IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A BENCHMARK IMAGE DATABASE FOR ISOLATED TELUGU HANDWRITTEN CHARACTERS

[1]B. Vijaya Lakshmi, [2]Y.Neelima, [3]Y.Udayani, [4]L.Sri Satya, [5]S.Jyothi Parimala

[1]Asst Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1] Computer Science and Engineering

[1] Gayatri Vidya Parishad College of Engineering For Women, Visakhapatnam, India.

Abstract: In the present work, we present an image database of isolated Handwritten Telugu Characters, used in the standard Telugu literature. There are 18 vowels, 36 consonants in Telugu language. Of which, only around 52 character classes are commonly used. These classes form unique pattern shapes and some of these are often written in multiple styles. Altogether 10,400 isolated character images, belonging to 200 different individuals are collected in our work. The database is divided into training and test sets in 7:3 ratio for each pattern class. A Convolutional Neural Network has been designed and the test set recognition performance is reported using this network. We have achieved a recognition accuracy of 80% on the test database consisting of 52 classes.

Key Words: OCR, Handwritten Character Recognition, Telugu Isolated Character, Image Database, Convolutional Neural Network.

I. INTRODUCTION

Since early days of machine learning research, Optical Character Recognition (OCR) has remained as a popular and well – studied problem domain. Character recognition is also considered as a benchmark problem in pattern recognition research, especially related to the recognition of handwritten text. Handwritten Character Recognition (HCR) is the process to recognize characters by the computer or any other electronic format machine or text based application which is written in our handwriting. Handwritten Character Recognition (HCR) is an inevitable part of Optical Character Recognition (OCR) and as very challenging field of pattern recognition.

Research efforts on Indian script OCR on Oriya, Bangla and Tamil have started to receive attention since last decade. Telugu is a Dravidian language predominantly in the Indian states of Andhra Pradesh and Telangana and the Union Territory of Puducherry (Yanam) by the Telugu people. It stands alongside Hindi and English as one of the few languages with primary official language states in more than one Indian state Telugu is also a linguistic minority in the state of Orissa, Karnataka, Tamil Nadu, Kerala, Chhattisgarh, Jharkhand and Maharashtra. It is one of the six languages designated a 'classical language of India' by the country's government.

Telugu ranks 3rd among the languages with the highest number of native speakers in India and fifteen in the Ethnologue list of most widely-spoken languages worldwide. Telugu is spoken by more than 80 million people worldwide. The Optical Character Recognition (OCR) of the Telugu script has wide ranging applications including education, health-care, administration etc.

The Telugu script is an abugida consisting of 60 symbols-16 vowels, 3 vowel modifiers and 41 consonants. Telugu has a complete set of

[1].

letters that follow a system to express sounds. The Telugu script is written from left to right and consists of sequences of simple and /or complex characters. An abugida or alpha syllabary, is a segmental writing system in which consonant-vowel sequences are written as a unit; each unit is based on a consonant letter, and vowel notation is secondary.

Benchmark databases are essential to compare the performance of different handwritten character recognition techniques over an even test bed. The first attempt to use neural networks for Telugu OCR to our knowledge was in Sukhaswami , Seetharamulu and Pujari(1995). They train multiple neural networks, and pre-classify an input image based on its aspect ratio and feed it to the corresponding network. This reduces the number of classes that each sub-network needs to learn. But this is likely to increase error rate, as failure in pre-classification is not recoverable. Later work on Telugu OCR primarily followed the featurization-classification paradigm. Negi, Bhagvati and Krishna, 2001 provided a combinations like ink-based features with nearest class centroid. Ink-gradients with nearest neighbors by Lakshmi and Patvardhan, 2002. Jawahar. Kumar and Kiran, 2003 provided principal components with support vector machines (SVM) and Pujari et al..2004 also provided wavelet features with Hopfield nets.

