ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 12, December 2013

Handwritten Character Recognition for Telugu Scripts Using Multi Layer Perceptrons (MLP)

C. Vikram C. Shoba Bindu C. Sasikala

Abstract-- **Handwritten character recognition is constantly a frontier area of study in the field of pattern recognition and image processing and there is a large requirement for OCR on hand written credentials. Even though, adequate studies have performed in foreign scripts like Chinese, Japanese and Arabic characters, simply a very little work can be traced for handwritten disposition recognition of Indian scripts particularly for the South Indian scripts.**

A hybrid HMM/ANN (Hidden Markov Model/ Artificial Neural Network) system for recognizing unconstrained offline handwritten text lines has been proposed by Salvador Espanaet. Al [13]. The key features of the recognition system are the novel approach to pre-processing and recognition, which are both based on ANNs. The pre-processing is based on using MLPs (Multilayer Perceptron): To clean and enhance the images, to automatically classify local extreme in order to correct the slope and to normalize the size of the text line images, and to perform a non uniform slant correction. The recognition is based on hybrid optical HMM/ANN models, where an MLP is used to estimate the emission probabilities. In this paper Multilayer Perceptron model is applied to handwritten TELUGU script.

I. INTRODUCTION

TELUGU is one of the scripts in India and Asia, with an increase of more than 846 million speakers. The improvement of OCR especially in Asian and Indian scripts is really at 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 of combinations of 16 vowels and 36 consonants. A practical OCR system for Telugu script was developed and proposed by Negi et al [3], where in actuality the complexity of 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 a robust recognition strategy which first uses the pixel distributions of the script and later exploits the structural information of Telugu orthography.

II. 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.

2.1 Input Page Description

Input page is shown as follows in Fig. 2.

i) Forms are made with appropriate letters on the pages. It is made so that mechanical extraction is probable.

ii) Each row contains 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 to each feature or row for identifying the initial column or row positions in the appearance of the pixel.

vi) Each block contains the width of 50 pixels and a length of 60 pixels.

2.2 Preprocessing

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

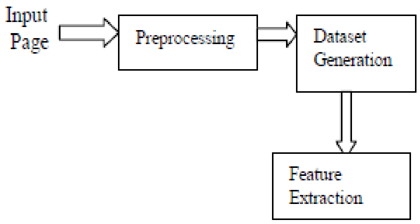


Fig. 1: Phases of the OCR system

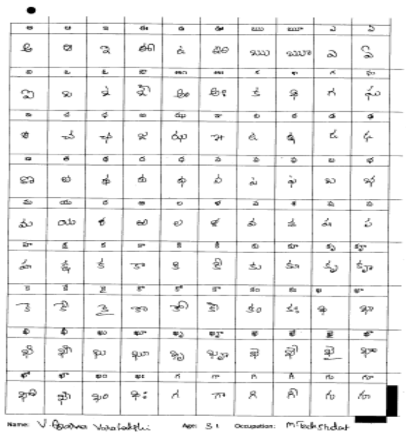


Fig. 2: Gray level image scanned with 300 dpi

2.3 Dataset Generation

i) It creates the 983 folders for 983 characters

ii) It takes the contribution of the image and check for the 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.

2.4 Feature Extraction

2.4.1 Candidate Search (zoning):

For a candidate search we utilize the way of measuring density of the 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. Accord 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 neighbours (n). The distance measure is the Euclidean Distance between the feature vectors.

2.5 Image acquisition

Character recognition in general entails scanning a file and saving it in the computer system that is utilized as an input picture of the character recognition issue. However the input to the system is completely different in the proposed work. In the proposed technique, any fundamental Telugu character from the given image is selected and pixels along the border of the character are identified.

The pre-processing is based on using MLPs (Multi-Layer Perceptron):

1. Image is cleaning.
2. Slope Removal.
3. Normalization.

2.6 Recognition Performance Measuring

In terms of Word Error Rate the recognition Performance is measured. It is calculated by comparing the output of the recognizer with the reference transcription.

$$\text{WER} = 100 * \frac{\text{insertions} + \text{substitutions} + \text{deletions}}{\text{total number of words}}$$

Similarly the Character Error Rate (CER) is calculated by replacing characters instead of words.

2.7 Implementation procedure

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

b) The Record has attained by separating each attribute by a comma and pursue by the class label.

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

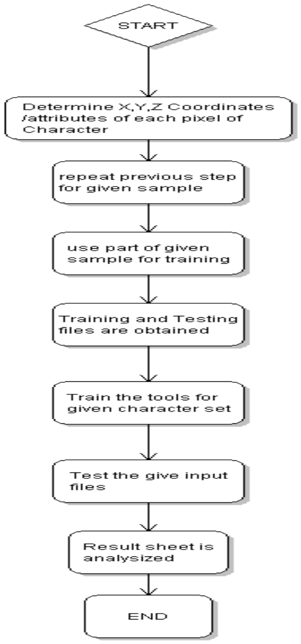
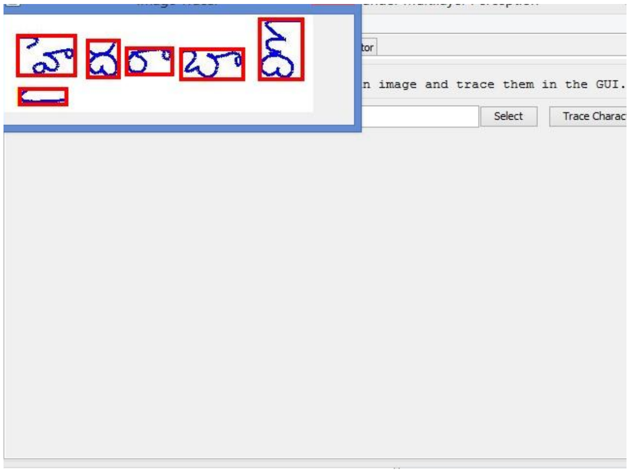

