Identification of *Devnagari* and *Roman* Scripts from Multi-script Handwritten Documents

Pawan Kumar Singh*, Ram Sarkar, Nibaran Das, Subhadip Basu, and MitaNasipuri

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
{raamsarkar,nibaran,bsubhadip,mitanasipuri}@gmail.com,
pawansingh.ju@gmail.com

Abstract. In a multilingual country like India it is a common scenario that a handwritten text document may contain more than one script. This causes practical difficulty in digitizing such a document, because the language type of the text should be pre-determined, before feeding it into a suitable Optical Character Recognition (OCR) system. In this paper, an intelligent feature based technique is reported, which automatically identifies the scripts of handwritten words from a document page, written in *Devnagari* script mixed with *Roman* script. The word-level script identification is performed by applying Multi layer Perceptron (MLP) based classifier with 39 distinctive features. The technique is tested on 100 handwritten document pages containing both *Devnagari* and *Roman* script words and 99.54% of words are identified with their true class.

Keywords: Script identification, Multi-script handwritten pages, Optical Character Recognition, Convex-hull feature, MLP classifier.

1 Introduction

India is a multi-lingual country with 25 official languages derived from 12 different scripts. A document page like railway reservation forms, question papers, language translation books and money-order forms etc. may contain words in more than one script/language. Each script has its own characteristics which is very different from other scripts. It is perhaps impossible to design a single recognizer which can identify a variety of scripts/languages. Thus, it is necessary to identify the language/script of the documents before feeding it to the corresponding OCR system. Script identification aims to extract information presented in digital documents namely articles, newspapers, magazines and e-books.

Besides, being the national language, *Hindi* is the most popular language in Indian sub-continent. *Devnagari* script is used to write *Hindi*, *Nepali*, *Marathi* and *Sindhi* languages. *English*, written in *Roman* script, has proven to be the binding language due to the diversity of languages and scripts in India and other countries in the world. For this reason, in the Indian sub-continent, people frequently use *English* words along with their native language in their writing. So, a script identification system to identify the script of the documents is of pressing need. This has motivated us to

* Corresponding author.

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 509–514, 2013.

design an intelligent script identification technique which can successfully serve the immediate required purpose.

From the literature, it is revealed that in the context of Indian languages, some amount of research work on script/language identification has been reported [1-8]. Despite these research contributions, relatively few works [3, 5-8] are found on word-level script identification from handwritten bilingual script containing *Devnagari* script mixed with *Roman* script. In one of the previous works, R. Sarkar et al. [8] discussed about word-level script identification from *Bangla* and *Devnagari* handwritten script mixed with *Roman* script by using MLP classifier. In this paper, we propose a word-level script identification scheme based on 39 distinct features to identify scripts from the documents containing both *Devnagari* and *Roman* scripts by using MLP classifier.

2 Design of Feature Set

In the present work, 39 distinct features are designed for the identification of *Devnagari* and *Roman* script words using MLP classifier. These feature values, used with suitable normalization, are described as follows:

2.1 Matra/Shirorekha Feature

If the longest horizontal run of black pixels on the rows of a word is computed then such run length for *Devnagari* script will be much higher than that of *Roman* script. This is because characters in a word are generally connected by headline/Shirorekha/Matra in *Devnagari*. This information [14] is used to separate the words written in *Roman* script from *Devnagari* script.

2.2 Segmentation Based Feature

Here, two distinct features are considered *viz.*, number of Matra pixels and number of segmentation-point pixels [14].

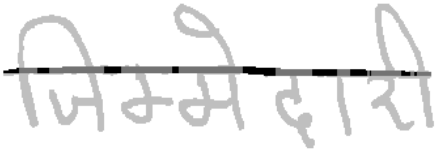

Fig. 1. Illustration of Matra pixels (light grey) and segmentation-point pixels (dark grey) are shown on a sample *Devnagari* word. [9]

Since *Roman* script words do not have Matra region, these pixels counts are significantly lower in case of *Roman* script in comparison to the words written in *Devnagari* script. Matra and segmentation-point pixels (see Fig. 1), are obtained by applying the character segmentation algorithm developed in [9].

2.3 Foreground-Background Transition Feature

The changeover of foreground and background pixels (transition point count) [13] is considered as feature value along 5 row positions in a particular word. For this purpose, the top row and bottom row of the word image are selected as R_1 and R_5 respectively. The row with maximum horizontalness is selected as R_2 as described in an earlier work on character segmentation[13].Then, the foreground-background changeover has been counted on 5 different row positions (see Fig. 2) such as R_2, $R_4=(R_2+R_5)/4$, $R_3=(R_2+R_4)/3$, $R_{12}=(R_1+R_2)/2$ and $R_{13}=(R_{12}+R_2)/2$ [8].

