Offline Handwritten Script Identification in Document Images

Mallikarjun Hangarge
Department of Computer Science
Karnatak Arts, Science and
Commerce College, Bidar,
Karnataka, India

B.V.Dhandra
P.G. Department of Studies and
Research in Computer Science,
Gulbarga University, Gulbarga
Karnataka, India

ABSTRACT
Automatic handwritten script identification from document images facilitates many important applications such as sorting, transcription of multilingual documents and indexing of large collection of such images, or as a precursor to optical character recognition (OCR). In this paper, we investigate a texture as a tool for determining the script of handwritten document image, based on the observation that text has a distinct visual texture. Further, K nearest neighbour algorithm is used to classify 300 text blocks as well as 400 text lines into one of the three major Indian scripts: English, Devnagari and Urdu, based on 13 spatial spread features extracted using morphological filters. The proposed algorithm attains average classification accuracy as high as 99.2% for bi-script and 88.6% for tri-script separation at text line and text block level respectively with five fold cross validation test.

General Terms
Pattern Recognition, Document Image Analysis

Keywords
Script Identification, offline handwritten documents, Optical character reader, cross validation

1. INTRODUCTION
A very important area in the field of document analysis is that of optical character recognition (OCR), which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form. To date, many algorithms have been presented in the literature to perform this task for a specific language; however, such OCR will not work for multilingual documents. Therefore, to make a successful multilingual OCR, script identification is very essential before running an individual OCR system. In this direction, most of the published work on automatic script identification of Indian scripts, deals with printed documents and very few articles were found for handwritten script identification. Most of the published work has identified a number of approaches for determining the script/ language of printed and handwritten documents and they could be typically classified into four categories: (a) the methods based on the analysis of connected components, (b) the methods based on analysis of characters, words and text lines, (c) the methods based on text blocks, (d) the methods based on analysis of hybrid information of connected components, text lines etc. Spitz [15] proposed a method for distinguishing between Asian and European languages by examining the upward concavities of connected components. Tan e*t al.* [9] proposed a method based on texture analysis for automatic script and language identification from document images using multiple channel (Gabor) filters and Gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Koreans, Malayalam, Persian and Russian. Hochberg, *et al.* [5, 6] described a method of automatic script identification from document images using cluster-based templates for printed documents and he also proposed an algorithm for script and language identification from handwritten document images using statistical features based on connected component analysis. Tan [16] developed rotation invariant features extraction method for automatic script identification for six languages. Wood *et al.* [17] described projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters. Chaudhuri *et al.* [1] discussed an OCR system to read two Indian languages scripts Bangla and Devnagari (Hindi). Chaudhuri *et al.* [2] described a complete printed Bangla OCR. Pal *et al.* [10] proposed an automatic technique of separating the text lines from 12 Indian scripts. Gaurav *et al.* [3] proposed a method for identification of Indian languages by combining Gabor filter based techniques and direction distance histogram classifier for Hindi, English, Malayalam, Bengali, Telugu and Urdu. Dhanya *et al.* [4] proposed an algorithm for automatic script identification by separating the English and Tamil words present in a bilingual document using spatial spread features and Gabor filters. Pal *et al.* [11] proposed an algorithm for word-wise script identification from document containing English, Devnagari, and Telugu text, based on conventional and water reservoir features. Padma *et al.* [8] described a method of identification and separation of text words of Kannada, Hindi and English languages using discriminating features. Peeta Basa pati *et al.* [13] discussed about word-wise script identification of three bilingual documents of Hindi, Tamil and Odiya using Gabor filters. Sanjeev *et al.* [7] proposed Kannada and English word separation from bilingual document using Gabor features and Radial basis function neural network. Basavaraj *et al.* [12] proposed a neural network based system for script identification of Kannada, Hindi and English. Nagabhushan *et al.* [14] discussed an intelligent pin code script identification methodology based on texture analysis using modified invariant moments. K. Roy *et al*. [19] proposed a system for word-wise handwritten script identification for Indian Postal automation. Recently, Lijun Zhou *et al*. [20] proposed an automatic handwritten script identification of destination address blocks of envelop images.

