text that conforms to a pre-defined writing constriction, e.g. all the text words in a document image will be discrete and non-touching.

Whereas unconstrained handwriting refers to the fact that the document image may contain discrete and cursive handwriting or a mixture of both, with no restriction on the writers while they write. Apart from this, difficulties inherent in recognizing handwritten scripts pose huge challenges than its printed form. Similarity among different scripts is quite common when the documents are handwritten. The styles of writing for handwritten scripts are more diverse than printed fonts. Also, problems such as existence of ruling lines, noise, skew, quality of ink, age of the document, etc. are commonly seen in handwritten documents. As mentioned earlier, script identification can be achieved at either page level, text-line level or word level. Sometimes, identifying scripts at page-level can be sometimes too convoluted and protracted due to large computational complexity. On the other hand, identifying words written in different scripts using a few characters is definitely a challenging task because the number of characters presents in a single word may not produce significant amount of discriminative information required for identification. Therefore, considering the complexities of the scripts, it would be better to identify the scripts at text-line level compared to page or word-level.

3 Data Collection and Pre-processing

There are no standard databases, considering either handwritten or printed *Indic* script documents, available in public domain which can be successfully used for this experimentation. Hence, we have prepared in-house database of handwritten documents. Different educated people were requested to write few text-lines of his/her choice inside A-4 size pages. Handwritten text-lines were written in 12 official scripts of India as mentioned earlier. It is to be noted that writers involved in the data collection drive belong to different professions. These pages are then scanned at 300 dpi resolution and saved as gray tone images. The noisy pixels therein, if any, are removed by Gaussian filter [23]. It is worth mentioning that the *inter*-word and *intra*-word spaces are very non-uniform in the handwritten text-lines. Numerals of any script which may be present in the text document are not considered for the present work. A text–line whose width is at least 50% of the page width is considered here. A sample snapshot of text-line images written in 12 different scripts is shown in Fig. 1. Otsu's global thresholding approach [24] is used to convert them into two-tone images (0 and 1) where the label '1' represents the object and '0' represents the background. However, the dots and punctuation marks appearing in the text- lines have not been eliminated, since these may also contribute to understand the text in a meaningful way. Finally, 2400 handwritten text-line images are prepared with precisely 200 text-lines per script.

