Script Identification In Trilingual Indian Documents

R. R. Aparna *skandhur@gmail.com*
Research Department of computer Science
S.D.N.B.Vaishnav college
AFF. Madras University
Chennai, 600043, India

R. Radha *radhasundar1993@gmail.com*
Research Department of computer Science
S.D.N.B.Vaishnav college
AFF. Madras University
Chennai, 600043, India

Abstract

This paper presents a research work in identification of script from trilingual Indian documents. This paper proposes a classification algorithm based on structural and contour features. The proposed system identifies the script of languages like English, Tamil and Hindi. 300 word images of the above mentioned three scripts were tested and 98.6% accuracy was obtained. Performance comparison with various existing methods is discussed.

Keywords: Script, Dilation, Boundary, Centroid, Zone, Blob, Bounding Box.

1. INTRODUCTION

India being a multi-lingual country, most of the documents contains two (bilingual) or three language (trilingual) scripts. English and Hindi are the most prevalent language being used in India along with the state language. Important documents like, government documents and question papers of every state contains two or three languages. Scrip is referred as the graphics part of the writing system. The writing style and the characters are combined in the script class. A script can be used by only one language or many languages. Indian and South East Asian languages are based on Brahmic scripts. North Indian languages are based on Devanagri, Bengali, Manipuri, Gurumukhi, Gujarati and Oriya scripts. South Indian languages are based on Tamil, Telugu, Kannada and Malayalam scripts. Automatic identification of scripts is very essential before feeding into multi-script optical character recognition (OCR) system. Many methods exist in the literature for script identification. In section 2 existing methods in this field are discussed. Properties of English, Hindi and Tamil are discussed in section 3. In section 4 the proposed methodology is explained. Results and performance comparison with various existing methods are detailed in section 5. Section 6 deals with the conclusion and future work.

2. LITERATURE SURVEY

Various methods exist in the literature for multi-script identification. The script identification is performed in two ways as local approach and global approach. Identification of script at text line level or word level or character level is defined as local approach. In global approach the entire text block is considered [1]. The following methods identify multi-scripts at text line level. [2], [18] used shape features, statistical features and Water Reservoir based features and classified using a rule based classifier.[3] have used horizontal projection profile (HPP) and derived a rule based classifier by calculating the threshold using the successive maxima. Structural features, K-Nearest Neighbour (KNN) and support vector Machine (SVM) was used by [4] for classification. Profile based features were used by [5] and classified using KNN. Structural, topological, contour and Water Reservoir based features were used by [6] and classified using rule based classifier.

Character level script identification technique using Gabor and Gradient features and SVM for classification [7]. The following methods were performed at word level. Eight features were obtained by [8] based on top and bottom max row, top horizontal line, tick component, top and bottom holes, vertical lines and bottom component and classified using KNN. [9] performed Devanagri script identification using the presence of headlines as [2] and Kannada script identification using morphological operation. Their method was restricted to font style and size. [10] have used 400 and 64 dimensional features and SVM was used for classification. [11] has used Neural networks (NN) for classification and features based on histogram mean distance, pixel value and vertical zero crossing. [12] used Zone based Gabor features and SVM classifier was used.

3. PROPERTIES OF ENGLISH, HINDI AND TAMIL SCRIPTS

English language uses Latin or Roman script. English follows the Alphabetic system. Hindi language uses Devanagri script. The characters of the Hindi words are connected using a headline called “Shirorekha”. Printed Tamil and English words contain isolated characters. The structure of the three scripts is divided into three zones (Fig.1). The upper zone (UZ), middle zone (MZ) and lower Zone (LZ). The character set of English constitutes 52 characters comprising 26 uppercase and 26 lowercase characters. The uppercase English characters stay inside the middle zone. The lowercase English characters like b, d, f, h, k, l, t containing ascenders occupies UZ and those containing desecenders like g, j, p, q, y occupies the LZ. But the presence of pixels in the LZ and UZ by lower case English alphabets is of symmetrical pattern (Fig1.a). Tamil character contains 12 vowels and 18 consonants. 216 compound characters were formed by combining vowels and consonants. Almost all the character of Tamil occupies UZ and LZ. Tamil character that occupies the UZ and LZ has varied pixel density and patterns (Fig1.b). Hindi character contains 34 consonants and 10 vowels and form combined characters. Since Hindi characters are connected using headline, the probability of pixel density in the LZ and UZ is very less (Fig1.c).