In our previous work [22], we have developed an algorithm for handwritten script identification of text words and numerals of English, Kannada and Devnagari scripts, based on global and local features. In this paper, we extended our previous idea to identify the script of handwritten text blocks of Kannada, English and Urdu. A number of experimentations were carried out on these scripts with global and local features separately and in combination. The results indicate that global features were efficient when the image size is large (block level), whereas glocal (local + global) features are efficient when the image size is smaller (word or character level). Therefore, this paper presents a scheme for handwritten text blocks script identification based on 13 global spatial features.

In Section 2, the brief overview of data collection and pre-processing is presented. In Section 3, Segmentation and feature extraction method is discussed. The experimental details and results obtained are presented in Section 4. Conclusion is given in Section 5.

2. DATA COLLECTION

A sample of 150 writers is chosen from schools, colleges and professionals for collecting the handwritten documents. The writers are not imposed by any constraint like type of pen and style of writing etc., and the purpose of data collection is also not disclosed. Writers are provided with the unrolled papers and are asked to write 10 lines of text matter in Devnagari, English and Urdu scripts. A total of 300 handwritten document images are created from 150 handwritten document pages.

The collected documents are scanned using HP Scanner at 300 DPI, which usually yields a low noise and good quality document image. The digitized images are in gray tone and we have used Otsu's global thresholding approach to convert them into two-tone images. Otsu's method chooses the threshold to minimize the interclass variance of the thresholded black and white pixels. The two-tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The small objects (less than are equal to 40 pixels) like, single or double quotation marks, hyphens and periods etc. are removed using morphological opening. The next step in pre-processing is skew detection and correction and is performed using the algorithm [21].

3. FEATURE EXTRACTION

3.1 Segmentation

Initially, 128 x128 text blocks are segmented manually from the document images of Kannada, Devnagari and Urdu and created 300 text blocks. Further, 10 pages of Devnagari and Urdu handwritten documents are used for line-wise segmentation using horizontal projection profile and obtained 200 lines. However, the touched line segmentation is not attempted here. As the standard database is available for English handwritten text lines; therefore, we have used 200 text lines from IAM database.

***Devnagari*:** Most of the characters of Devnagari script have a horizontal line at the upper part. In Devnagari, this line is called sirorekha. However, we shall call them as headlines. When two or more Devnagari characters sit side by side to form a word, the sirorekha or headline touch one another and generates a big headline [10] in case of printed documents, whereas in handwritten documents, these lines are usually drawn after the word is written.

Roman (English): The important property of the Roman (English) script is the existence of the vertical strokes in its characters and has less number of horizontal strokes as compared to Devnagari and Urdu scripts. The right and left diagonal strokes are also plays an important role in distinguishing Roman from Devnagari and Urdu scripts.

***Urdu*:** The Urdu characters have strong base line as well as right diagonal strokes. Urdu script has less number of holes as compared to other two scripts.

These directional visual discriminating features are extracted from the image or pattern for discrimination of proposed scripts. In the following, we describe the features and their method of computation. To extract the strokes in vertical, horizontal, right and left diagonal directions, we have performed the opening operation on the input binary image with the line-structuring element. The length of the structuring element is experimentally fixed for text block as veritical-10, horizontal-7, left and right diagonal -5 each. For line wise feature extraction the structuring element length is thresholded to 70% of the average height of the connected components of an image (empirically fixed).

Stroke density: The stroke length is defined as the number of pixels in a stroke as the measure of its length [18], for the strokes in vertical, horizontal, right and left diagonal directions of the image. Further, the stroke density is defined as the total length of all the strokes divided by the size of the image. Throughout the discussion N is referred as number of on pixels. The values of 13 features extracted here, are real numbers. The average feature vector of 25 sample images is shown by a line chart in Fig 1, to visualize the strength of the feature set for discriminating the proposed scripts.