START
Determine X,Y,Z Coordinates /attributes of each pixel of Character
repeat previous step for given sample
use part of given sample for training
Training and Testing files are obtained
Train the tools for given character set
Test the give input files
Result sheet is analysized
END

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 Telugu are grouped into one group namely Group-1. These are (a, aa, ala, bra, la, ta, tha, khaa).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 "aa" 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.



n image and trace them in the GUI.

Fig 4: Tracing the Telugu Characters

Figure 4 clearly shows the tracing done by the proposed scheme where each and every letter is rounded off with the red colour.

III. RESULTS AND DISCUSSIONS

3.1 Establishing the reliability of the data acquisition method used

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

3.2 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 are lots of similarity and hence the confusion of 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 a large amount of confusion within the standard XY plane, where as in the YZ plane, the patterns are wholly distinct from one another. Images in Figure-7 are samples of the characters pertaining to an individual sub-group with maximum confusion.

IV. RESULT SHEET

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:

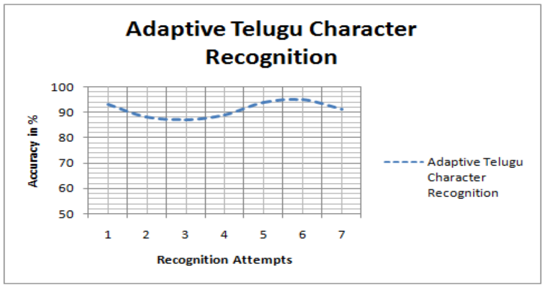


FIG 5: ACCURACY IN %

An example set of 29 distinct Telugu characters is analysed and 27 out of 29 of these were properly recognized.

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

In order to show better performance the proposed scheme is compared with another scheme called i2-OCR. We provided both the schemes with 100 words as input and using the Word Error Rate (WER) formula as a base we calculated the number of correctly classified words and results are tabulated in table 1.

Table 1: Results of Comparison

OCR Type	NO. Of Inputted Words	No. Of Correctly classified words	Accuracy
i2 OCR	100	76	76%
OCR Using MLP	100	85	85%

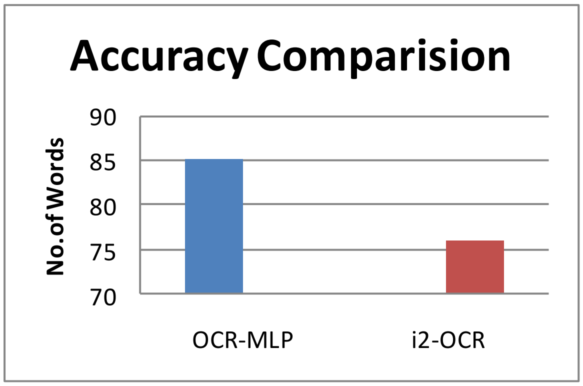


Fig6: Comparison of OCR accuracy with i2-OCR

Figure 6 shows the comparison in the form of a graph where proposed scheme has recognition accuracy of 85% while i2-OCR has only 76%.

V. CONCLUSIONS

HTC (Handwritten Telugu Character) recognition has been reported little in Literature.

This paper describes a novel system to classify Telugu (a south Indian language) characters using MLP. Hand written characters were extracted by measuring X, Y and Z coordinates of pixels of each and every character reproducibly. Further, in the next thing, contemplating the Coordinates of X, Y, Z of each pixel of 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 Telugu Characters in precisely the same subgroup are successfully recognized and classified. A tree structure right along with the various nodes showing the various Telugu characters is obtained.

This process is employed just to the basic Telugu characters and can really be extended for a combination of 2 or even more basic characters also in future studies. The system of data collection can be enhanced in future by an automated procedure of measurement like using a laser technique in place of manual data collection mostly for the Z dimension (depth info).

REFERENCES

[1] Atul Negi, Chandra Kanth Chereddi. Candidate Search and
Elimination approach for Telugu OCR, TENCON Conference, 2003.

[2] Atul Negi, Chakravarthy Bhagavati, and B. Krishna. An
OCR System for Telugu. In Proceedings 6th ICDAR, Seattle, USA, 2001.

[3] Atul Negi, Nikhil Shanker, and Chandra Kanth Chereddi
Localization, Extraction and Recognition of text in Telugu
Document Images. In Proceedings of the 7th ICDAR, Edinburgh, Scotland, 2003.

[4] 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

[5] 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.

[6] Bunke H, Roth. M, Schukat-talamazzini. E.G. 1995. Offline Cursive Handwriting recognition using Hidden Markov Models. Pattern Recognition. Pergamon. pp. 1399-1413.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] Atul Negi, “Dataset Generation and Feature Extraction for Telugu Hand-Written Recognition”, 2012.

[13] Salvador Espan˜a-Boquera and Maria Jose Castro-Bleda, “Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 4, April 2011, page.no:767-779.