International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.6, pp : 730-733 1 June 2014

# Handwritten Character Recognition for English and Telugu Scripts Using Multi Layer Perceptions (MLP)

**P.V.Manoj, A.K.Sahoo, Samudra Gupt Maurya, Rohit Kumar**

Dept of Computer science, Sharda University,Greater Noida
pvmanoj45@gmail.com, ashok.sahoo@sharda.ac.in, samgpt1@gmail.com@gmail.com,
rkrohitkumar214@gmail.com

***ABSTRACT:*** ***Handwritten character recognition is frequently a frontier area of study in the field of prototype recognition and image processing and there is a large involve for OCR on hand written credentials. Even though, passable studies have perform in foreign scripts like Chinese, Japanese and Arabic characters, simply a very few work can be traced for handwritten disposition recognition of entire world scripts. In upcoming days, character recognition system might dish up as a key factor to create paperless environment by digitizing and processing existing paper documents. In this paper, we have provided the detail study on existing methods for handwritten character recognition.***

*Keywords- DataSet Generation, Feature Extraction, OCR for English & Telugu Hand-Written Text and English & Telugu Hand-Written Text*

## INTRODUCTION

English is among the scripts in around the world, with an increase on millions of speakers. The improvement of OCR in especially and Asian Indian scripts is really in a comparatively nascent stage, while it really is seen that OCR technology is in a mature stage of growth for English and other Roman / Latin scripts. Among the reasons is the sophistication of the orthography, particularly in Telugu. While possibly 10000 syllables are frequently used within the language, the orthographic units are composed by combinations of 16 vowels and 36 consonants. A practical OCR system for English &Telugu script was developed and proposed by Negi et al [3], where in actuality the complexity of English & Telugu script and tactics for its reduction were proposed. Their approach includes recognition and identification of connected components.

Within this paper we propose a better and robust recognition strategy which first uses the pixel distributions of the script and later exploits the structural information of English & Telugu orthography. Within this paper we don't discuss layout related problems for the isolation of English & Telugu text regions that will be taken up elsewhere.

## PROPOSED APPROACH

Following the strategy of Negi [1], we focus on recognizing from the sequence of 983 distinct glyphs, which are extracted as connected components from the input image. We tried a method which isn't greatly influenced by the measurement of the training set. Nevertheless, this would imply a method predicated on a candidate search and elimination technique. Our system is composed of the phases as shown in Fig. 1.

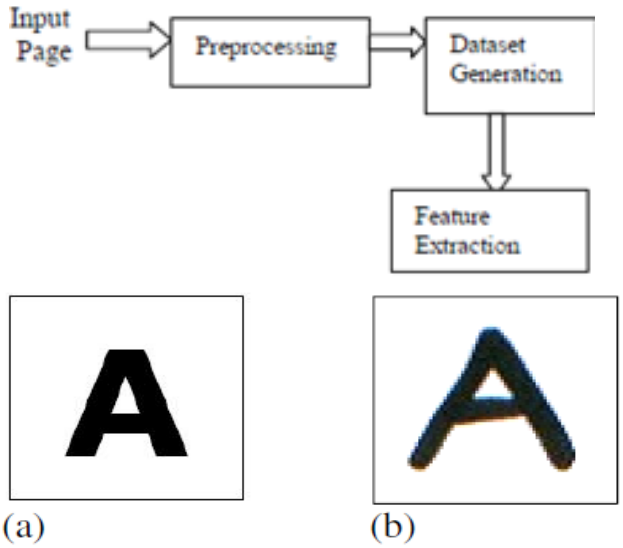
### *A. Input Page Description*

Input page is shown as follows in Fig. 2.

i) Forms are made with appropriate of letters on pages. It is made so that mechanical extraction is probable.
ii) Each row containing 10 characters excluding last page each page enclosed 90 characters. Totally 983 Telugu characters are to be overflowing with single handwriting.
iii) Each page enclosing a circle at the top of the page which denoting the side number of the one set of handwriting.
iv) Each page containing the parallel and vertical block at bottom right corner for in place of the right orientation of the check image.
v) Every vertical and horizontal line extends each feature or row for identifying the initial column or row position in the appearance of pixel.
vi) Each block is containing the width of 50 pixels and length of 60 pixels.

### *B. Preprocessing*

Preprocessing will improve the image clarity by cleaning- up the image and increasing the entrance value. The reprocessed image will give input to the subsequently phase.




