International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 1869-1880

ONLINE TELUGU HANDWRITTEN CHARACTER RECOGNITION USING EFFICIENT MACHINE LEARNING APPROACHES

[1]P.V. Ramana Murthy, [2] P.Andrews Hima Kiran, [3]S. Ajay Kumar, [4]Pattola Srinivas

[1]Research Scholar, Rayalaseema University & Assoicate Professor, CSE Department,Malla Reddy Engineering College (A),Hyderabad, Telangana, India
[2]Assoicate Professor, CSE Department,Malla Reddy Engineering College (A),Hyderabad, Telangana, India
[3]Research Scholar, KLEF Vaddeswaram &Assoicate Professor, CSE Department,Malla Reddy Engineering College (A),Hyderabad, Telangana, India
[4]Assoicate Professor, CSE Department,Malla Reddy Engineering College (A),Hyderabad, Telangana, India

Abstract

In Pattern recognition, Online Telugu Handwriting Recognition (OTHWR) has become one of the recent research areas of interest due to exponential use of resources such as paper documents, photographs, Smartphone, and iPods. Telugu Handwriting differs from person to person and it is a difficult task to recognize the Telugu characters. Telugu Handwriting Recognition is categorized into two ways - online and offline. Online Telugu character recognition involves conversion of digital pen-tip movements into a list of coordinates. Telugu character recognition is considered as one of the most critical components which enable a data processor to distinguish letters and digits possibly using the contextual data. Various attempts in resolving this problem by using different selections of classifiers and features have been established and still the problem is remaining challenging. In the proposed work, we have used Optical Character Recognition System and various machine learning technique i.e., Convolution Neural Network (CNN) and Support vector machines. In SVM, we have constructed the stroke recognition engine and the characters have been represented as a sequence of strokes and features have been extracted and classified. Support vector machines for Telugu language (south Indian) character recognition algorithm with high recognition accuracy and minimum training and classification of time. A qualified analysis has performed to test the efficiency of the proposed models against previous methods on an interesting dataset. In observations, it was found to have improved than that of some of the recent expressions made in literature usage for the identification of online handwritten Telugu handwriting characters.

Keywords*: Pattern Recognition,Optical Character Recognition System, CNN,SVM*

I. INTRODUCTION

Online Telugu character Recognition has become one of the active and challenging research areas in the field of image processing and pattern recognition due to the exponential use of the resources such as paper documents, photographs, smart phone, and ipads. Telugu is a well-known language spoken in the southern part of India. It has

ISSN: 2233-7857 IJFGCN

16 vowels and 36 consonants. A single Telugu character (Aksharam) in an Indian language typically represents an entire syllable. A single Telugu character can be a) pure vowel or b) One or more consonants. Because of this reason, the total numbers of vowels and consonants have been numbered to be 52.The complexity in character recognition varies among different languages due to distinct shapes, strokes and number of characters. In India, Telugu (South Indian language) has been ranked to be third by the number of native speakers. Telugu language mostly contains many similar shaped characters. In many cases a Telugu character differ in conveying various shades of nasal sounds from its similar full-zero (anusvāra) (◌ం), half-zero (arthanusvāra or candrabindu) (◌ఁ) and visarga(◌ః).Hence, because of this it makes a difficult task for achieving a better performance in Telugu character recognition.

II.OBJECTIVES

The development and advantages of computer technology has been extensively used nowadays and the usages of computers become more and more demanding in our everyday life. Machine simulation of human functions has become a challenging research field due to this advent of digital computers[4]. An automated online Telugu language character recognition technique stands as a solution, which will interpret characters automatically. The automatic OTHWR is a challenging problem since there are number of a variation of same character has been notices due to the change of fonts and sizes in Telugu language.

The differences in font types and sizes made the recognition task difficult. In result, the recognition of Telugu character process has become tough. Keeping in view of this problem, it has been observed t penlight as a set of objectives.

2.1 Objectives:

Step 1: Analyze Telugu script for building a model script for better documentation for an easy understanding.
Step 2: Develop a system to recognize Telugu characters online and offline using machine learning models such as Optical Character Recognition, Support vector machine, Convolutional Neural Network respectively
Step 3: Performance analysis of these machine learning models for different Telugu language character datasets.

2.3 Telugu Language

Telugu, a language is primarily spoken in the states of Andhra Pradesh and Telangana India. It has the third largest number of native speakers in India [5].Telugu language is written in a script originated from the Brahmi script, Telugu is a south-central Dravidian language influenced by Sanskrit and Prakrit, as well as Urdu.

