Structural Feature Based Approach for Script Identification from Printed Indian Document

Sk Md Obaidullah
Dept. of Computer Sc. & Engg.
Aliah University, Kolkata, W.B
sk.obaidullah@gmail.com

Anamika Mondal
Dept. of Computer Sc.
West Bengal State University, W.B
anamikamondal2009@gmail.com

Kaushik Roy
Dept. of Computer Sc.
West Bengal State University, W.B
kaushik.mrg@gmail.com

***Abstract* - Script identification is a complex real life problem for automation of printed or handwritten document processing. The task becomes more challenging when it comes about a multi script/lingual country like India. For the development of OCR for a particular language the script needs to be identified first. That is why development of a script identification system is a pressing need. Till date no such work is available considering all 13 official Indian scripts. In this paper we present a scheme for script identification from printed document for 10 official Indian scripts namely Bangla, Devnagari, Roman, Oriya, Urdu, Gujarati, Telegu, Kannada, Malayalam and Kashmiri. Total 459 document pages are considered and 62 dimensional feature set is computed for the present work. Finally using simple logistic classifier with 5 fold cross validation an average identification rate of 98.9% is found.**

Keywords - OCR, Printed Script Identification, Feature Set, Simple Logistic Classifier

I. INTRODUCTION

Optical character recognition is an active area of research since many years. It is useful for converting the physical document into digital form for making a paperless world a reality in future. Document digitization also helps for better storage, distribution, indexing and retrieval of huge volume of data available in modern society. For a multi lingual and multi script country like India the work is more relevant. Here in India we have officially 13 different scripts including Roman and 23 different languages including English [1]. There are many languages which use same script for writing. As an example, Bangla is a popular script in the eastern part of India which is used to write Bangla, Assamese and Manipuri languages. Whereas Devnagari is a popular script, is used to write different languages like Hindi, Marathi, Nepali, Konkani etc. So here it is not possible to develop a general purpose OCR targeting a particular language. Before inputting the particular language to the OCR, script need to be identify first. That is why development of a script identification system is an essential requirement. Again another problem arises when a document is written using multiple scripts. For example postal document, pre printed filled up application form, many commercial advertisement etc. are example of such multi script document. In a real life postal document we can see the city name is written in one script whereas pin code is written in another script. In these cases it is impossible to determine the nature of target OCR for the particular language a priori. So script identification is must.

Script identification can be classified into two broad categories namely printed script identification and handwritten script identification. Again handwritten can be categorized namely offline script identification and online script identification. Few works are reported in literature on script identification based on Indian scripts and some other based on non Indian mixed with Indian scripts. Spitz [2] in his work identified Latin, Han, Chinese, Japanese, and Korean scripts by using features like upward concavity distribution, optical character density etc. He carried out his work on document level. Lam et al. [3] identified some non Indian scripts using horizontal projection profile, height distribution, presence of circles, ellipse, and presence of vertical stroke etc. features. Hochberg et al. [4] identified six scripts namely Arabic, Armenian, Devnagari, Chinese, Cyrillic, Burmese using some textual symbol based features. L Zhou et al. [5] identified Bangla and English scripts using connected component based features from both printed and handwritten document. Jayashree et al. [6] identified printed Gujrati script using cluster based templates. B. Patil and N.V. Subbareddy [7] proposed a tri script identification technique on English, Kannada and Hindi using neural network based classification. They performed their work at word level. A. M. Elgammal, M. A. Ismail [8] proposed a block level and line level script identification techniques from Arabic and English scripts using Horizontal projection profiles and run-length histograms analysis. Dhandra et al. [9] proposed a word level script identification technique from Kannada, Hindi, English and Urdu using morphological analysis. Chaudhuri B. B. and Pal U. [10] proposed a line based script identification techniques from roman bangle and devnagari scripts. Chew Lim Tan et al. [11] proposed a mixed script identification techniques considering Chinese, Latin and Tamil using upward concavity based features. Chaudhury S. and Sheth R. [12] proposed Gabor filter based script identification techniques from