(a). Optical Character, (b).Handwritten Character
Fig. 1. Phases of the OCR system

Fig. 2. Optical Character, .Handwritten Character Gray level image scanned with 300 dpi

*C. Dataset Generation*

i) It creates the 983 folders for 983 characters

ii) It takes the contribution of the image and check for direction of the page. If the page is not in right direction rotate 180 degrees

iii) Check for the circle location of the image for identifying what is the authentic character starting number

iv) Find the first row and column location of the page

v) Find the box synchronize

vi) Check for the bounding box quality coordinates

vii) Crop the character and put it on the individual folder

viii) Repeat (v) to (vii) until last box coordinates decision out

*D. Feature Extraction*

Candidate Search (zoning): For a candidate search we utilize the way of measuring density of pixel distribution in different zones of the input glyph for a feature vector. First the input glyph is broken into zones by superimposing a grid and the percent of the number of foreground pixels is calculated as in Fig. 2.

A code publication of this feature vector is pre-computed from the training set. The feature vector of the input glyph is computed and searched in the codebook to obtain k (5 in our case) nearest neighbors (n). The distance measure is Euclidean Distance between the feature vectors.

*E. Image acquisition*

Character recognition in general entails scanning a file and saving it in the computer system, that is utilized as an input picture to the character recognition issue. However the input to the system is completely different in the proposed work. In the proposed technique, any fundamental English & Telugu character from the given palm leaf is selected and pixels along the border of the character are identified. The (x,y) co ordinates for each pixel is measured using digital measuroscope (XY coordinate measuring instrument). Dial Indicator Plunger Assembly, having an accuracy of 0.01 microns least count, is employed to gauge the depth information at every pixel, which varies from 10 microns to 150 microns for assorted pixels across the contour.

*F. Correlation co-efficient for printed characters*

Initially for reference set, the printed English & Telugu characters are considered for grouping, using two dimensional correlation coefficient. Probably the most similar letters are grouped together which had correlation coefficient of 0.75 and above.

The two-dimensional correlation coefficient between two matrices A and B may be obtained provided An and B are matrices or vectors of exactly the same size.

$$r = \frac{\sum_m \sum_n (A_{mn} - A)(B_{mn} - B)}{\sqrt{(\sum_m \sum_n (A_{mn} - A)^2)(\sum_m \sum_n (B_{mn} - B)^2)}}$$

Where $\boldsymbol{A}$ = mean2 (A), and $\boldsymbol{B}$ = mean2

*G. Implementation procedure*

The values of X, Y, Z (numerical values) are recognized as the attributes of the pixel.

The Record is attain by separating each attribute by a comma and pursue by the class label.

The test and the information set for each character is made by concatenating the records of each disposition.

Figure 3. Flow chart showing general steps of application

Flow chart indicating general measures of application is described by Figure-3. Probably the most similar characters of English & Telugu are grouped in to one group namely Group-1. For each pixel in a character, there are 3 traits which are numerical as well as a Group name. Here the Coordinates of the pixel are X, Y where as Z is the aspect showing the depth of the indentation at that pixel that is proportional to the pressure applied by the Scriber at that point. By way of example in case a pixel of the character "a" has 1.091, 0.159, 24 since the values for X,Y and Z, then the Record for this pixel is 1.091, 0.159, 24, a. Working Out and test data set for every single character is created by concatenating these records of each character.

## SOFTWARE AND PROGRAMS USED IN THIS WORK

All the experiments are carried on a PC machine with P4 3 GHz CPU and 512MB RAM memory under JAVA 7.0 platform. Following the data acquisition, the (X, Y) data is utilized to plot the English & Telugu character and create exactly the same on the computer using MS Word. Further in the preprocessing stage. Photoshop computer software is used to normalize as well as the concept of minimal rectangle fit is employed to get all the characters into the size of 50x50 pixels size.

## RESULTS AND DISCUSSIONS

*Establishing the reliability of the data acquisition method used*

For establishing the reliability of the data acquisition system used within this work Printed English & Telugu characters are considered initially. By utilizing the correlation coefficient theory, a correlation matrix between each one of the Test English & Telugu character is created [8] with every other English & Telugu character and tabulated the correlation coefficient in XY plane, using the technique described in the recognition model before.

*Confusion of similar characters in XY plane*

Thinking about the grouping results obtained for the printed characters, as discussed previously within this section, it really is quite clear that there is lot of similarity and hence the confusion for those characters on XY plane in the same group. This is true even yet in the case of handwritten characters, and therefore the end result for character recognition obtained within this method is very low.

Three letters of one group having a correlation coefficient of 0.75 and above showed large amount of confusion within the standard XY plane, where as in the YZ plane, the patterns are wholly distinct from one another.