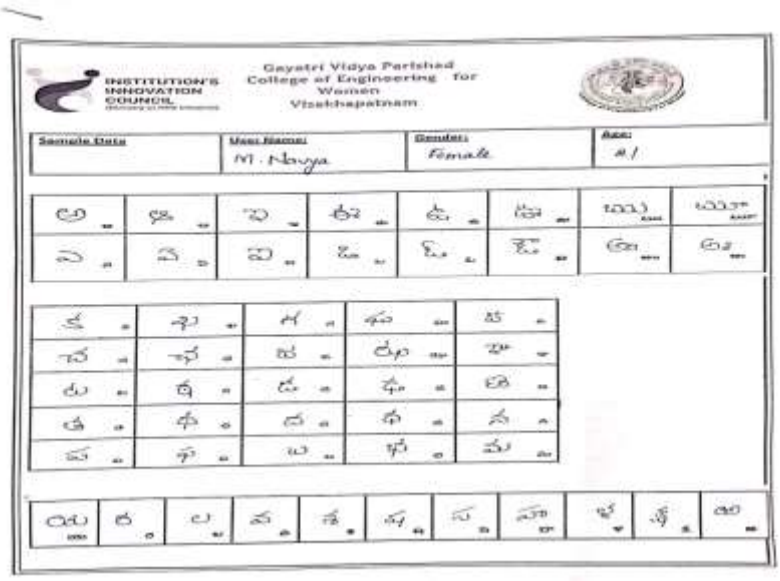


Fig1.1 Handwritten telugu characters

The Fig1.1 shows sample images of handwritten Telugu basic characters.

The more recent work is centred around improving the supporting modules like segmentation, skew-correction and language modeling was provided by Kumar et al..2011.

As mentioned before, the primary objective of the present work is to create a standard database for handwritten Telugu character. We also report a benchmark recognition performance on this newly prepared database. We hope our effort may help the research community to develop a much awaited, complete handwritten Telugu character recognition in near future.

II. DATASET PREPARATION UNDER PRESENT WORK

A. Survey on the isolated Telugu characters

The choice of the dataset is the key for OCR systems. Telugu scripts has 36 consonants and 18 vowels of which 35 consonants and 13 vowels are in regular practice and made available in TIFF files. Telugu handwritten style is an non-cursive and therefore pen-up typically divides the basic graphic symbols although not always. Hence the graphic symbols i .e vowels , consonants, consonant modifiers and diacritical signs are included in the symbol set. Some consonant-vowels are also included which dissembling be easily subdivide. Additionally, the symbol set also comprises certain symbols which do not have a dialectal interpretation, but have an unchanging outline across writers and help lessen the total number of symbols to be collected.

Unfortunately, there are very few works on Telugu character datasets. The work by Pramod et al,. has 1000 words and on an average of 32 images per category. They used the most frequently occurring words in Telugu but were unable to cover all the words in Telugu.

Here, we propose a dataset which takes into consideration all 52 categories of isolated characters and nearly 200 samples per class. All the images are of size 32*32. There are 5371 training samples, 1791 validation samples and 3070 testing samples which add up to of 1 million images (10 GB). Our dataset is novel because unlike other datasets which only take into account the commonly occurring permutations of characters, we have spanned the entire Telugu alphabets.

B. Data collection and Pre-processing

Data collection is one of the difficult task in any pattern recognition research. It becomes even more challenging when the number of classes is too high. The hand written text is usually written in either structured document (like pre-formatted forms) or in unconstrained fashion. In many cases, semi-structured layouts are used for data-entry where words or text lines are written in rectangular boxes.

We have used the CR which is in form of offline process is used in this project for analyses of static documents 52 isolated characters we have taken and each character consisting of 200 samples for exertion. We have taken these characters set from different oldness man and woman. The character set whatever we have taken created dataset by scanning and stored image file format. The Fig2.1 shows the isolated characters which are used for experiment.