FIGURE 1: a) English script b) Tamil script c) Hindi script.

4. PROPOSED METHODOLOGY

The input image is converted to greyscale image and binarized using Ostu method [17], (Fig.2). The proposed algorithm works by dividing the multi-script words into three zones and the pixel densities are calculated in the UZ above Headline (HL) and LZ below base line (BL) for discriminating multi-script (fig.1). The proposed algorithm is based on the contour features and classifies using a rule based classifier. First it checks whether the script belongs to Hindi by using the pixel density of UZ and LZ. Classification of English and Tamil script is performed using the contour features of the boundary pattern such as upper boundary (UB) and lower boundary (LB) pixel density calculation.

FIGURE 2: Binarized images: a) Tamil b) English c) Hindi.

4.1 Classification Phase - I

The given word image is classified in two phases. In phase - I the image is classified to check the presence of Hindi word. In phase - II, it is classified as English or Tamil.

Step 1: The heights (h_i) and widths (w_i) of all the blobs are calculated for the image (Fig2). Where i=1 to N, N is the number of blobs.

Step 2: Calculate the average width $avg(\mathrm{w}_i)$ and $avg(\mathrm{h}_i)$ average heights of all blobs.

Step 3: Divide every h_i by $avg(\mathrm{h}_i)$ and store in an array.

$$v_1 = \frac{h_i}{avg(h_i)} \tag{1}$$

Step 4: Choose the $\min(v_1)$ as it is closer to the average $avg(\mathrm{h}_i)$.

Step 5: Extract the $\min_x$, $\min_y$ and ht for the chosen h_i (blob).

Step 6: Calculate the height of the MZ using the following formula. (Fig.3)

$$mh = round\left(min_y + ht\right) \tag{2}$$

FIGURE 3: Middle Zone (MZ) Height.

Step 7: Count the pixels row wise for all rows of the entire image and store it in an array (s1).

Step 8: Count the pixels above HL and below BL using

$$\left(s2 = \sum_{i=1}^{\min_y} s1, \mathrm{s}\,3 = \sum_{i=mh}^{\max_y} s1 \right) \tag{3}$$

Step 9: Count the no of blobs

The word is classified as Hindi if pixels in BL and HL is zero and the total number of blobs is equal to one as the Hindi word is connected by a headline otherwise it is classified as English or Tamil in the next phase.

4.2 Classification Phase – II

Step 1: Extract the boundary features using Morphological operations
Step 2: Dilate the image.(Fig.4a)

Step 3: Calculate the centroid value c_x and c_y .

Step 4: Extract the boundary co-ordinates and store in the array, x_{ub} , y_{ub} and x_{lb} , y_{lb} . (Fig. 4 & 5)

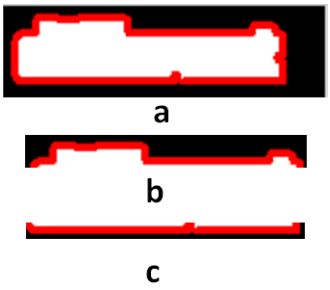

FIGURE 4: a) Obtained Boundary (English) b) UB c) LB.

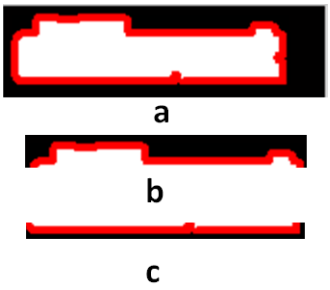

FIGURE 5: a) Obtained Boundary (Tamil) b) UB c) LB.

Step 5: Extract the upper boundary pixel co-ordinates (ub_y) and lower boundary (lb_y) pixel co-ordinates using the following formula,

$$ub_y = \sum_{i=1}^{\max(y_{ub})} y_{ub}, \mathrm{y}_{ub}(i) < c_y \quad (4)$$

$$lb_y = \sum_{i=1}^{\max(y_{lb})} y_{lb}, \mathrm{y}_{lb}(i) > c_y \quad (5)$$

Step 6: Calculate the total count of pixels in the upper boundary and lower boundary. If the total pixels is greater than the threshold then it is classified as Tamil else English. The basic idea is that, the upper boundary of Tamil words (Fig.6a) contains unsymmetrical patterns (peaks and valleys) by accumulating more pixels. But in English word (Fig.6b) the boundary shape pattern is more symmetrical without jerks. Hence the pixels between min_y and $\max_y$ will be zero or less for English. Always the characters of Tamil font will be little bigger in size compared to the English character of same font size.