English, Hindi, Telegu and Malayalam scripts. They performed the work at block level. In another work M.C. Padma and P.A Vijaya [13] proposed a work based on wavelet transform considering seven Indian and non Indian scripts namely English, Chinese, Greek, Cyrillic, Hebrew, Hindi, and Japanese. Using Multi Channel Log Gabor filter based features Joshi et al. [14] proposed a block level script identification technique from English, Hindi, Telegu, Malayalam, Gujrathi, Kannada, Gurumukhi, Oriya, Tamil, Urdu scripts. Dhanya et al. [15] proposed a word level script identification technique from Roman and Tamil scripts using Multi Channel Gabor Filters and Discrete Cosine Transform (DCT) based feature.

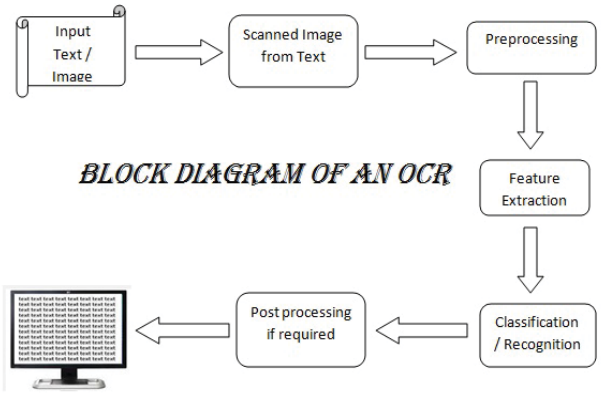

Fig. 1 *Block diagram of an OCR*

In this paper we propose a simple but efficient approach to identifying script written in any one of the ten different scripts namely Bangla, Devnagari, Roman, Oriya, Urdu, Gujarati, Telegu, Kannada, Malayalam and Kashmiri. The performance is evaluated using standard classifier and results are compared with other available works. The following Figure 1 shows block diagram of an OCR system, an Indian currency showing different official languages/scripts in India is given in figure 2. In Figure 3 different Indian scripts are shown form our sample database.

Fig. 2 Part of an Indian currency showing different official languages/scripts of India [22]

Fig. 3 Sample from our database of (a) Bangla, (b) Devnagari, (c) Roman (d) Oriya (e) Urdu (f) Gujarati (g) Telegu (h) Kannada (i) Malayalam (j) Kashmiri script documents.

The paper is organized as follows: In section II data collection and Pre-processing are described. In Section III Feature extraction techniques are discussed and the classification procedure with experimental result is described in section IV. Finally, Conclusion and scope of future work are given in section V.

II. Data Collection and Pre processing

One of the most important tasks is data collection. There is a problem of availability of standard database in this field of research. We have tried to collect real life printed script data from different sources like books, articles etc. Total 459 printed document pages are collected. Those documents are digitized using HP scanner. Out of 459 pages there are 60 Bangla, 60 Devnagari, 60 Roman, 58 Gujrati, 20 Oriya, 60 Telegu, 60 Kannada, 22 Kashmiri, 29 Malayalam and 30 Urdu script images. Some handwritten documents are also used for better evaluation of our system. The original images are in gray tone and digitized at 300 dpi. After completion of digitization preprocessing was done using the following steps.

Preprocessing:

- A two stage based approach is used to convert the images into binary (0 and 1). At first stage pre-binarization [16] is done based on local window algorithm in order to get an idea of different ROI.
- Run Length Smoothing Approach (RLSA) is applied on the pre-binarized image, to overcome the limitations of the local binarization method used earlier.

- Finally, using component labeling, each component is selected and map them in the original gray image to get respective zones of the original image and the final two tone binary image is obtained using histogram based global binarization algorithm [16] on these regions of the original image.

After this step a 62 dimensional feature set is computed. Details about the feature extraction process are described in the next section.