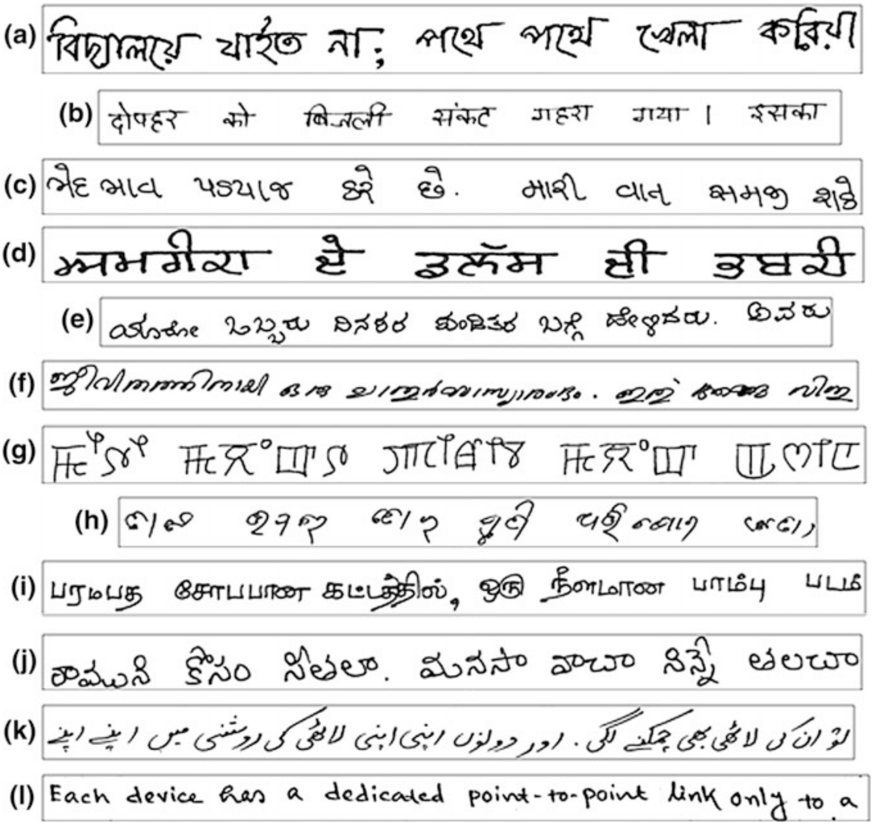

Fig. 1 Sample text-line images written in: **a** *Bangla*, **b** *Devanagari,* **c** *Gujarati,* **d** *Gurumukhi,* **e** *Kannada,* **f** *Malayalam,* **g** *Manipuri,* **h** *Oriya,* **i** *Tamil,* **j** *Telugu,* **k** *Urdu,* and **l** *Roman* scripts respectively

4 Proposed Work

Every script/language, consisting of a finite set of characters, has a distinct visual appearance, which serves as useful visual clues to recognize the script. The current research is inspired by this simple observation of the human beings which also motivates the researchers to design different texture based features. Usually, texture features are designed to capture the granularity and repetitive patterns of local regions seen within an image. Some well-known texture features relying on GLCM and Gabor filter bank consider multiple scales and orientations for feature extraction which in turns involves a high computation cost. The conventional statistical textural features utilized in this paper, are Neighborhood Gray-Tone Difference Matrix (NGTDM) and Gray-level Run Length Matrix (GLRLM). These features are illustrated below in the following subsections.

4.1 Neighborhood Gray-Tone Difference Matrix (NGTDM)

A NGTDM [25] defines the texture measures which are very much correlated with human perception of textures. It calculates the texture using neighborhood intensity differences which will be helpful to describe the local features. The NGTDM are based on the differences between each pixel and the neighboring pixels in the adjacent regions. A NGTDM is basically a column vector of G elements. This vector is populated by computing the difference between the intensity values of a pixel and the mean intensity calculated over a square shaped window centered at that pixel. Suppose the image intensity level $f(x,y)$ at location (x,y) is i, $i=0,1,2,\ldots,L-1$. The mean intensity value of the window centered at (x,y) can be written as:

$$f_i = f(x,y) = \frac{1}{W-1}\sum_{m=-K}^{K}\sum_{n=-K}^{K} f(x+m, y+n) \tag{1}$$

where, K denotes the window size and $W=(2K+1)^2$. The i-th entry of the gray-tone difference matrix is given by:

$$g(i) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} |i - f_i| \tag{2}$$

for the pixels whose intensity value is i. Otherwise, $g(i)=0$.

Five different features are derived from the NGTDM, described below, to quantitatively estimate the following perceptual texture properties:

- **Coarseness**. It finds out the presence of any texture in an image and is measured by the size of the primitives which form the texture. Generally, a coarse texture comprises large sized primitives which are typified by the degree of neighboring uniformity of gray-levels. On the other hand, fine texture can be defined by small primitives and these are described by the degree of neighboring variations of gray-levels.

$$F_{cos} = \left(\epsilon + \sum_{i=0}^{L-1} p_i g(i)\right)^{-1} \tag{3}$$

where, ε is a small number which avoids the coarseness coefficient to become infinite and p_i is the estimated probability of the occurrence of the intensity values i such that

$$p_i = N_i/n \tag{4}$$

with N_i denoting the number of pixels having level i, and $n=(N-K)(M-K)$.

- **Contrast**. It quantifies the amount of clarity with which the different primitives in a texture can be differentiated. A well contrasted image is defined by the primitives which are visible as well as distinguishable. Among the factors that influence contrast, the gray-levels, the ratio of white and black pixels and the frequency of intensity changes of gray-levels are important.

$$F_{con} = \left[\frac{1}{N_t(N_t-1)}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_i p_j (i-j)^2\right]\left[\frac{1}{n}\sum_{i=0}^{L-1} g(i)\right] \tag{5}$$

- **Busyness**. It measures the change of intensity from any pixel to its locality. If the intensity changes are quick and rush then it is called busy texture, whereas if the same are slow and gradual then it is called a non-busy texture. There is a relationship of busyness with the spatial frequency of the intensity changes in an image. Along with that, busyness is also affected by the amplitude of the intensity changes.