ISSN: 2233-7857 IJFGCN

Script: Onamaalu or the Telugu alphabet consists of 56 symbols - 16 vowels, 36 consonants, and 4 other symbols. It is highly conducive for Phonetics. Today only 12 vowels, 31 consonants are being used. In our model, we have included all 56 characters. The symbols in the Telugu language are broadly divided into the following classes.

1. Vowels
2. Consonants
3. Vowel diacritics
4. Conjunct consonants

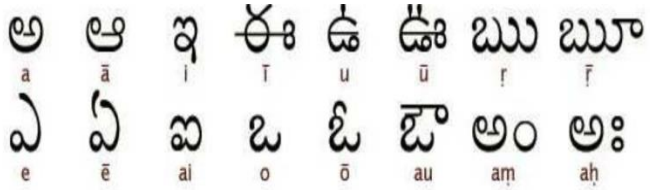

Figure 1: Vowels in Telugu language

Figure 2: Consonants in Telugu language

III. TELUGU CHARACTER RECOGNITION SYSTEM.

It is divided into two types

1) Offline Handwritten Character Recognition

2) Online. Handwritten Character Recognition.

1. Offline Handwritten character recognition is a procedure of the recognition the scanned handwritten image or document.[7]

2. Online handwritten character is a method of recognizing characters by a machine while the user writes; here handheld devices are used for identifying the character by recording the (x, y) coordinates of the track of the character. Teluge character images are pointed by some special digitizers[1] or PDA or any touch screen devices, a sensor picks up the pen-tip movements as well as pen-up/pen-down switching and user's written strokes are taken into consideration by captured sampling the pen's (x, y) coordinates at evenly spaced time intervals pen up and pen down of pointing Teluge character images.

ISSN: 2233-7857 IJFGCN

Figure 3 : Some of the character recognition tools

IV.METHODOLOGY

4.1. Optical Character Recognition System

Optical Character Recognition System is the process of extracting text from the document images. The input to the system is a scanned document, and the output of the system is Unicode or text in the document. The process begins with pre-processing the scanned image.[8]

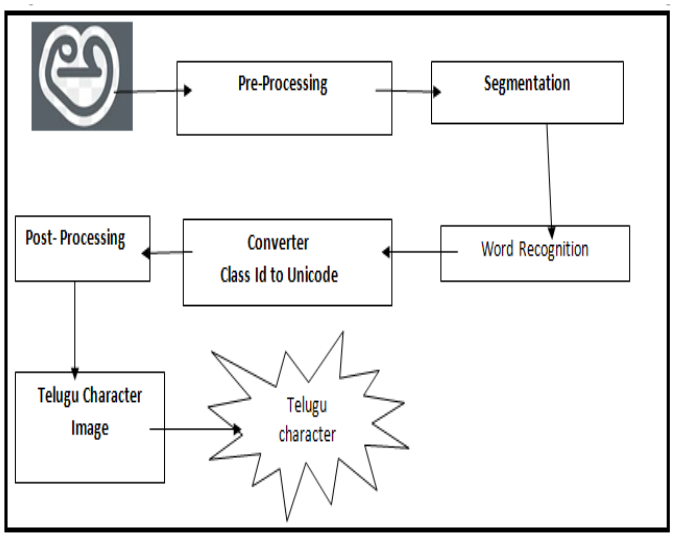

Figure 4. Optical Character Recognition Process

i. **Pre-Processing module:** It includes modules like noise removal, converting a color image into a gray scale image, thresholding the gray scale image into a binary image, and finally skew-correction of the image.

ii. **Segmentation Module:** After preprocessing Lay-out analysis is done with various levels of segmentation like block/paragraph level segmentation, line level segmentation[2], word level segmentation and component/character level segmentation. [9]

ISSN: 2233-7857 IJFGCN

iii. **Word Recognition module:** The input to the recognition module is a Telugu word image.

iv. **Telugu character** symbol are passed as an input to the Word recognizer module and the features are extracted. The recognizer has a base classifier with a very high performance. It recognizes the isolated symbols.

v. **Feature Extraction: It** is used to extract the maximum features of the available raw data. Here we concentrate mostly on the pen tip position vector velocity and density, Aspect ratio, percent of pixels above and below the axis average distance from the image centre if we can maximum feature from the input then the probability of recognizing the character improves.

vi. **Classification: The last and final big step in the online Telugu character recognition process is the classification.**

Classifiers:

- Nearest neighbor classifier is one of the most popular classifiers.
- K-nearest neighbor (KNN)[3] is a supervised learning algorithm, which is an extension of the nearest neighbor classifier.
- Decision tree classifier
- Naives Bayes (NB) classifier
- Neural network classifiers: Multi-Layer Perception (MLP) and Convolution Neural Network (CNN) Classifier.
- Support Vector Machines (SVM) Classifier
- Support Vector Machines have become very popular for high accuracies and its ability to generalize.

