Script Identification of Machine Printed Documents Using NN Classifiers

[1]Ankita Ahuja, [2]Renu Singla
[1] M.Tech ,CSE, [2]Assistant Professor, CSE
SRCM

***Abstract* - In recent years there are many multimedia documents captured and stored with the advances in computer technology and hence the demand for recognizing and retrieval of such documents has increased tremendously .In such environment the large volume of data and variety of scripts make manual identification unworkable. In such cases the ability to automatically determine the script, and further the language of a document would reduce the time and cost of document handling. So the development of script identification from multilingual document image systems and then retrieving document image by matching with a query image (input image) has become an important task. It is noted that the research in this field is relatively thin and still more research is to be done, particularly in case of machine printed and handwritten documents. Here present a method developed to identify the script in machine printed document images automatically without manual intervention Particularly in case of machine printed and handwritten documents. The objective of this paper is to develop procedure to identify different text portions of a document. In this work eight feature namely top max row, bottom max row, top horizontal lines, vertical lines, bottom components, tick components, top holes and bottom holes have been used to identify the script type. Using these features two methods that is heuristic based algorithms and KNN approach proposed to identify the script type with the scripts of Telugu, Hindi , English, Bangla. There are a large number of different approaches to recognize the scripts currently available in OCR System. In this paper we look to identify the script of multilingual documents. In the proposed script identification system, we have considered different Indian languages such as English, Devanagari, Kannada, Gurumukhi, Bangla Script.**

***Keywords* - Multi Script Document, Script Identification, OCR, Tick components, Bottom components, KNN,** Heuristic Search

I. INTRODUCTION

Nowadays, we are living in the age of computer era. So we need to store paper documents in the form of the electronic documents to storage and facilitate easy communication. Generally, the usage of paper documents is still required in most of the communications. Basically, to the communication in the world wide we used the fax machine. Also, the important point is that, normal paper is a very comfortable and easy medium to do the communication. So, we need to have software, those can extract automatically and maintain data and information from paper documents. That information can be used for later retrieving those documents from the stored database. The approaches to solve these kinds of tasks are integrated under the general heading of document image analysis, which has been a fast growing area of research in recent years.The major operation of document image analysis is automatic extraction of text information from the paper document image. That can be achieved by a software tool that is Optical Character Recognition (OCR), which can be defined as the mechanism of manipulating the optically scanned text by the system. So we need an OCR system that can identify multi script because in multi-lingual country like India that covers 18 regional languages derived from 12 different scripts.

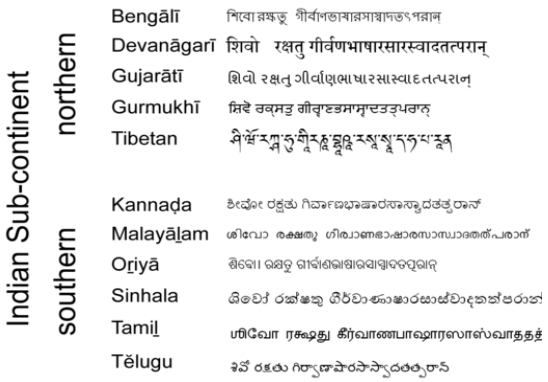

Figure 1 Sample of Multilingual document

II. EXISTING WORK

Existing works on automatic script identification are classified into either local approach or global approach. Local approaches extract the features from a list of connected components like line, word and character in the document images and hence they are

well suited to the documents where the script type differs at line or word level. In contrast, global approaches employ analysis of regions comprising of at least two lines and hence do not require fine segmentation. Global approaches are applicable to those documents where the whole document or paragraph or a set of text lines is in one script only. The script identification task is simplified and performed faster with the global rather than the local approach. Ample work has been reported in literature on both Indian and non-Indian scripts using local and global approaches.

From the literature survey, it has been revealed that some amount of work has been carried out in script/language identification. . Gopal et al. [1] have presented a scheme to identify different Indian scripts through hierarchical classification which uses features extracted from the responses of a multichannel log-Gabor filter.Peake and Tan [4] have proposed a method for automatic script and language identification from document images using multiple channel (Gabour) filters and gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Korean, Malayalam, Persian and Russian. Tan [1] has developed rotation invariant texture feature extraction method for automatic script identification for six languages: Chinese, Greek, English, Russian, Persian and Malayalam.