$$F_{bsuy} = \frac{\sum_{i=0}^{L-1} p_i g(i)}{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\left|ip_i - jp_j\right|} \forall p_i \neq 0,\ p_j \neq 0 \tag{6}$$

- **Complexity**. This is the visual information of texture. A texture is said to be complex when its information content is very high. This depends on the number of diverse primitives and average intensity values. Complexity is the sum of normalized differences between intensity values measured in pairs. These are weighted by the sum of the elements in the NGTDM corresponding to any two intensity values. Mathematically, it can be written as:

$$F_{com} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\frac{|i-j|}{n\left(p_i + p_j\right)}\left[p_i g(i) + p_j g(j)\right] \forall p_i \neq 0,\ p_j \neq 0 \tag{7}$$

- **Texture Strength**. Strength integrates and summarizes the concepts of busyness and coarseness. An image with a strong texture is composed by easily definable and clearly visible elements. It can be expressed as:

$$F_{str} = \frac{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\left(p_i + p_j\right)(i-j)^2}{\epsilon + \sum_{i=0}^{L-1} g(i)} \forall p_i \neq 0,\ p_j \neq 0 \tag{8}$$

For feature extraction purpose, each of the text-line images, written in different scripts, are firstly divided into 4 sub-images using 2-level quad tree decomposition approach and the five features are then computed from each of these sub-images. Two distances $d = 1$ and $d = 2$ are used in feature computation, corresponding to neighborhood sizes of 3×3 and 5×5 respectively. So, a feature vector of size 40 (F1-F40) is extracted for each text-line images using NGTDM. In the computation of F_{cos} and F_{str}, the value of ϵ is taken as 10^{-7}.

4.2 Gray-Level Run Length Matrix (GLRLM)

The application of a run length matrix for the purpose of texture feature extraction is proposed by Galloway [26]. Let, there is a given image of size $M \times N$, then a run-length matrix $p(i, j)$ is determined as the number of runs of pixels having gray-level i and run length j.

- Short Run Emphasis (SRE):

$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2} = \frac{1}{n_r} \sum_{j=1}^{N} \frac{p_r(j)}{j^2} \tag{9}$$

- Long Run Emphasis (LRE):

$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot j^2 = \frac{1}{n_r} \sum_{j=1}^{N} p_r(j) \cdot j^2 \tag{10}$$

- Gray-Level Non-uniformity (GLN):

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{i=1}^{M} \left[p_g(i) \right]^2 \tag{11}$$

- Run Length Non-uniformity (RLN):

$$RLN = \frac{1}{n_r} \sum_{j=1}^{N} \left(\sum_{i=1}^{M} p(i,j) \right)^2 = \frac{1}{n_r} \sum_{j=1}^{N} [p_r(j)]^2 \tag{12}$$

- Run Percentage (RP):

$$RP = \frac{n_r}{n_p} \tag{13}$$

  In the above equations, n_r is the number of runs whereas n_p is the number of pixels in the image. It is noticed that most of the features are only functions of $p_r(j)$, which do not consider the gray-level information of $p_g(i)$. Chu et al. [27] estimated two features to calculate gray-level information in the matrix.

- Low Gray-Level Run Emphasis (LGRE):

$$LGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2} = \frac{1}{n_r} \sum_{i=1}^{M} \frac{p_g(i)}{i^2} \tag{14}$$

- High Gray-Level Run Emphasis (HGRE):

$$HGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 = \frac{1}{n_r} \sum_{i=1}^{M} p_g(i) \cdot i^2 \tag{15}$$

Further, Dasarathy et al. [28] described another four feature estimation functions based on the concept of combined statistical measure of gray-level and run length, as follows:

- Short Run Low Gray-Level Emphasis (SRLGE):

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 \cdot j^2} \tag{16}$$

- Short Run High Gray-Level Emphasis (SRHGE):

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot i^2}{j^2} \tag{17}$$

- Long Run Low Gray-Level Emphasis (LRLGE):

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot j^2}{i^2} \tag{18}$$

- Long Run High Gray-Level Emphasis (LRHGE):

$$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 \cdot j^2 \tag{19}$$

These features are all based on intuitive reasoning, in an attempt to capture some apparent properties of run-length distribution. For each of the 11 measurements, defined above, the values of $\theta \in 0°, 45°, 90° \, and \, 135°$ lead to a total of 44 (F41-F84) features using GLRLM. Finally, a set of 84 (i.e. 40 + 44) statistical textural features are extracted using both NGTDM and GLRLM for the text-line level classification of twelve different handwritten scripts.