1 Vertical Stroke Density (VSD):

$$VSD = \frac{\sum_{i}^{N} Onpixel(Pattern\_verticle_i)}{Size(Pattern\_verticle)} \quad (1)$$

2 Horizontal Stroke Density (HSD):

$$HSD = \frac{\sum_{i}^{N} Onpixel(Pattern\_horizontal_i)}{Size(Pattern\_horizontal)} \quad (2)$$

3 Right Diagonal Stroke Density (RDSD):

$$RDSD = \frac{\sum_{i}^{N} Onpixel(Pattern\_right\_digonal)}{Size(Pattern\_right\_digonal)} \quad (3)$$

4 Left Diagonal Stroke Density (HSD):

$$LDSD = \frac{\sum_{i}^{N} Onpixel(Pattern\_left\_digonal_i)}{Size(Pattern\_left\_digonal)} \quad (4)$$

Pixel Density of an image after fill holes: This is the ratio between the number of on pixels left after performing fill hole operation on input pattern, to its size. For fill holes, we choose the marker image (erode image), f_m, to be 0 everywhere except on the image border, where it is set to 1-f. Here f is the image of a connected component.

$$f_m(x,y) = \begin{cases} 1 - f(x,y), & \text{if } (x,y) \text{ is on the border of f} \\ 0, & \text{otherwise} \end{cases}$$

Then $g = [R_f^c (f_m)]^c$ has the effect of the filling the holes in f. Where, R_f^c is the reconstructed image of f.

5. **Pixel Density** of the pattern after fill holes is defined as

$$PHD(pattern) = \frac{\sum_{i}^{N} Onpixel(g_i)}{Size(g)} \qquad (5)$$

The remaining (sixth to thirteenth) features are extracted by top-hat and bottom-hat morphological filtering (transformations) in four directions. The features are computed in similar way as discussed in equation (1)-(4). The "top-hat" transformation, due to F.Meyer, aims to extract the objects that have not been eliminated by the opening. *It can be defined as the residue between the identity and an opening.* This transformation is preferred here to decompose an input image in four directions at three levels to extract fine textural primitives for discrimination of scripts.

The sample feature vector of English, Devnagari and Urdu is given below.

Devnagari = [0.0021 0.0443 0.0554 0.0114 0.0350 0.0167 0.0984 0.0068 0.0396 0.0065 0.0096 0.0368 0.0039]

English = [0.0047 0.0222 0.0175 0.0011 0.0258 0.0140 0.0354 0.0092 0.0178 0.0035 0.0085 0.0184 0.0040]

Urdu = [0.0003 0.0274 0.0156 0.0014 0.0264 0.0078 0.0287 0.0097 0.0180 0.0008 0.0022 0.0256 0.0013]

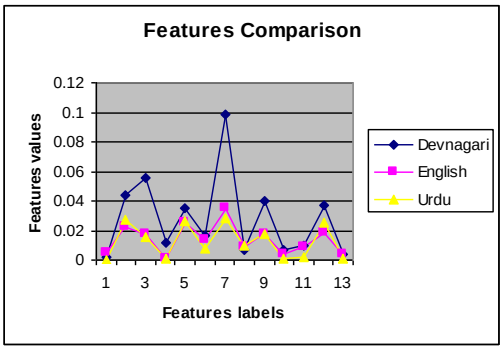

Figure 1. Average of 25 features of (a) Devnagari (b) English and (c) Urdu scripts is illustrated by line chart

(a) (b) (c)

Figure 2. Shows horizontal top hat transformation (a) input Devnagari text block (b) horizontal opening of (a), (c) horizontal top hat transformation of (a).

4. SCRIPT CLASSIFICATION

Experimentations are carried out with KNN classifier by varying the number of neighbours (*K*= 3, 5, 7, 9, 11, 13, 15) and the performance of the algorithm is found optimal when *K* = 5 for text blocks and k=3 for text lines respectively. To evaluate the performance of the classifier the data set containing 300 text blocks and 400 text lines are randomly divided into five groups and a 5-fold cross validation was done for 100 iterations to get optimum results.