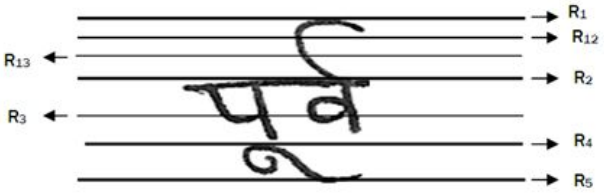

Fig. 2. Selection of specific rows for estimation of foreground-background feature values. [8]

2.4 Convex-Hull Based Feature

The Convex hull of a set of points S in n dimensions is the intersection of all convex sets containing S. For $\boldsymbol{N}$ points $\boldsymbol{P_1, P_2, \ldots, P_N}$, the convex hull C is then given by the expression

$$\mathbf{C} = \left\{ \sum_{\mathbf{j=1}}^{\mathbf{N}} \boldsymbol{\lambda_j p_j} : \boldsymbol{\lambda_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^{N} \lambda_j = 1} \right\} \quad (\mathbf{1})$$

In the current work, we have used Graham scan algorithm [13] for computing the convex hull of binary pattern of each word image. From the Green's theorem [14], it can be shown that the area A of convex hull is given by

$$\mathrm{A} = \frac{1}{2} \sum_{\mathrm{i=1}}^{\mathrm{L}} (\mathrm{x_i\, y_{i+1}} - \mathrm{x_{i+1} y_i}) \quad (2)$$

where, L is the number of order vertices, (x_i, y_i) coordinates of the order vertices forming polygon.

Also, the centroid (C_x, C_y) of the convex hull can be expressed as:

$$C_x = \frac{1}{6A} \sum_{i=1}^{L} (x_i + x_{i+1})\,(x_i y_{i+1} - x_{i+1} y_i) \quad (3)$$

$$C_y = \frac{1}{6A}\sum_{i=1}^{L}(y_i + y_{i+1})\,(x_i y_{i+1} - x_{i+1} y_i) \tag{4}$$

Now, the deficit of convexity can be defined as the set of pixels inside the convex hull of any object pattern which does not belong to the said object. There may be two types of convex deficiencies *viz*., regions totally enclosed by the object, called *lakes* and regions lying between the convex hull perimeter of the object, called *bays* [10]. It has been seen from the convex hull structure of words that the number of *lakes* found in case of *Devnagari* words is more than that of the *Roman* words. On the other hand, in case of *Roman* script words, number of *bays* is more than that in *Devnagari* script words. In the current work, we have extracted 31 basic topological features (like *bays* attributes, *lakes* attributes, etc.) from the convex hull of handwritten word images. For extracting local information, each such word pattern is further divided into four sub-images based on the *centroid* of its convex hull. After that, new convex hulls have been constructed for each such sub-image.

In the present work, d_{cp} are calculated as the column and row wise distances of data pixels from the convex hull boundary taken from the top, bottom, right and left boundaries of any image. Now, the seven topological features are calculated as maximum d_{cp}, total number of rows having $d_{cp} > 0$, average d_{cp}, mean row coordinate having $d_{cp} > 0$, total number of rows having $d_{cp} = 0$, number of visible bays from left to right direction and, convex hull perimeter feature (total number of convex hull pixels having $d_{cp} = 0$ from four sides). From the top, bottom, right and left boundaries of the image (7x4=28) such features are calculated. Finally, the remaining three features *viz*., total numbers of *lakes*, *bays* and convex hull perimeter pixels having $d_{cp} > 0$ are taken under consideration. This makes the total feature count as 39.

3 Experimental Results

For the experimental purpose, a database of 100 handwritten pages comprising of 13941 words are considered. Handwritten document pages have been collected from different types of sources, viz., class notes of students of different age-groups, and the document pages, written by different persons, on request. The document pages written under supervision were collected from various persons with varied textual contents of the books containing both *Hindi* and *English* vocabularies. The documents are digitized by a HP scanner at 300 dpi. The digitized images are in gray tone and we have used a histogram-based thresholding approach to convert them into two-tone images. Both the *Hindi* and *English* word images from the document pages are then manually cropped for the evaluation of the present technique.

We have used 3-fold cross validation scheme for evaluating the script recognition algorithm. The detail of the experimental setup is shown in Table 1. For developing a trained network for each of the MLP based classifiers, several runs of BP learning algorithm with learning rate (η) = 0.7, momentum term (α) = 0.7 and adjustment factor =0.9 are executed for different number of neurons in its hidden layer. The accuracies of 3 different runs of script identification scheme are 99.27%, 99.54% and 99.40% respectively.