5 Experimental Evaluation and Discussion

The performance of the present script identification scheme is evaluated on a dataset of 2400 preprocessed text-line images as described in Sect. 3. For each 200 text line images of a particular script, 135 images are applied for training and the rest 65 images are applied for testing purpose. Seven well-known classifiers *namely*, Naïve

Table 1 Recognition performances of the proposed script identification technique using seven well-known classifiers (best case is shaded in gray and styled in bold)

	Classifiers						
	Naïve Bayes	**Bayes Net**	**MLP**	**SVM**	**Random Forest**	**Bagging**	**MultiClass Classifier**
Success Rate (%)	89.74	90.95	**97.69**	95.87	94.6	91.18	93.37
95% confidence score (%)	91.19	93.67	**99.85**	97.7	97.39	93.83	95.52

Bayes, Bayes Net, MLP, Support Vector Machine (SVM), Random Forest, Bagging and MultiClass Classifier are used to select the best classifier suitable for the present experimental setup. The recognition performances and their corresponding scores achieved at 95% confidence level are shown in Table 1.

As observed from Table 1 that MLP classifier produces the highest identification accuracy of 97.69%. In the present work, detailed error analysis of MLP classifier with respect to some well-known statistical parameters *namely*, Kappa statistics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-measure, Matthews Correlation Coefficient (MCC) and Area Under ROC (AUC) are also computed. The values of Kappa statistics, mean absolute error, root mean square error of MLP classifier for the present technique are found to be 0.9748, 0.0056 and 0.0557 respectively. Table 2 provides a statistical performance analysis of the remaining parameters for each of the aforementioned scripts.

Table 2 Statistical performance measures along with their respective means (shaded in gray and styled in bold) achieved by the proposed technique for twelve handwritten scripts

Scripts	**TPR**	**FPR**	**Precision**	**Recall**	**F-Measure**	**MCC**	**AUC**
Bangla	0.812	0.000	1.000	0.812	0.896	0.893	0.906
Devanagari	0.990	0.001	0.990	0.990	0.990	0.989	0.999
Gujarati	0.990	0.004	0.962	0.990	0.976	0.973	1.000
Gurumukhi	1.000	0.005	0.953	1.000	0.976	0.974	0.999
Kannada	1.000	0.003	0.971	1.000	0.985	0.984	1.000
Malayalam	0.990	0.001	0.990	0.990	0.990	0.989	0.999
Manipuri	1.000	0.000	1.000	1.000	1.000	1.000	1.000
Oriya	1.000	0.006	0.935	1.000	0.967	0.964	1.000
Tamil	1.000	0.005	0.953	1.000	0.976	0.974	1.000
Telugu	1.000	0.002	0.981	1.000	0.990	0.989	1.000
Urdu	1.000	0.000	1.000	1.000	1.000	1.000	1.000
Roman	0.940	0.000	1.000	0.940	0.969	0.967	0.980
Weighted Average	**0.977**	**0.002**	**0.978**	**0.977**	**0.976**	**0.975**	**0.990**