In the context of Indian languages, some amount of research work on script/language identification has been reported [8, 10, 11, and 13]. Pal and Choudhuri [5] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Devanagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya, Punjabi, Tamil, Telugu and Urdu) using ten triplets formed by grouping English and Devanagari with any one of the other scripts. Santanu Choudhuri, et al. [2] have proposed a method for identification of Indian languages by combining Gabour filter based technique and direction distance histogram classifier considering Hindi, English ,Bengali, Telugu and Urdu. Basavaraj Patil and Subbareddy [6] have developed a character script class identification system for machine printed bilingual documents in English and Kannada scripts using probabilistic neural network. Pal and Choudhuri [7] have proposed an automatic separation of Bangla, Devanagari and Roman words in multilingual multiscript Indian documents. Nagabhushan et.al. [10] have proposed a fuzzy statistical approach to Kannada vowel recognition based on invariant moments. Pal et. al. [9] have suggested a word-wise script identification model from a document containing English, Devanagari and Telugu text. Chanda and Pal [8] have proposed an automatic technique for word -wise identification of Devanagari, English and Urdu scripts from a single document. Spitz [18] has proposed a technique for distinguishing Han and Latin based scripts on the basis of spatial relationships of features related to the character structures. Pal et al. [12] have developed a script identification technique for Indian languages by employing new features based on water reservoir principle, contour tracing, jump discontinuity, left and right profile. Ramachandra et al. [13] have proposed a method based on rotation- invariant texture features using multichannel Gabor filter for identifying six (Bengali, Kannada, Malayalam, Oriya, Telugu and Marathi) Indian languages. Hochberg et al. [14] have presented a system that automatically identifies the script form using cluster-based templates The previous research work in the area of document script/language identification shows that much of them rely on script/languages followed by other countries and few from our country, but hardly few attempts focus on these three languages Telugu, Hindi and English.

III. PROPOSED SCRIPT IDENTIFICATION METHODS

In previous Optical character recognition system, if we provide a Multi-script document it would not identify that script with more accuracy and provides a degraded output in more times. In this research our main focus on the implementation of an approach that can help the Optical character recognition system in multi--script identification of a machine printed document.

Proposed Framework**:** To identify the multi-script input sample document general framework.

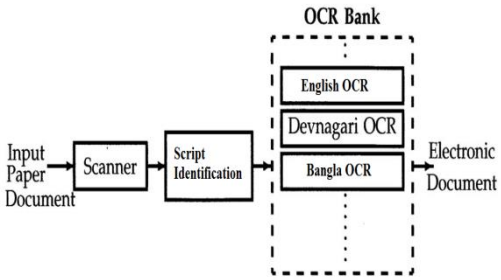

Figure 2 Proposed approach frameworks

In this paper, we discuss script identification methods namely heuristic approach and KNN based approach. These methods use various visual features to identify the script type from multilingual documents. Some of the visual features are:

- *Top max row (feature (F1)*
- *Bottom max row*
- *Top horizontal line*
- *Tick component*
- *Bottom component*
- *Vertical lines*
- *Top holes*
- *Bottom holes*

Based on these feature we proposed two methods namely heuristic and KNN based algorithm, to identify the script type from the given multilingual documents consisting of Telugu, Hindi and English scripts.

A. *Heuristic method*

This method uses simple heuristics that are formulated from the experimentation. If a word contains more than 60% of top horizontal lines or top max and bottom max rows are same then it can be considered as Hindi script. If a word contains tick like components or top and bottom holes it is considered as Telugu script and if it contains vertical lines then it is treated as English word. If the word doesn't fall in one of these scripts then it is considered as belonging to other class.

B. *KNN (K-Nearest Neighbor) based identification*

This method consists of learning and testing stages. In the training stage it learns the average values for each feature and in the second step these average values are used for classifying the unknown words. The learning
where N is the number of features in the feature vector f, $f_j(x)$ represents the j^{th} feature of the test sample X and $f_j(M)$ represents the j^{th} feature of M^{th} class in the knowledge base. Then, the test sample X is classified using the k-nearest neighbour (K-NN) classifier. In the K-NN classifier, a test sample is classified by a majority vote of its k neighbours, where k is a positive integer, typically small. If K =1, then the sample is just assigned the class of its nearest neighbour. It is better to choose K to be an odd number to avoid tied votes. So, in this method, the K -nearest neighbours are determined and the test image is classified as the script type of the majority of these K-nearest neighbours.