For experimentation, 300 handwritten document pages obtained from 150 writers are used with an assumption that the document pages contain only text lines. These document pages are scanned using a flatbed HP scanner at a resolution of 300 dpi. A sample image of size 128x128 pixels is selected manually from each document image and created 300 text block images. Out of these 300 images Devnagari, English, and Urdu are 100 each. The accuracy of the classification achieved for script identification at text block level as well as at text lines is presented in Tables 1 to 4. From the experimentation, we noticed that the text blocks of Devnagari containing connected components of weak headlines and without headlines are miss classified as Roman script (see Fig. 3). The English script miss classified as Devnagari due to the text blocks containing the connected components of strong horizontal stroke at the top of the character. Urdu script miss classified as Devnagari due to long base lines used by some writers at the bottom of the characters. The algorithms proposed by Hochberg [5], for identification of six scripts and Lijun Zhou [20], for two scripts have shown an accuracy of 88% and 95% respectively. The algorithm proposed in this paper achieves the maximum average accuracy of 97.50% for the combination of Roman and Urdu scripts. The minimum average recognition accuracy is 89.00% for Roman and Devnagari scripts. However, overall accuracy of the proposed algorithm is as high as 88.6667% and 97.5% for tri-script and bi-script classification. Further, we observed that when the size of the image increases the results of recognition also increases and hence the text line wise script identification results are high as compared to text block script identification results. The text line level bi-script and tri-script identification results are shown in table 3 and 4. The performance comparison of the proposed algorithm with other methods is presented in Table 5. The proposed algorithm is implemented in MATLAB 6.1. The average time taken to recognize the script of a text block image is 0.2187 seconds and for text lines 0.8734 seconds on a Pentium-IV with 128 MB RAM based machine running at 1.80 GHz.

Table 1 Text block level bi-script identification results

Scripts	For K=3
English Vs. Devnagari	90.00%
	88.00%
Average	**89.00%**
English Vs Urdu	96.00%
	99.00%
Average	**97.50%**
Devnagari Vs. Urdu	95.00%
	98.00%
Average	**96.5%**

Table 2. Text block level tri-script identification results

Scripts	For K=5
English	86.00%
Devnagari	83.00%
Urdu	97.00%
Average	**88.67%**

Table 3 Text line level bi-script identification results

Scripts	For K=3
English Vs. Devnagari	98.37%
	95.00%
Average	**96.68%**
English Vs Urdu	99.45%
	98.94%
Average	**99.20%**
Devnagari Vs. Urdu	99.00%
	98.94%
Average	**98.97%**

Table 4. Text line level tri-script Identification Results

Scripts	For K=3
English	97.83%
Devnagari	93.00%
Urdu	95.78%
Average	**95.54%**

Table 5. The comparative study of text block level script identification

Proposed Algorithm	Scripts	Acc.	TC	VT
Hochberg	A, Ch, Cy, D, R and J	88.00%	N.R	RP
Lijun Zhou	R and B	95.00%	N.R	NR
Proposed	R and D	89.00%	0.2086 sen	RP
	R and U	97.50%		
	D and U	95.50%		
	R, D and U	88.67%	0.2187 Sen.	RP

A- Arabic, Ch-Chinese, Cy-Cyrillic, D-Devnagari, R-Roman, J-Japanese, B-Bangla, U-Urdu, Acc-Accuracy, T.C-Time complexity, V.T- validation test, N.R-not reported, RP-reported, sen-seconds.

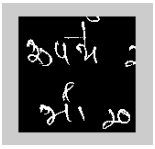

Figure 3. A sample miss classified Devnagari text block as English script

5. CONCLUSION

In this paper, we have proposed a very simple method for handwritten script identification of three major India scripts. The aim of the paper is to facilitate the multilingual handwritten OCR, script based retrieval of offline-handwritten documents and postal envelop sorting based on the scripts of the address blocks. Proposed algorithm decomposes the image in four directions at three levels by extracting fine texture primitives for discrimination. During the extraction of features, the connected components of size less than are equal to 40 pixels are removed from the image prior to features computation. Thus, the approach is robust with respect to noise. It is clear that this algorithm is insensitive to writing style, ink, and size, inter-line; inter-word spacing and slope of text lines as well as characters slant.

ACKNOWLEDGMENTS

This work is carried out under the UGC sponsored minor research project (ref: MRP(S):661/09-10/KAGU013/UGC-SWRO, dated, 30/11/2009)

The authors are grateful to Dr. P. Nagabushan, and Dr. D. S. Guru, University of Mysore, for their comments and suggestions.