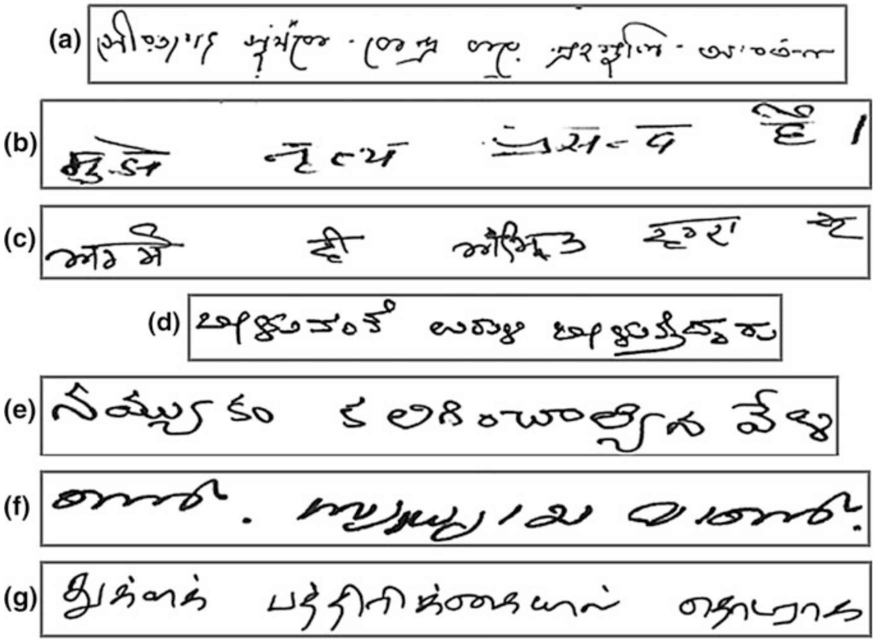

Fig. 2 Sample text-line images written in **a** *Bangla*, **b** *Devanagari*, **c** *Gurumukhi*, **d** *Kannada*, **e** *Telugu*, **f** *Malayalam* and **g** *Tamil* misclassified as *Gujarati*, *Gurumukhi*, *Devanagari*, *Telugu*, *Kannada*, *Tamil* and *Malayalam* scripts respectively

Though Table 2 shows impressive results but some misclassifications have been found during the experimentation. The main reasons are: (a) presence of speckled noise, (b) existence of multi-skewed words present in some text-lines, and (c) occurrence of irregular spaces within text words, punctuation symbols, etc. in the text–line images. The structural resemblance in the character set of some of the Matra based scripts like *Devanagari* and *Gurumukhi* and non-Matra based scripts like *Kannada* and *Telugu* as well as *Malayalam* and *Tamil* cause similarity in the contiguous pixel distribution which in turns misclassifies them among each other. Figure 2 shows some samples of misclassified text-line images.

6 Conclusion and Future Work

We have proposed a robust method for handwritten script identification at text-line level for all the official scripts of India. The main intention of this paper is to facilitate the multilingual handwritten OCR and script based retrieval of offline handwritten documents. A set of 84 features are extracted using the combination of NGTDM and GLRLM. NGTDM aims to extract information about spatial changes in intensity which can be obtained by looking at the difference between the gray tone of each image pixel and the gray tones of its neighbors. On the contrary, GLRLM contains great discriminatory information which in turn preserves much of the texture information in run-length matrices. Experimental results have shown that an accuracy rate of 97.69% is achieved using MLP classifier which is quite acceptable taking the complexities and shape variations of the scripts under consideration. This work is first of its kind presuming the number of official scripts into account. Our future endeavor will be to modify this technique to perform the script identification from handwritten document images containing more number of Indian languages. As the key feature used in this technique is mainly texture based, in future, the technique will be applicable for recognizing non-*Indic* scripts in any multi-script environment. Focus will be also to increase the size of the text-line script database to incorporate larger variations of writing styles belonging to writers from speckled backgrounds which, in turn, would devise our technique as writer independent.