In our proposed model we have used Convolution Neural Network (CNN) Classifier and Support Vector Machines (SVM) Classifier to map the extracted features to different classes for identifying the online Telugu characters

vii. **Post processing:** The input of this module is the class labels from the classifier module. The output of this module is the Unicode of the word. The class labels are then reordered according to the language, they are then mapped into Unicode.[10]

OCR Model Evaluation:

By using OCR technique for identifying the Telugu character on real data. I have chosen different character images in such a way that most of the classes are covered of Telugu character images and trained data and achieved with a symbol accuracy rate of 90.78%.

Result:

Method	**Accuracy**
Optical Character Recognition System for Telugu character	90.78%

Dataset

The choice of the dataset is the key for OCR systems. Here, we propose a dataset which takes into consideration all possible combinations of vattu and guninthas with all

ISSN: 2233-7857 IJFGCN

possible categories and nearly 560 samples per class. All the images are of size 32x32. Our dataset is original because unlike other datasets which only take into account the commonly occurring permutations of characters and vattus, we have spanned the entire Telugu alphabets and their corresponding vattu and guninthas.

4.2 **Use of Convolutional Neural Networks (CNN's) for Telugu character recognition**

In the proposed work, we used a deep learning construction for recognizing the Telugu handwritten characters.

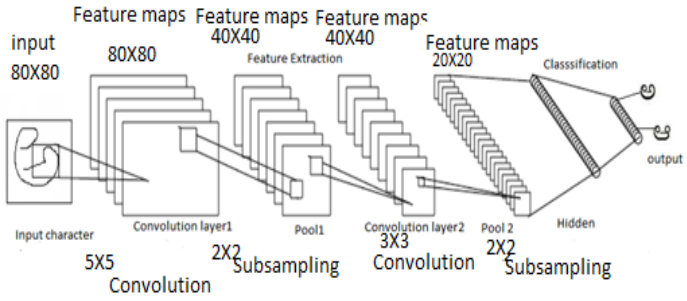

Figure 5: CNN Layers using Telugu characters

This deep learning network consists of convolutional layer, rectified linear unit, pooling layer and fully connected layer continued by an output layer as shown in Figure.[11]
In CNN there are three types of layers:
1) Convolutional layer
2) Pooling layer
3) Fully-connected layer

1. Convolutional layer: In this layer few parameters
Like number of filters, size of filters, stride, etc. Small window filter slid along with the dimensions of input data and performs dot products between the values stored in the filter and the input data points

2. Pooling layer: This layer reduces the dimensionality of the input data which reduces the computations, number of parameters and therefore reduces over fitting. Typically, the pooling layer is inserted between convolutional layers. It discards the activations of previous layers and hence forcing the next convolutional layers to learn from a limited variety of data

3. Fully-connected layer: In this layer neurons that are connected to all neurons of the previous layer as explained it.
Step1: During the Identification phase select a hand written character image for classification.
The input layer holds the raw pixel values of the selected image of height and width 80 X 80 for Telugu characters.

ISSN: 2233-7857 IJFGCN

Step 2: The selected input image is passed to the convolution layer. In this layer a random number of filters are used to proceed along the height and width of the image to yield a feature map. During this phase a feature is obtained by sliding each filter across the height and width of the image, compute the dot products between the input volume and filter during the forward pass.

The output of the first convolution layer creates 32, 4 such feature maps in Telugu and then it is transformed to the next layer through a differentiable function.

Step 3: Lastly, the output is of 3D (80 X 80 X 32) and (32 X 32 X 4) which is transformed to first pooling layer and the image is down-sampled along the spatial dimensions resulting in an output volume of (40 X 40 X 32).