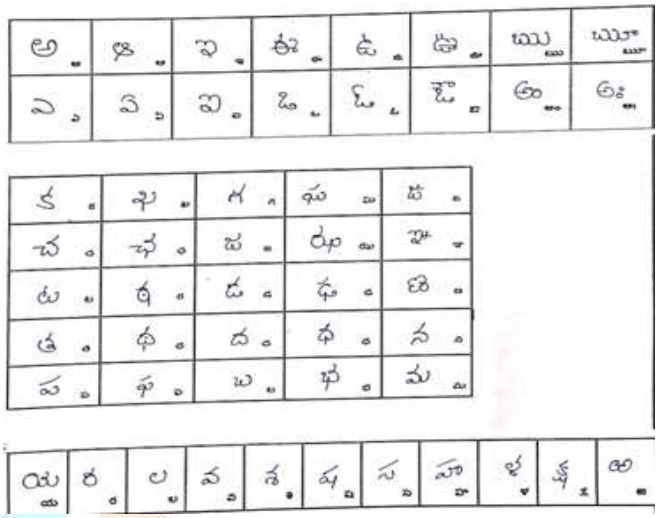


Fig 2.1 isolated telugu characters

The fig2.2 shows the cropped isolated characters using python software, these two hundred images are altered.

Fig 2.2 cropped isolated characters

Algorithm for Telugu handwritten isolated characters recognition:

Step1: Input image
Step2: Scanned the input image
Step3: Pre-processing image optioned
Step4: Image segmentation
Step4: Feature extraction
Step5: isolated character recognition.

Pre-processing: After image acquisition immediate step is pre-processing which is used or image processing. Here processing we will apply noise removal method to remove noise from the image. Image is prepared for next stage processing, if pre-processing is performed well the segmentation and feature extraction process becomes easy. while scanning , the noise may be created on the images, so for good processing pre-processing is required.

Filtering is carried out on binarized image to remove noise present in image. To rejoin the broken character in the image dilation processing is used.

C. Description of the present dataset

In the current work, total two hundred of persons voluntarily participated during the data collection drive. A total two hundred sheets were processed using a fore mentioned methodology to generate the isolated hand written character images. Many hand written samples had to be discarded because of writing errors like use of inconsistent glyphs, overwriting etc. all the extracted characters are then divided randomly into train and test sets. The final training set consisting of 5371 characters, where as the test set consists of 3070 characters.

Table 2.1 less frequently used characters in dataset

Class name	Less frequently used characters
7	
8	
11	
21	
25	
26	
52	

Table 2.1 shows less frequently used characters in telugu language. Most of the telugu words does not include these characters.

Table 2.2 differently written telugu characters

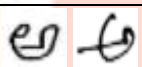
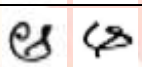
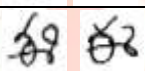
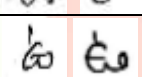
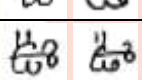
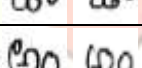
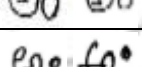
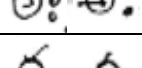
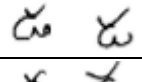
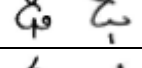
Class name	Differently written characters
1	
2	
4	
5	
6	
15	
16	
28	
29	
30	
32	

Table 2.2 shows the characters in telugu language which can be written in two or more ways. Table 2.2 shows the characters that can be written in any of the shapes specified.

III. PERFORMANCE ANALYSIS ON THE DEVELOPED DATASET

Some research has already been done on recognition of hand written basic characters, vowels and consonants of Telugu scripts. In one of the recent works, Atul Negi et al., proposed an OCR system for Telugu script. The two important features of this system are practicability and simplicity. The connected component algorithm used for segmenting words into recognizable units and the template matching used to recognize the components. Some more research contributions related to this topic are also available in the literature, T. Sitamahalakshmi et al., conducted performance comparison of radial basis function networks (RBFN) and probabilistic neural networks for Telugu character recognition, presented two classification models i.e radial basis function network, probabilistic neural networks . the accuracy of all the classes areas above 90%' in the both methods. Recently, Chandra B V et al., carried out a script independent approach to recognize Telugu and Devanagari hand written numeral characters with probabilistic neural networks (PNN).

