Research Paper Engineering

Bilingual OCR for Printed English and Devnagari Text

* Aarti G. Ambekar ** Chhaya S. Hinge
*** Samidha S. Kulkarni

* ME Student,Electronics Dept.,K.J. Somaiya College of Engg.,Mumbai

** ME Student,Electronics Dept.,K.J. Somaiya College of Engg.,Mumbai

*** Associate Professor, Electronics Engineering Department, K J Somaiya College of Engineering, Mumbai

ABSTRACT

In this paper, a bi-lingual optical character recognition (OCR) for printed documents containing English and Devnagari based text is presented. The proposed OCR uses statistical based features namely, Zoning and Projection profiles for discriminating the characters. The classification is done with KNN classifier. The key contribution of the paper is to perform script identification along with font recognition which is further followed by respective character recognition. The proposed system provides satisfactory results

Keywords: Bilingual OCR, Script identification, Font recognition, Statistical Features, KNN classifier

1. Introduction.

Optical character recognition (OCR) is a field of research in pattern recognition, artificial intelligence and computer vision. It is mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. Recognition of printed characters is itself a challenging problem since there is a variation of the same character due to change of fonts. In a multi-script multi-lingual environment like India, a document may contain text lines in more than one script or languages. It is necessary to identify different script regions of the document in order to feed the document to the OCRs of individual language. Hence an OCR method is proposed here which is used for bilingual character recognition with script identification and font detection. Here English and Devnagari (consonants) printed characters are used.

2. English and Devnagari Script

The script Devnagari is mainly based on phonologically, and it is written from left to right. The national language of India that is Hindi along with Sanskrit, Nepali and Marathi language also use Devnagari script for writing. In Devnagari there are 12 vowels, 33 consonants with modifiers and some characters with somewhat similar shapes due to which its OCR design is a difficult task. English script consists of 26 capital letters and small letters respectively. Shirorekha or head line which can be seen as the main feature of Devnagari script as per Fig. 1 is absent in English [3]. Hence this can be considered as a main distinguishable feature of Devnagari and English script for script identification. Both scripts can be divided into three fictitious zones namely upper with ascenders and header line specifically for Devnagari, middle zone with most text part and lower zone with descenders. They are separated by four virtual lines as shown in Fig 1.

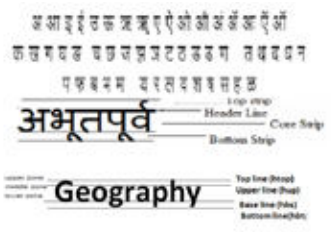

Fig 1: Devnagari characters & Zones

Height of character varies according to font size which can be considered as an important feature for font recognition. Heights estimating the four virtual lines used for separation of zones are as follows: -

htop = Height of character at top line
hup = Height of character at upper line
hbs = Height of character at base line
hbt = Height of character at bottom line

3. Basic steps in OCR:-

3.1 Preprocessing: - Under this mainly image acquisition is done with flat bed scanner normally with resolution of 600 to 1200 dpi. The digital acquired image is then converted into binary image using thresholding under binarization step. For binarization different

algorithms are used like Niback ,Ostu etc. Ostu algorithm is preferably used for getting better results. During scanning noise like salt and pepper may get introduced, which has to be removed by different filtering mask. Preferably median filters are used. If scanned image is tilted, skew detection and correction has to be applied. Normalization is done for getting a fixed size of image.

3.2 Segmentation: - In this step

line, word and character segmentation is done with the help of Horizontal and vertical projection histograms. Projection histograms count the number of object (black) pixels in each column and row of a character image. For background white pixel, dip occurs which can be used as thresholding points for segmentation [1, 2]. Projections of letter 'a' are shown in Fig 2.

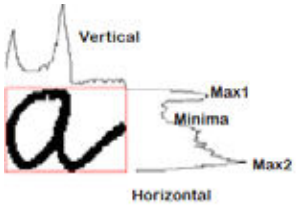

Fig 2:- Projections of letter 'a'

3.3 Feature extraction:-
This can be categorised in three types as [1, 2] :-

1. Statistical: Statistical properties like pixel density of character are used for discrimination.

2. Structural:- This uses structural features like aspect ratio, cross points of character for discrimination

3. Global transformations and moment:- Normally Fourier transform and invariant moments are used under this category for discrimination

3.4 Classification: -
Classification is done based on the extracted features of characters. Normally used classifiers are KNN, SVM, HMM etc. One can also go for neural network or combinations of classifiers.