Mathematical representation:

$$x_i^l = f\left(\sum_{i \in M_j} x_i^{l-1} k_{ij}^{\ l} + b_j^l\right)$$

In the same manner second convolutional procedure creates 32, 4 different feature. A size of 2 X 2 and 4 X 4 filters results a feature map size of 40 X 40 down sampled into 20 X 20. Further down-sampling in the pooling layers produces resizing feature maps of size 5 X 5.

Down-sampling mathematical function

$$x_j^l = f\left(\beta_j^{\ l} down(x_j^{l-1}) + b_j^l\right)$$

4.3 CNN Model Evaluation:

By using Convolution Neural Networks (CNN's) for Telugu character recognition I tried different optimizers with different learning rates by changing the number of layers and number of filters and filter size. I have achieved a test accuracy of 91% and training accuracy of 95% on Telugu character dataset with 20 epochs, by increasing the number of epochs, the accuracy will also increase to further level. Finally, I have achieved with accuracy rate of 98.3%.

Result:

Method	**Accuracy**
Convolutional Neural Networks (CNN's) for Telugu character recognition	98.3%

5. Use of Support Vector Machine (SVM) for Telugu character recognition:

ISSN: 2233-7857 IJFGCN

On-line Telugu character recognition approach involves automatic conversion of text into letter codes / Unicode. In online text reorganization approach, characters are dynamically obtained on digitizer or tablet PC. During reorganization phase sensor picks up the co-ordinates of the trajectory of the stroke X(t), Y(t) from the pen-tip movements as well as the instances of pen-down and pen-up. Telugu script consists of 16 vowels and 36 consonants. The characters in Telugu script combination give nearly 18,000 unique characters. All these unique characters in Telugu can be represented as a combination of set of 235 strokes. In online recognition of characters, the recognition problem is divided into recognizing strokes in each tier separately, and put together the strokes to determine the character. Character strokes can be divided based on the position of the stroke into 3 tiers – Top stroke, Bottom stroke and baseline auxiliary stroke. A stroke is considered to be a set of coordinates form one pen-down to next pen-up. Each character is collected as a bunch of one or more strokes. The online stroke data contains both the temporal information (writing process) and the spatial shape information of the characters.
A character recognition engine is used for both these stroke information to attain robust performance.

5.1 Stroke Pre-Classifications: A typical Telugu character can be divided vertically into three tiers: Top, middle and down. Each character has a baseline stroke and usually one or more attached strokes at the top, bottom and side of the base stroke. Also there can be a lot of size variation in the strokes, with some of the strokes having very few constituent points for proper identification.

Strokes are classified into four categories:

1. Main stroke
2. Baseline auxiliary
3. Top stroke and
4. Bottom stroke.

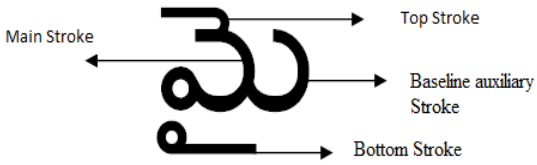

Figure 6: Pre-Classification of a Telugu

Out of the total number of 235 strokes, there are 149 main strokes, 19 baseline strokes, 26 top strokes and 41 bottom strokes. The first stroke written in a character is considered to be as the main stroke, baseline auxiliary, bottom and top strokes can be written in any order. So to classify baseline auxiliary, bottom and top strokes, two methods are adopted.

Method: 1

The baseline auxiliary stroke is identified by a histogram of the y-co-ordinates of the stroke. The baseline stroke is the stroke closest to the interval with the least number of points.

ISSN: 2233-7857 IJFGCN

The top and bottom strokes are identified by considering their average vertical displacement from the horizontal line.

Method 2 – SVM: This method is based on a SVM. Here I constructed a feature vector by concatenating the main stroke and with the stroke that is being pre-classified. Both strokes are re-sampled, so that either stroke has only 16 points. These Feature vectors are used to train a SVM-based pre-classifier.

5.2 Feature Extraction

Once the strokes are pre-classified, feature vectors for stroke recognition are constructed. Different types of features are explored for best classification results like 1) X and Y coordinate points, 2) Fourier transforms, 3) Hilbert transform logarithm of the spectral density, 4) Wavelet features.

5.3 Stroke Recognition using SVMs:

The extracted features from baseline auxiliary, bottom and top stroke pre-classes are given as input to SVMs with Gaussian kernel to recognize the specific stroke.

This method uses 3 SVMs for classifying the main stroke. The feature vector extracted from the main stroke is passed to two different SVMs: Vowel classifier and Consonant classifier. Output vectors of the Vowel and Consonant SVM classifiers are concatenated and passed to third SVM for classifying the main stroke.