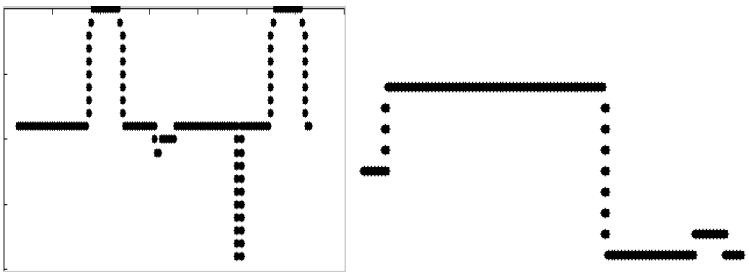

FIGURE 6: Scatter plot for UP of Tamil and b) Scatter plot for UP of English.

5 RESULTS AND DISCUSSION

Samples were collected from the trilingual documents. 300 word samples of Hindi, English and Tamil were tested using the proposed algorithm and 98.6% accuracy was obtained. In cases where the Tamil words were very small and having symmetrical pattern was misclassified as English. The error rate is very low as always the number of pixels will be greater for Tamil compared to English.

Script	Hindi	English	Tamil
Hindi	**100**	0	0
English	0	**98**	2
Tamil	0	2	**98**

TABLE 1: Accuracy of Script Identification.

There is lack of any comparative analysis of the results with most of the reported works in script recognition. Experimental results of every proposed method have not been compared with other benchmark works in the field [1]. As there was no standard evaluation measures available in the literature. The results were evaluated as using the following statistical measures (Table. 2 & 3). The proposed algorithm is evaluated using the statistical measures Recall, Precision, Accuracy and Fscore based on the following metrics. (H-Hindi, E-English and T-Tamil)

True Positive (TP): Correctly identified: (H, E, T identified as H, E, T).
False Positive (FP): Incorrectly identified: (E, T incorrectly identified as T, E).
True Negative (TN): Correctly rejected.
False Negative (FN): Incorrectly rejected.

5.1 Statistical Measure

Recall: This gives the positive cases obtained. It is same as the sensitivity to identify positive results.

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

Precision: This measure gives the percentage of positive predictions.

$$\mathrm{Precision} = \frac{TP}{FP + TP}$$

Accuracy: This measure gives the predictions that were correct.

$$\mathrm{Accuracy} = \frac{TN + TP}{TP + FN}$$

Fscore: This measure is used as a single measure of performance test by combining all the above results.

$$\mathrm{Fscore} = 2\left(\frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call}\right)$$

	Predicted -ve	Predicted +ve
-ve case	TN: 0	FP: 4
+ve case	FN: 0	TP: 296

TABLE 2: Statistical Results.

Measure	Accuracy
Accuracy	100%
Recall	100%
Precision	98.6%
Fscore	99%

TABLE 3: Accuracy of Classification.

5.2 Performance Comparison with the Existing Methods

The proposed algorithm is compared with many of the existing methods in the literature for script identification of Hindi (H), English (E), Kannada (K), Devanagri (D), Oriya (O),Telugu (Te), Gurumukhi (G), and Tamil (T) using local approach.

Author & year	Script	Local Approach	Technique Used	Accuracy	Remarks
[14]-2002	E, H & K	Word-wise	NN	98%	Fixed font size and computationally complex.
[12]-2003	12 scripts	Text line	Local features and water reservoir based feature	98%	Many features were extracted from individual characters
[13]-2010	K, H & E	Text line	Top and bottom profile, KNN classifier	99.75%	Dependent on font size, time consuming
[16]-2002	T & E	Word-wise	Gabor features, SVM classifier	96%	Computationally complex
[5]-2005	D, E & U	Word-wise	Profile based features, Water reservoir method	97.51%	Character level features were extracted.
[9]-2009	E, D & B	Word-wise	64 & 400 dimensional features and SVM as classifier	98.51%	Error rate increases when number of character is less than three
[3]-2013	H, B, Te & K	Text line	Projection profile and rule based classifier	97.83%	Fixed font size was used
[6]-2013	G & E	Character level	Gabor, Gradient features and SVM classifier	99.45%	Complex and Time consuming
[10]-2013	D	Word-wise	Mean distance, pixel value and zero crossing, NN	90%	Computationally complex and execution time is very high
[7]-2011	H &E	Word-wise	Many features,	95%	

			KNN		
[8]-2011	K, E & H	Word-wise	Features based on Morphology and KNN	95.54%	Suitable only for fixed font size and font style
[15]-2002	D, T & O	Word-wise	Gabor filters	97.33	Computationally complex
Proposed method-2014	**H, E &T**	**Word-wise**	**Zonal and boundary features using rule based classifier**	**98.6%**	**Simple, fast, efficient and accurate**

TABLE 4: Performance Comparison.