4. Proposed Method

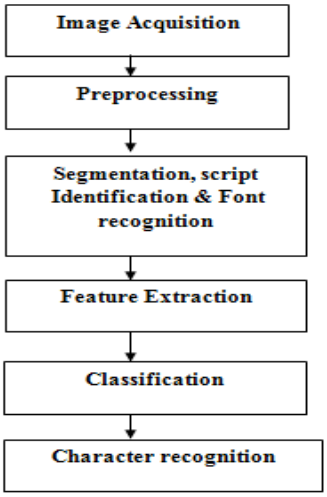

Fig 3 Proposed Method

1. Image acquisition is done for digitization of image with the help of flat bed scanner of resolution 600dpi.
2. Binarization of scanned image is done by using Ostu algorithm.
3. Horizontal and vertical projection profiles are obtained by scanning the image horizontally as well as vertically.
4. Script identification is done by extracting the features from horizontal projection profile as shown in Fig.2 as follows:-

a. Locate two maxima of object pixels, and find their average
b. Locate minima next to first maxima.
c. Find range R of object pixel as

$$R = \frac{\text{Minima}}{\text{Average of maxima}}$$

For Devnagari this range is observed as between 0.1 to 0.5 and for English it is greater than 0.5.

d. With the help of same profile based features font recognition can be done as follows[4]:

i. Different heights of the horizontal projection profiles as shown in Fig. 1 like hbt, htop, hbs & hup are located.
ii. Features used for font detection are calculated as follows:

- ht1= htop - hbt.
- ht2 = htop - hbs
- ht3= hup - hbs

iii. For font weight density of object pixel is calculated in between middle zone of character as:-

$$D = \frac{\text{Total Object pixel in middle zone}}{\text{Total number of pixel in middle zone}}$$

iv. Line word and character segmentation is done from horizontal and vertical projections by selecting threshold at valley points.

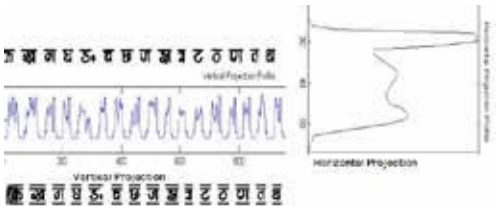

Fig.4 Segmentation of one of datasets

v. For character recognition required features are extracted using statistical approach based on zoning. For bringing the uniformity normalization of segmented character is done by resizing it to 48x48 pixel size. Normalized image is divided into 'n' number of zones as n= 4, 9, 16, 32. Pixel density of each zone is calculated by taking the ratio of total number of object pixels to total number of pixels in the zone. These 64 feature values are store as feature vector and used as input data for classifier.
vi. For classification KNN Euclidean distance classifier is used. In KNN classifier, training patterns are plotted in dimensional space, where required features are present. These patterns are plotted according to their observed feature values and are labeled according to their known class.

5. Experimental Results and Conclusions
For testing the proposed algorithm data sets are prepared for English small and capital letters with three fonts namely Times new Roman, Veranda and Courier New with font size from 12 to 28 having font weight bold and normal and for Devnagari consonants three fonts are used namely Shivaji01, Shivaji 05 and Kautilya (these are designed for Marathi language which uses Devnagari script) with font size from 12 to 28 having font weight bold and normal .These are manually created. For both script database is created using 10 samples of each character. As all the digitized images are noise free and without any skew hence noise removal and skew detection are not required. The summarized percentage results are tabulated in Table 1 as given below

Language	Script Identification	Font	Character Recognition
English	98%	90%	95%
Devnagari	97%	91%	93%

Table1. Percentage Results
The Bilingual OCR system has been built. The system includes segmentation modules, feature extraction modules, script & Font identification modules, English OCR system and Devnagari OCR system.

6. Future Scope:
The proposed system can be used for other scripts also. Here only printed texts are considered one can also go for handwritten text. Combinational classifier can be used for better recognition results.

REFERENCES

[1]. Holambe, Thool, Jagade "A Brief Review and Survey of Feature Extraction Methods for Devnagari OCR' IEEE transaction, Ninth International conference, pp99-104, 2011IEEE | [2]. N. Palrecha Rai, Kumar, Srivastava Tyagi, "'Character Segmentation for Indic Multilamngual Indic and Roman Scripts" IEEE 7th International Conference on Signal | Processing and applications, | pp 45-49, 2011 | [3] Chanda , Pal, Franke, "Two stage Approach for Word wise Script Identification", IEEE 10th International Conference, pp 926-93, 2009 | [4] Martin Solli and Reiner Lenz," A Font Search Engine for Large Font Databases", Electronic Letters on Computer Vision and Image Analysis 10(1):24-41, 2011 | [5] Anoop M. and Anil K.J., (2004) | "Online Handwritten Script Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.26, No.1, | pp.124-130. | [6] Wood. X. Yao. K.Krishnamurthi and Dang "Language Identification For Printed Text Independent Of Segmentation," Proc. Of Int'l. Conf. On Image Processing, | pp. 428-431, 1995. |