6. REFERENCES

[1] B.B.Chaudhuri and U.Pal," An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)", *Proc. of 4th ICDAR*, Uhn. 18-20 August 1997.

[2] B.B.Chaudhuri and U.Pal, "A complete printed Bangla OCR," *Pattern Recognition*, vol. 31, pp. 531-549, 1998.

[3] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, R.B.Shet," Identification of scripts of Indian languages by Combining trainable classifiers," *Proc. of ICVGIP*, Dec-20-22, Bangalore, India, 2000.

[4] D.Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," *Sadhana*, vol. 27, part-1, pp. 73-82, 2002.

[5] J. Hochberg, P. Kelly, T Thomas and L Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.19, pp.176-181, 1997.

[6] Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, "Script and language identification for hand-written document images," *IJDAR*, vol.2, pp. 45-52, 1999.

[7] R. Sanjeev Kunte1 and R.D. Sudhaker Samuel, **"**On Separation of Kannada and English Words from a Bilingual Document Employing Gabor Features and Radial Basis Function Neural Network", *Proc. of ICCR*, pp. 640-644, 2005.

[8] M.C.Padma and P. Nagabhushan," Script Identification and separation of text words of Kannada Hindi and English languages through discriminating features," *Proc. of NCDAR-2003*, pp. 252-260. 2003.

[9] G.S.Peake and Tan, "Script and language identification from document images," *Proc. of Eighth British Mach. Vision Conf.*, vol.2, pp. 230-233, Sept-1997.

[10] U.Pal and B.B.Chaudhuri, "Script line separation from Indian Multi-script documents," *5th ICDAR*, pp.406-409, 1999.

[11] U.Pal, S.Sinha and B.B Chaudhuri, "Word-wise Script identification from a document containing English, Devnagari and Telgu Text," *Proc. of NCDAR*, PP 213-220, 2003.

[12] S. Basavaraj Patil and N.V.Subbareddy, "Neural network based system for script identification in Indian documents," *Sadhana*, vol. 27, part-1, pp. 83-97, 2002.

[13] Peeta Basa pati, S. Sabari Raju, Nishikanta Pati and A.G. Ramakrishnan, "Gabor filters for document analysis in Indian Bilingual Documents," *Proc. of ICISIP*, pp. 123-126, 2004.

[14] P**.** Nagabhushan**,** S.A**.** Angadi and B.S. Anami," An Intelligent Pin code Script Identification Methodology Based on Texture Analysis using Modified Invariant Moments," *Proc. of ICCR*, pp. 615-623, 2005.

[15] A.L.Spitz, "Determination of the script and language content of document images," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp.234-245, 1997.

[16] T.N.Tan, "Rotation invariant texture features and their use in automatic script identification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp.751-756, 1998.

[17] S. Wood. X. Yao. K.Krishnamurthi and L.Dang "Language identification for printed text independent of segmentation," *Proc. of Int'l. Conf. on Image Processing*, pp. 428-431, 1995.

[18] Anoop M, Namboodri, Anil K. Jain," Online handwritten script identification", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26,no.1, pp.124-130, 2004.

[19] K. Roy, A. Banerjee and U. Pal, "A System for Word-wise Handwritten Script Identification for Indian Postal Automation", In Proc. *IEEE India Annual Conference* 2004, (INDICON-04), pp. 266-271, 2004.

[20] Lijun Zhou, Yue Lu, Chew Lim Tan, "Bangla/English Script Identification based on Analysis of Connected component Profiles", In Proc. 7th IAPR workshop on Document Analysis System, New land, pp. 234-254,13-15, Feb-2006,

[21] B.V.Dhandra, V.S.Malemath, Mallikarjun Hangarge, Ravindra Hegadi, "Skew detection in Binary image documents based on Image Dilation and Region labeling Approach", In Proceedings of ICPR 2006, V. No. II-3, pp. 954-957

[22] B.V. Dhandra, Mallikarjun Hangarge, "Global and Local Features Based Handwritten Text Words and Numerals Script Identification," iccima, vol. 2, pp.471-475, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), 2007