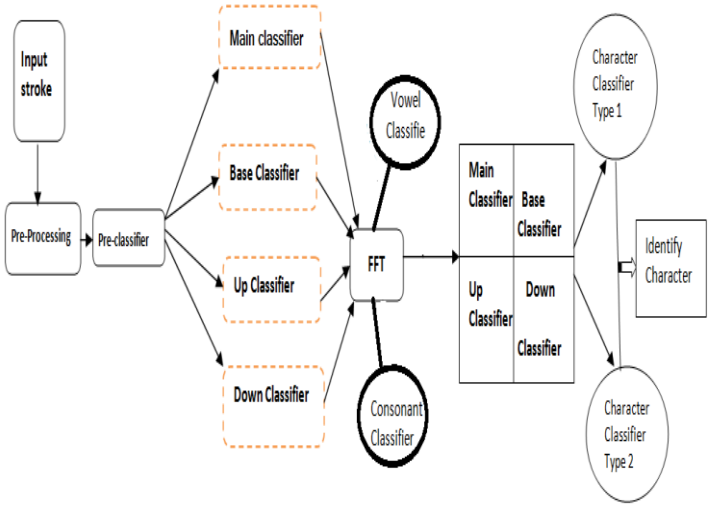

Figure 7: Telugu character identification using stroke model-SVM

5.4 Character Recognition Process:

Here we use two SVM classifiers to recognize the output Telugu character

1. Telugu character Classifier Type 1
2. Telugu character Classifier Type 2

Information obtained in Classifier Type 2 is always relates to the bottom strokes of the Telugu characters.

Main stroke, base stroke, top stroke and bottom stroke neglecting Classifier Type 2 strokes are taken as a feature vector to recognize consonant and vowel character. If a Classifier Type 2 bottom stroke is present, then the corresponding Classifier Type 2

ISSN: 2233-7857 IJFGCN

bottom stroke-based feature vector is constructed and passed to the Classifier Type 2 character classifier.

In cases where there are multiple copies of the stroke, the component is set to the number of copies.

5.5 SVM based Classification Engine

Online and offline Telugu character recognition approaches involve different representations of a hand written character. A proper combination of both approaches may be used to improve the accuracy for Telugu character recognition. In offline Telugu character recognition approach algorithms extract the shape of handwriting from the character image and use these for recognizing a Telugu character. In online Telugu character recognition approach data is converted into an image and is used for classification. To minimize the complexity classification problem is split into multiple networks.

For Telugu character recognition Deep Belief Networks (DBN) and Convolutional Neural Networks (CNN) with feedback learning networks are observed as best to learn features by themselves.

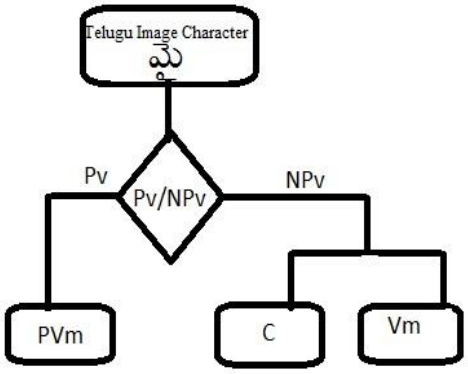

Figure 8. SVM based Classification Engine

5.6. Classification Process:

- First categorize the character as a vowel or a consonant – network (PV)
- If Vowel
- Recognize the vowel – network (PVin)
- If Consonant
- Recognize the consonant – network (C)
- Recognize the vowel modifier for the consonant – network (Vm)

5.7. SVM Model Evaluation:

I compared my proposed SVM stroke recognition approach to similarly old systems developed for online Telugu character recognition and observed my main stroke recognition schema approach achieved higher performance with a overall stroke recognition accuracy of 96.69%.