6. CONCLUSION AND FUTUREWORK

The existing methods are computationally complex and have used more features based on structural and pixel density feature. But the proposed method is very simple, efficient, fast and accurate for discriminating the scripts in bilingual or trilingual documents. Majority of the Indian bilingual or trilingual documents contains 2 or 3 scripts. The documents contain Hindi, English and the state language. Hence the proposed algorithm can be used for script identification of trilingual documents containing Tamil, Hindi and English or Bilingual documents containing any two combinations of these three scripts. The future work will be concentrated towards script identification of Kannada, Telugu, and Malayalam in trilingual or bilingual documents with the combination of Hindi and English as the other two languages.

7. REFERENCES

[1] D.Ghosh, T.Dube and A.P.Shivaprasad. “Script Recognition: A review,” IEEE Transactions on pattern analysis and machine Intelligence, vol. XX, 2009.

[2] U.Pal and B.B.chaudhuri. “Identification of different script lines from multi-script documents,” Image and Vision Computing, vol. 20, pp. 945-954, 2002.

[3] O.Prakash, V.Shrivastava and A.Kumar. “An efficient approach for script identification,” International journal of computer Trends and Technology (IJCTT), vol. 4, 2013.

[4] Rajesh Gopakumar, N.V.SubbaReddy, Krishnamoorthi Makkithaya and U.Dinesh Acharya. “Script Identification from multilingual Indian documents using structural features,” Journal of computing, vol. 2, 2010.

[5] S.Chanda and U.Pal. “English , Devnagari and Urudu text Identification,” in Proc. International conference on Document Analysis and Recognition, pp. 538-545, 2005.

[6] R.Rani, R.Dhir and G. Lehal. “Script Identification of pre-segmented multifont characters and digits,” in Proc. ICDAR, 2013, pp.1150-1154.

[7] M.Swamy Das, D.Sndhya Rani, C.R.K.Reddy and A.Govardhan. “Script identification from multilingual Telugu, Hindi and English Text documents,” International Journal of Wisdom based computing, vol.1, 2011.

[8] B.V.Dhandra and M.Hangarge. “Morphological reconstruction for word level script identification,” International journal of computer science and security, vol.1, 2011.

[9] S.Chanda, S.Pal, K.Franke, U.Pal. “Two-stage approach for word wise script Identification,” in Proc. ICDAR, 2009.

[10] D.Yadav, S.Sanchez-cuadrado and J.Morato. “Optical acharacter recognition for Hindi language using a NN approach,” Journal of information processing systems, vol. 9, 2013.

[11] R.Rani, R.Dhir and G.Lehal. “Modified Gabor feature extraction method for word levelscript identification –Expermentation with Gurumukhi and English scripts,” International journal of signal processing, image processing and pattern recognition, vol. 6, pp. 25-38, 2013.

[12] U.pal and B.B.Chaudhri. “Indian script Character recognition: A survey,” Pattern recognition, vol. 37, pp.1887-1889, 2004.

[13] M.C.Padma and P.A Vijaya. “Script identification from trilingual documents using profile based feature,” International journal of computer science and applications, vol.7, pp.16-33, 2010.

[14] S.B. Patil and N.V.Subbareddy. “Neural network based system for script identification in Indian documents,” Sadhana, vol. 27, pp. 83-97, 2002.

[15] P.B.Pati, S.S.Raju, N.Pati and A.G.Ramakrishnan. “Gabor filters for Document analysis in Indian Bilingual Documents,” in Proc. IEEE ICISIP, 2004.

[16] D.Dhanya, A.G.Ramakrishnan and P.B.Pati, “Script identification in printed bilingual documents,” Sadhana, vol. 27, pp. 73-82, 2002.

[17] R.C.Gonzalez, R.E.Woods and S.L.Eddins. “Digital Image Processing using Matlab,” NewDelhi, Tata Mcgraw-Hill, 2011.

[18] U.Pal and B.B.Chaudhri. “Script line separation from Indian Multi script documents,” in Proc. ICDAR, pp. 406-409,1999.