III. Feature Extraction

Selection of efficient features is the most important task in pattern recognition task. Performance of the system directly depends on the feature set chosen. Efficient features mean which are robust and easy to compute. Some features are computed based on visual observations also. In this work all the features are broadly classified into three major categories namely: (i) Structural Feature, (ii) Mathematical Feature and (iii) Morphological Feature. Altogether a set of 62 features are computed using Intel OpenCV [17] library. They are described as follows:

A. Structural feature

One of most important type of feature is structure based feature. This is calculated using connected component analysis. Initially outer and inner contours are extracted from the connected component, then features like: Circularity, Rectangularity, Convexity, Freeman Chain Code etc. are computed for each contour. Some of these features are discussed in the following sub sections.

1) Circularity

One of the key features among the structural category is circularity of a component. Many scripts like Oriya, Malayalam etc. have more circular nature than others. Following is the algorithm for calculation of circularity of a component.

Algorithm for calculating circularity of a component:

- *Minimum enclosing circle is drawn which will enclose the component minimally and the radius (r1) of the circle is being stored.*
- *Circle fitting is done. Circle fitting refers to the fitting of a circle in the component in as minimum manner as possible. Its radius (r2) is also stored.*
- *The difference of the two radii is stored to indicate the circularity of the component. The more the circularity of the component, the lesser will be the difference between the two radii.*

Calculate r1;
Calculate r2;
Calculate Circularity C = (r1-r2)
Store the result C;

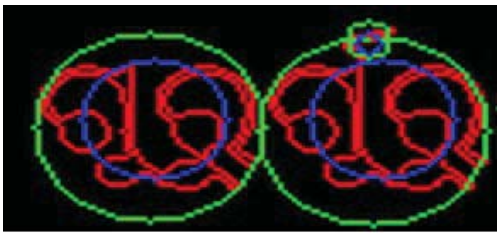

Fig. 4 Computation of Circularity of component on Oriya script using fitted circles (blue: minimum encapsulating & red: best fitted).

In fact the complete or almost complete circular components will have zero difference between the two radii or will have a difference tending to zero.

2) Rectangularity:

In Rectangularity structural feature we calculate the up-right bounding rectangle of the outer contours and inner contours of selected components. These evaluated rectangular boxes can be square, horizontal or vertical depending upon their length of height and width. The ratio of width and height determines whether the contour is square (ratio = 1), horizontal (ratio > 1) or vertical (ratio < 1).

Fig. 5 Computation of Rectangularity of component on Gujarati script showing by blue color

3) Freeman Chain Code based Feature:

Presence of horizontal and vertical line in the scripts is important feature for identification. In Bangla and Devnagari script horizontal line presence on the upper part of the scripts are called 'Matra' or 'Shirorekha'. This is a unique distinguishing feature of these two scripts from the rest. We use cvFindContours() function in OpenCV [17] in CV_CHAIN_CODE mode for identifying these lines as a sequence of integers as shown in figure below. Some slanting line presence in other scripts is also identified by the technique.

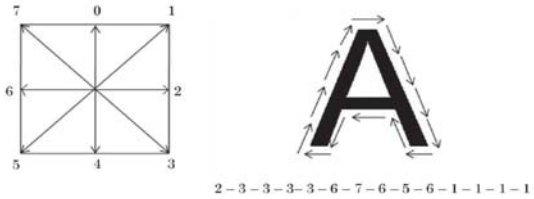

Fig. 6 Freeman chain code and an example

4) Component Based Feature

Using component analysis we have classified all the component into three categories namely (i) Large component, (ii) Medium component and (iii) Small component. Results are stored in LC, MC and SC variable and for all the targeted scripts we have counted number of LC, MC and SC. Here is the algorithm for calculating number of small component. The threshold value is considered as 5.