Result:

ISSN: 2233-7857 IJFGCN

Method	Accuracy
Online Handwritten Character Recognition for Telugu Language Using Support Vector Machines[6]	96.69%

Graphical view:

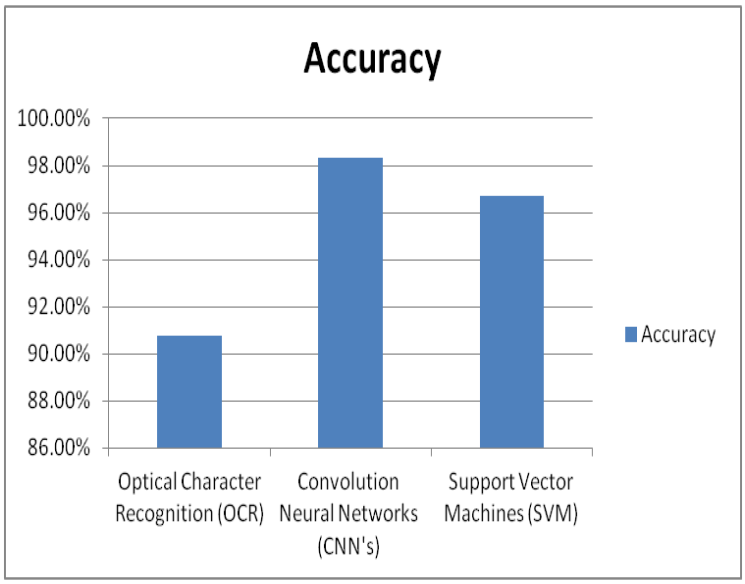

Figure 9: Accuracy result for OCR, CNN, SVM

6. CONCLUSION

Handwritten character recognition is a complex problem because of a variety of character in different languages. Our Research in this direction focused on various machine learning techniques for solving the character recognition problem. By using OCR technique for identifying the Telugu character on real data. I have chosen different character images in such a way that most of the classes are covered of Telugu character images and trained data and achieved with a symbol accuracy rate of 90.78%. Using Support Vector machine presented a schema and performed pre-classification. I compared my proposed SVM approach to similar old systems developed for online Telugu character recognition and observed my main stroke recognition schema approach achieved higher performance with a overall stroke recognition accuracy of 96.69%. Telugu character image data can be classified better by using machine leaning deep convolutional neural networks (CNN). Due to the complexity of Telugu character I found difficult in tuning the algorithm to best fit so I tried different optimizers with different learning rates by changing the number of layers and number of filters and filter size. Finally, I have achieved with an accuracy rate of 98.3%. Finally I can say there is a

ISSN: 2233-7857 IJFGCN

scope to develop an OTHCR system with more accuracy by creating a many training Telugu characters consisting of all font families and of all sizes.

REFERENCES

[1] D Jayaram ,CRK Reddy ,Kamakshi Prasad ,M Swamy Das,," An Overview of Optical Character Recognition Systems Research on Telugu Language " International Journal of Science and Advanced Technology (ISSN 2221-8386) Volume 2 No 9 September 2012

[2] http://en.wikipedia.org/wiki/telugulanguage.

[3] Anuja V.Nair, Bindu.V, "A Review on Indian Sign Language Recognition", International journal of computer applications, Vol. 73, pp: 22, (2013).

[4] Arica, Nafiz; Yarman-Vural Fatos T; An Overview of Character Recognition Focused on Off-Line Handwriting;IEEE Transactions on Systems, Man, and Cybernetics- PartC: Applications and Reviews, Vol. 31, No. 2, may 2001

[5] V. Bansal and R. M. K. Sinha. A devanagari ocr and a brief overview of ocr research for indian scripts. Proceedings of STRANS01, IIT Kanpur, 2001.

[6] Rajkumar.J, Mariraja K., Kanakapriya,K., Nishanthini, S. and Chakravarthy, V.S.,"Two schemas for online character recognition of Telugu script based on Support Vector Machines." In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, pp. 565-570. IEEE Computer Society, 2012

[7] D Jayaram ,CRK Reddy ,Kamakshi Prasad ,M Swamy Das, ," An Overview of Optical Character Recognition Systems Research on Telugu Language " International Journal of Science and Advanced Technology (ISSN 2221-8386) Volume 2 No 9 September 2012.

[8] A. Negi, C. Bhagvati, and B. Krishna. An ocr system for telugu. In Proceedings of International Conference on Document Analysis and Recognition, page 1110. IEEE, 2001.

[9] U. Pal and B. B. Chaudhuri. Indian script character recognition: a survey. Pattern Recognition, 37(9):1887–
1899, 2004.

[10] K. Mohan and C. V. Jawahar. A post-processing scheme for malayalam using statistical sub-character
language models. pages 493–500, 2010

[11] Raju Dara, UrmilaPanduga, "Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine", 2015 International Journal of Computer Applications.

ISSN: 2233-7857 IJFGCN