IV.RESULTS AND DISCUSSIONS

Proposed methods implemented in MATLAB. Since the standard dataset of Indian scripts is currently not available for the experimentation separate datasets comprising Telugu, English and Hindi are created from the internet news paper and text books.

Few separate data sets are used for training and others for testing. The heuristic method is tested with different number of Telugu, English and Hindi word. The Similarly for KNN based method the mean values of each feature for the three languages are computed as shown in Table below.

Language	No of script lines	Recognized correctly	Accuracy %
English	400	350	88
Hindi	350	320	91
Devanagri	200	180	90
Telgu	150	130	86

Table 1 Classification Result with Heuristic Based Method

Language	No of script lines	Recognized correctly	Accuracy %
English	300	270	85
Hindi	250	200	88
Devanagri	150	120	92
Telgu	100	93	93

Table 2 Classification results with KNN based method

The KNN based classifier could successfully identify the three script words (Telugu, English and Hindi) with an average accuracy of approx 90%. It is based on the average values of each feature and it is trained with large no of words of each script.

V.CONCLUSION AND FUTURE ENHANCEMENTS

The proposed heuristic approach could successfully identify the three types of script words (Telugu, English ,Devanagri and Hindi) with an average accuracy of approx 90%. The KNN based classifier could successfully identify the three script words (Telugu, English, Devanagri and Hindi) with an average accuracy of more than 85%. It is based on the average values of each feature. The system exhibits an overall accuracy of 87%. The work could be extended to word level script identification an for other Indian scripts. It can also be concluded that act a huge amount of work has been published on script identification. Also it can be noted that comparatively less amount of research has been done for script identification. This leaves us with ample opportunity to work in this field. These techniques can be extended to work with script identification of all the languages. It can also be extended for script identification of script's characters of old documents or which are written in curved shapes.

REFERENCES

[1] T.N.Tan, "Rotation Invariant Texture Features and their use in Automatic Script Identification", IEEE Trans. Pattern Analysis Sand Machine Intelligence, 20(7), 751- 756, (1998).
[2] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B.Shet, "Identification of Scripts of Indian Languages by Combining Trainable Classifiers", ICVGIP 2000, Dec., 20-22, Bangalore, India.
[3] M.C.Padma, P.Nagabhushan, "Horizontal and Vertical linear edge features as useful clues in the discrimination of multiligual

(Kannada, Hindi and English) machine printed documents", Proc. National Workshop on Computer Vision, Graphics and Image Processing (WVGIP), Madhurai, 204-209, (2002).

[4] G.S. Peake, T.N.Tan, "Script and Language Identification from Document Images", Proc. Eighth British Mach. Vision Conference., 2, 230-233, (1997).

[5] U.Pal, B.B.Choudhuri, "Script Line Separation From Indian Multi-Script Documents", Proc. 5th International Conference on Document Analysis and Recognition(IEEE Comput. Soc. Press), 406409, (1999).

[6] S.Basvaraj Patil, N.V.Subba Reddy, "Character script class identification system using probabilistic neural network for multi-script multi lingual document processing", Proc. National Conference on Document Analysis and Recognition, Mandya, Karnataka, 1-8, (2001). lingual document processing", Proc. National Conference on Document Analysis and Recognition, Mandya, Karnataka, 1-8, (2001).

[7] U.Pal B.B.Choudhuri, "Automatic Separation of Words in Multi Lingual multi Script Indian Documents", Proc. 4th *International Conference on Document Analysis and Recognition*, 576-579, (1997).

[8] S.Chanda, U.Pal, "English, Devanagari and Urdu Text Identification", *Proc. International Conference on Document Analysis and Recognition*, 538-545, (2005).

[9] U.Pal, S.Sinha, B.B.Choudhuri, "Word-wise script identification from a document containing English, Devanagari and Telugu text", *Proc. 2nd National Conference on Document Analysis and Recognition*, Karnataka, India, 213-220, (2003).

[10]P.Nagabhushan, S.A.Angadi, B.S.Anami, "A Fuzzy Statistical Approachto Kannada Vowel Recognition based on Invariant Moments", *proc. 2nd National Conference, NCDAR*, Mandya, 275-285, (2003).