Calculation of small component:

Initially SC=0;
For each component calculate pixel count.
If Number Of Pixel (NOP) <= 5
SC++;
End

It is observed that script like Urdu contains maximum small components compared to others.

B. Other Important Features

Besides structural features we have used other features like mathematical and morphological features. These futures also play an important role for distinguishing scripts from one another. Mathematical features are extracted using Gabor filter bank using varied orientation like 60º, 90º, 120º, 150º. In morphological operations dilation and erosion functions are used with different structuring elements like horizontal, vertical, right angular and left angular for extraction of important features. These features also played an important role during the classification.

IV. Classification and Experimental Result

Weka [18], a machine learning tool is used for classification of different scripts after computation of all the features. This built in tools can be called from own Java code or using the weka.jar file of the package or directly from GUI interface. Its user interactivity makes it very popular nowadays. It contains tools for various applications like data pre-processing, classification, clustering, regression, association rules, visualization etc. There are many classifiers available in this tool. In our work we have used Simple Logistic Classifier for classification of all the scripts.

Simple Logistic Model Classifier

One of the popular classifier in Weka is Simple Logistic Model. It is a classifier for building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection. For more detail refer [19, 20]. Overall 98.9% average accuracy is obtained using 5 fold cross validation.

TABLE I. Confusion Matrix
B: Bangla, D: Devnagari, R: Roman, G: Gujarati, O: Oriya, Te:Telegu, Ka:Kannada, Ks:Kashmiri, M:Malayalam, U:Urdu

Script Type	Recognition Rate (%)									
	B	D	R	G	O	Te	Ka	Ks	M	U
B	100	0	0	0	0	0	0	0	0	0
D	0	98.3	0	0	0	0	0	0	1.7	0
R	0	3.3	95	0	1.7	0	0	0	0	0
G	0	0	0	100	0	0	0	0	0	0
O	0	0	0	0	95	0	0	0	5	0
Te	0	0	0	0	0	100	0	0	0	0
Ka	0	0	0	0	0	0	100	0	0	0
Ks	0	0	0	0	0	0	0	100	0	0
M	0	0	0	0	0	0	0	0	100	0
U	0	0	0	0	0	0	0	0	0	100
Overall Avg. Accuracy Rate: 98.9%										

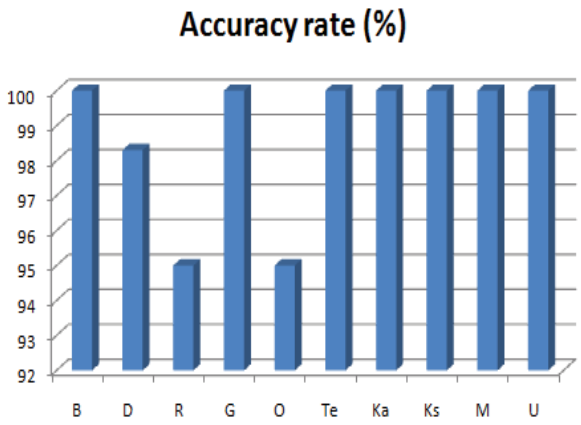

Fig. 7 Accuracy rate of different scripts

TABLE II. Comparative Study

Name of Algorithm	Total No. of Scripts	Scripts Considered	Avg. Acc. Rate (%)
Joshi et. al. [16]	10	Bangla, Devnagari, English, Gujarati, Kannada, Malayalam, Oriya, Gurumukhi, Tamil, Urdu	97.1
Dhandra et. al. [9]	04	Kannada, Hindi, English, Urdu	97
Proposed Method	10	Bangla, Devnagari, Roman, Gujarati, Oriya, Telegu, Kannada, Kashmiri, Malayalam, Urdu	98.9

V. Conclusion and Future Scope

A system for script identification considering ten popular Indian scripts is proposed here. Using simple logistic classifier with 5 fold cross validation an average identification rate of 98.9% is obtained. It is observed that the accuracy rate is not fully achieved because of the higher misclassification rate of Roman and Oriya scripts compared to others. We will try to handle this problem in future.