Only few words are misclassified during test. The main reasons for misclassification of *Devnagari* words are due to discontinuities in Matra and poor quality of documents due to presence of noise (see Fig. 3(a)). Some of the *Roman* script words are also misclassified (see Fig. 3(b)). The possible reason may be due to existence of Matra like component in the word, extracted feature values are identical to the word written in *Devnagari* script. In addition, the presence of some small component in the upper part for *Devnagari* and *Roman* scripts, sometimes they are misclassified among each other.

Table 1. Detail results of the present script identification technique

Scripts	Set #1		Set #2		Set #3	
	Number of words trained	Number of words tested	Number of words trained	Number of words tested	Number of words trained	Number of words tested
Devnagari	7500	1992	7508	2002	7430	2014
Roman	3410	1031	3412	1026	3520	1011
TOTAL	10910	3023	10920	3028	10950	3025
Success rate for test case (%)	99.27		**99.54**		99.40	

(a) *Devnagari* script (b) *Roman* script

Fig. 3. Sample word images where present technique fails to identify the actual script

4 Conclusion

Script identification has been the forerunner for developing a multi-script OCR system. In this paper, a scheme for word-level handwritten script identification technique is presented. Document pages, applied in the present work, are written in *Devnagari* script mixed with *Roman* script. An intelligent set of 39 features with MLP classifier is applied to identify the scripts of the words. With 3-fold cross validation of 13941 words, the present technique classifies 99.54% words with their true classes. In addition to this, the scheme also works well for single character words. The script identification technique, in future, can also be applied to other scripts like *Bangla*, *Gurmukhi*, etc. mixed with *Roman* script because of the similar script structure. Also, this scheme will be effective and efficient for different writing styles of handwriting which ensures generality of the present technique. In future, we aim to improve the accuracy of the technique by minimizing the script dependency of the features. In addition to this, we are also planning to design a general script identification module for the development of multi-script OCR system.

References

1. Pal, U., Choudhuri, B.B.: Script Line Separation From Indian Multi-Script Documents. In: Proc. of 5th International Conference on Document Analysis and Recognition, pp. 406–409. IEEE Comput. Soc. Press (1999)
2. BasvarajPatil, S., Subba Reddy, N.V.: Character script class identification system using probabilistic neural network for multi-script multi lingual document processing. In: Proc. of National Conference on Document Analysis and Recognition, Karnataka, pp. 1–8 (2001)
3. Pal, U., Choudhuri, B.B.: Automatic Separation of Words in Multi Lingual multi Script Indian Documents. In: Proc. of 4th International Conference on Document Analysis and Recognition, pp. 576–579 (1997)
4. Chanda, S., Pal, U.: English, Devnagari and Urdu Text Identification. In: Proc. of International Conference on Document Analysis and Recognition, pp. 538–545 (2005)
5. Pal, U., Sinha, S., Choudhuri, B.B.: Word-wise script identification from a document containing English, Devanagari and Telugu text. In: Proc. of 2nd National Conference on Document Analysis and Recognition, Karnataka, India, pp. 213–220 (2003)
6. Padma, M.C., Nagabhushan, P.: Identification and separation of text words of Kannada, Hindi and English languages through discriminating features. In: Proc. of 2nd National Conference on Document Analysis and Recognition, Karnataka, pp. 252–260 (2003)
7. Spitz, A.L.: Determination of the script and language content of document images. Proc. of IEEE Tran. on Pattern Analysis and Machine Intelligence 19, 234–245 (1997)
8. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: Word-Level Script Identification from Bangla And Devanagri Handwritten Texts mixed with Roman script. Journal of Computing 2(2) (2010)
9. Basu, S., Sarkar, R., Das, N., Kundu, M., Nasipuri, M., Basu, D.K.: A Fuzzy Technique for Segmentation of Handwritten Bangla Word Images. In: Proc. of International Conference on Computing: Theory and Applications (ICCTA), pp. 427–433 (2007)
10. Das, N., Pramanik, S., Basu, S., Saha, P.K., Sarkar, R., Kundu, M., Nasipuri, M.: Recognition of handwritten Bangla basic characters and digits using convex hull based feature set. In: Proc. of International Conference on Artificial Intelligence and Pattern Recognition, AIPR (2009)
11. Vysniauskaite, L., et al.: A Priori Filtration Of Points For Finding Convex Hull, Tede, vol. XII(4), pp. 341–346 (2006)
12. Lady, E.L.: (February 14, 2000), `http://www.math.hawaii.edu/~lee/calculus/green.pdf`
13. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: A two-stage approach for Segmentation of Handwritten Bangla word Images. In: Proc. of International Conference on Frontiers in Handwritten Recognition (ICFHR), Canada, pp. 403–408 (2008)
14. Sarkar, R., Malakar, S., Das, N., Basu, S., Kundu, M., Nasipuri, M.: Word Extraction and Character Segmentation from Text Lines of Unconstrained Handwritten Bangla Document Images. Journal of Intelligent Systems 20(3), 227–260 (2011)