## RESULT SHEET

(As observed after running the proposed Model)

Using the devised tool after training the system with the training data, the test data is used to test the validity of classification. After running the tool, the evaluation result obtained is as follows:

Figure 5 is a graph which gives the percent of detection accuracy on divergent character groups.

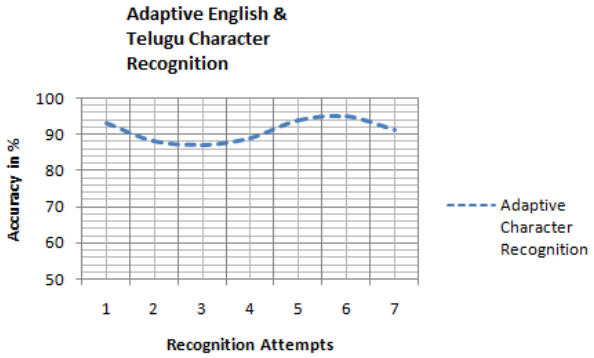



Figure 5: Accuracy in %

## CONCLUSIONS

This paper describes a novel system to recognize and classify English &Telugu characters. Hand written characters were extracted by measuring X, Y and Z co-ordinates of pixels of each and every character reproducibly. The intricacy of the characters in the language is reduced by dividing all of the characters into 6 sub-groups using printed characters and their correlation coefficient values. Z dimension could be the depth at that specific pixel that's directly proportional to the pencil / stylus pressure on the palm leaf.

Further, in the next thing, contemplating the Coordinates of X, Y, Z of each pixel of English & Telugu handwritten characters and proposed algorithm, the system is trained to distinguish and classify still another array of characters pertaining to these trained classes. Even the most similar English & Telugu Characters in precisely the same subgroup are successfully recognized and classified. A tree structure right along with the various nodes showing the various English & Telugu characters is obtained.

The proposed procedure can identify them definitely by use of 3D features in the recognition process, even though some of the letters like (a,e,i) (Pa, Ma, Va and Ya) are very similar to each other.

All of the characters of the test character file are analyzed with the database / skilled characters. 87% for X attribute and the Attribute Usage is found to be 100% for Z, whereas 93% for Y attribute. This tells that Z, I.e., the depth of indentation at any given pixel is proportional to the pen pressure used at that pixel point is an essential aspect for classification and identification. For this classification and character recognition, it takes very little quantity of time I.e. 0.1 Seconds.

This process is employed just to the basic English & Telugu characters and can really be extended for combination of 2 or even more basic characters also in future studies. The system of

data collection can be enhanced in future by automated procedure of measurement like using a laser technique in place of manual data collection mostly for the Z dimension (depth info).

## REFERENCES


i. *Shi Zhixin, Setlur Srirangaraj and Govindaraju Venu. 2005. Digital Image Enhancement Using Normalization Techniques and their Application to Palm Leaf Manuscripts. CEDAR. Center for Excellence for Document Analysis and Recognition. New York. U.S.A*

ii. *Ashwin T. V and Sastry P.S. 2002 A Font and Size- independent OCR system For Printed Kannada documents Using Support Vector Machines. Sadhana. Vol. 27, Part 1. 35-58.*

iii. *Bunke H, Roth. M, Schukat-talamazzini. E.G. 1995. Off line Cursive Handwriting recognition using Hidden Markov Models. Pattern Recognition. Pergamon. pp. 1399-1413.*

iv. *Gader Paul D, Keller James M, Krishnapuram Raghu, Chiang Jung-Hsien and Mohamed Magdi. A. 1997. Neural Methods in Handwriting Recognition. Research Feature. IEEE. pp. 79-85.*

v. *Joshi Niranjan, Sita G, Ramakrishnan A.G and Madhvanath Sriganesh. 2004 Tamil Handwriting recognition using subspace and DTW based Classifiers. Springer. Vol. 3316/2004. pp. 806-813.*

vi. *Rao P.V.S and Ajitha T.M 1995. Telugu Script Recognition-A feature Based Approach. ICDAR. IEEE. pp. 323-326*

vii. *Aradhya Manjunath. V.N, Hemantha Kumar. G and Noushat.S. 2007. Multilingual OCR system for South Indian Scripts and English Documents: An Approach based on Fourier Transform and Principle component Analysis. Engineering Applications on Artificial Intelligence. Elsevier.*

viii. *Panyam Narahari Sastry, Ramakrishnan Krishnan, Bhagavatula Venkata Sanker Ram. November 2008. Telugu Character Recognition-A three dimensional Approach. Technology Spectrum. Volume 2, No. 3, 19-26.*