Still lot of attention need to given on script identification techniques for Indian documents. Till now no work is available based on all 13 official Indian scripts. But there is a problem of unavailability of standard script database in this kind of work. In future we want to extend our work first by considering all 13 official Indian scripts. Secondly we want to test the performance of the system using many standard classifiers for a real life script identification problem in both printed and handwritten domain.

References

[1] http://www.rajbhasha.nic.in/8thschedulehin.pdf

[2] A.L. Spitz, "Determination of The Script and Language Content of Document Images," IEEE Trans. Pattern Analysis & Machine Intelligence, vol. 19, no. 3, pp. 235-245, Mar. 1997.

[3] L. Lam, J. Ding, and C.Y. Suen, "Differentiating Between Oriental and European Scripts by Statistical Features," International Journal of Pattern Recognition & Artificial Intelligence, vol. 12, no. 1, pp. 63-79, Feb. 1998.

[4] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, "Automatic Script Identification from Document Images Using Cluster-based Templates," IEEE Trans. Pattern Analysis & Machine Intelligence, vol. 19, no. 2, pp. 176-181, Feb. 1997.

[5] Lijun Zhou, Yue Lu , Chew Lim Tan "Bangla/English Script Identification Based on Analysis of Connected Component Profiles", Lecture Notes in Computer Science,2006, Volume 3872/2006, 24354, DOI: 10.1007/11669487_22.

[6] Jayashree R. Prasad, U. V. Kulkarni, Rajesh S. Prasad, "Template Matching Algorithm for Gujrati Character Recognition," Second International Conference on Emerging Trends in Engineering & Technology, pp.263-268, 2009.

[7] Basavaraj Patil and N.V. Subbareddy. "Neural network based system for script identification in Indian documents", *Sadhana,* Vol. 27, part-i1, pp 83-97, 2002.

[8] A. M. Elgammal and M. A. Ismail, "Techniques for Language Identification for Hybrid Arabic-English Document Images", Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp.1100-1104, 2001.

[9] Dhandra.B.V, Nagabhushan. P, Mallikarjun Hangarge, Ravindra Hegadi, and Malemath. V.S, "Script Identification Based On Morphological Reconstruction In Document Images", The 18th International Conference On Pattern Recognition, 2006.

[10] Chaudhuri.B.B and Pal.U, "An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)", In Proc. 4th ICDAR, August, 1997.

[11] Chew Lim Tan, Peck Yoke Leong, Shoujie He, "Language Identification in Multilingual Documents", 2003.

[12] Chaudhury.S and Sheth.R, "Trainable script identification strategies for Indian languages", In 5th International Conference on Document Analysis and Recognition, pp. 657- 660, 1999.

[13] M.C. Padma, P.A Vijaya, "Wavelet Packet Based Texture Features for Automatic Script Identification", International Journal of Image Processing, vol. 4(1).

[14] Joshi, G.D., Garg, S., Sivaswamy, J., "Script Identification from Indian documents", In Seventh IAPR Workshop Document Analysis and Systems, New Zealand, pp. 255–267, 2006.

[15] D Dhanya, A G Ramakrishnan and P. B. Pati, Script identification in printed bilingual documents, Sadhana, vol. 27(1), pp. 73–82, February 2002.

[16] K. Roy, "On the Development of an Optical Character Recognition System for Indian Postal Automation", PhD Thesis, Jadavpur University, 2008.

[17]http://www.software.intel.com/sites/oss/pdfs/OpenCVreferencemanual.pdf

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Vol. 11, pp. 10-18, 2009.

[19] Niels Landwehr, Mark Hall, Eibe Frank, "Logistic Model Trees", Machine Learning, Vol. 95(1-2):161-205, 2005.

[20] Marc Sumner, Eibe Frank, Mark Hall, "Speeding up Logistic Model Tree Induction", In 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 675-683, 2005.