References

1. Singh, P.K.: Script identification from multi-script handwritten documents. M. Tech Thesis, CSE Department, Jadavpur University (2013)
2. Language in India. http://www.languageinindia.com/feb2011/vanishreemastersfinal.pdf. Accessed 05 Feb 2016
3. Singh, P.K., Sarkar, R., Nasipuri, M.: Offline script identification from multilingual indic-script documents: a state-of-the-art. Comput. Sci. Rev. **15–16**, 1–28 (2015)
4. Dhandra, B.V, Nagabhushan, P., Hangarge, M., Hegadi, R.: Script identification based on morphological reconstruction in document images. In: IEEE International Conference of Pattern Recognition, Hong Kong, pp. 950–953 (2006)
5. Padma, M.C., Vijaya, P.A.: Global approach for script identification using wavelet packet based features. Int. J. Signal Process. Image Process. Pattern Recogn. **3**, 29–40 (2010)
6. Padma, M.C., Vijaya, P.A.: Wavelet packet based texture features for automatic script identification. Int. J. Image Process. **4**, 53–65 (2010)
7. Pal, U., Chaudhuri, B.B.: Identification of different script lines from multi-script documents. Image Vis. Comput. **20**, 945–954 (2002)
8. Padma, M.C., Vijaya, P.A.: Identification of Telugu, Devnagari and English scripts using discriminating features. Int. J. Comput. Sci. Inf. Technol. **1** (2009)
9. Padma, M.C., Vijaya, P.A.: Script identification from trilingual documents using profile based features. Int. J. Comput. Sci. Appl. **7**, 16–33 (2010)
10. Joshi, G.D., Garg, S., Sivaswamy, J.: Script identification from Indian documents. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) LNCS **3872**, 255–267 (2006)
11. Jindal, M., Hemrajani, N.: Script identification for printed document images at text-line level using DCT and PCA. IOSR J. Comput. Eng. **12**, 97–102 (2013)
12. Pal, U., Chaudhuri, B.B.: Automatic separation of words in multi lingual multi script indian documents. In: 4th International Conference on Document Analysis and Recognition (ICDAR). pp. 576–579 (1997)
13. Sinha, S., Pal, U., Chaudhuri, B.B.: Word-wise script identification from Indian documents. LNCS **3163,** 310–321 (2004)

14. Hassan, E., Garg, R., Chaudhury, S., Gopal, M.: Script based text identification : a multi-level architecture. In: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pp. 11:1–11:8 (2011)
15. Dhandra, B.V, Mallikarjun, H., Hegadi, R., Malemath, V.S.: Word-wise script identification from bilingual documents based on morphological reconstruction. In: IEEE International Conference on Digital Information Management, pp. 389–394 (2006)
16. Pati, P.B., Ramakrishnan, A.G.: Word level multi-script identification. Pattern Recogn. Lett. **29**, 1218–1229 (2008)
17. Dhanya, D., Ramakrishnan, A.G., Pati, P.B.: Script identification in printed bilingual documents. Sadhana Acad. Proc. Eng. Sci. **27**, 73–82 (2002)
18. Singh, P.K., Dalal, S.K., Sarkar, R., Nasipur, M.: Page-level script identification from multi-script handwritten documents. In: 3rd IEEE International Conference on Computer, Communication, Control and Information Technology (C3IT), pp. 1–6 (2015)
19. Hangarge, M., Dhandra, B.V: Offline handwritten script identification in document images. Int. J. Comput. Appl. **4** (2010)
20. Singh, P.K., Sarkar, R., Nasipuri, M.: Line-level script identification for six handwritten scripts using texture based features. In: 2nd Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, pp. 285–293 (2015)
21. Roy, K., Pal, U.: Word-wise Handwritten Script Separation for Indian postal automation. In: International Workshop on Frontiers in Handwriting Recognition, La Baule, pp. 521–526 (2006)
22. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: Word level script identification from Bangla and Devnagari handwritten texts mixed with Roman scripts. J. Comput. **2**, 103–108 (2010)
23. Singh, P.K., Sarkar, R., Das, N., Basu, S., Nasipuri, M.: Identification of Devnagari and Roman scripts from multi-script Handwritten documents. In: 5th International Conference on Pattern Recognition and Machine Intelligence (PReMI), pp. 509–514 (2013)
24. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice-Hall, India (1992)
25. Amadasun, M., King, R.: Textural features corresponding to textural properties. IEEE Trans. Syst. Man Cybern. **19**, 1264–1274 (1989)
26. Galloway, M.M.: Texture analysis using gray level run lengths. Comput. Graph. Image Process. **4**, 172–179 (1975)
27. Chu, A., Sehgal, C.M., Greenleaf, J.F.: Use of gray value distribution of run lengths for texture analysis. Pattern Recogn. Lett. **11**, 415–420 (1990)
28. Dasarathy, B.R., Holder, E.B.: Image characterizations based on joint gray-level run-length distributions. Pattern Recogn. Lett. **12**, 497–502 (1991)