Panyam Narahari Sastry et al., done Telugu hand written character recognition using Zoning features, the recognition accuracy was 78%. Chandra Prakash et al., developed OCR for Telugu: Database, Algorithm and designed an android application for ease of use. Prameela et al., proposed a very comprehensive and practical OCR system for Telugu language . this system is shape and font dependent that requires pre-processing and feature extraction. Despite such efforts, their generated a real time implementation of Telugu character recognition using Raspberry Pi but, it requires more number of samples that should be trained. HMM-based online handwriting recognition for Telugu symbols, the accuracy generated for top 1 is 91.6% and for top 5 is 98.7%.

Elastic matching of online hand written Telugu scripts using local features provides accuracy of 90.6%. In one of the earlier works, the technique we use for Telugu hand written isolated character recognition is two dimensional fast fourier transform and support vector machine.Now coming to some other researches, their comes online handwritten character recognition of Devanagari and Telugu characters using support vector machine, the accuracy depends on the stroke recognition. Two schemes for online hand written character recognition of Telugu scripts based on support vector machine is of 96.42%. Recently, pre-processing of HP data set Telugu strokes in online hand written Telugu character recognition, in this characters are written by native Telugu writers. Their generated a hybrid model for recognition of online Telugu hand writing in Indian scripts with accuracy of 93.10%.

IV. PROPOSED SYSTEM

Image Acquisition is input for recognition process. Pre-processing is the most import and first step in the Telugu character recognition. After obtaining input , here we correct the input by binarization, smoothing, filtering, sampling, normalization, which will remove the deficiencies in the input which may have occurred due to the device error or scanning or limitations of the sensor. Segmentation, this is the next step after pre-processing step. Segmentation is the process where we divide image or input or document into sub groups, in other words, we are going to divide a single entity into multiple parts so that it will be easy to recognize the natural hand written word written.

Feature extraction, after the segmentation, we will use feature extraction to extract maximum features of the available raw data. Classification, it is the last and final big step in the handwriting character recognition process is the classification. In this step various techniques or models are used to map the extracted features to different classes and thus identifying the characters or words the features represent the classifier to be used is decided based on various factors taking into consideration the real world problems.

We are using Convolutional Neural Networks (CNNs) for the recognition of handwritten Telugu characters. Convolutional Neural Networks (CNNs) have been shown to be very successful for image recognition tasks over a broad spectrum of problems.CNN's were introduced to exploit the two-dimensional correlation structure of image data. They are a special type Neural Networks that works in the same way of a regular neural network except that it has a convolution layer at the beginning. A Convolutional Neural Network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, RELU layer i.e. activation function, pooling layers, fully connected layers and softmax layers.

IMAGE ACQUISITION
PRE PROCESSING
SEGMENTATION
FEATURE EXTRACTION
CLASSIFICATION
CHARACTER RECOGNITION

Fig 4.1 stages in handwritten character recognition.

EXPLANATION OF KEY FUNCTIONS

Convolutional Layer: In full generality, input to a CNN is a 3D-image with two spatial dimensions and one frequency dimension. For colour images, each sample is a stack of three maps; one each for red, blue and green. Hyper-spectral images could have many more, while gray-scale images have only one map. In our case, the input is a 32 X 32 single-map binary image.

2D-Convolution: A 2D-convolution operation is parameterised by a 2D-kernel of size, k=2l + 1. Typically k is set to three or five (for smaller input images, it could be set to even two). At each pixel in the interior of the input image, this kernel is applied. That is, we take the dot-product of the kernel with the l- neighbourhood of the target pixel. Convolution with a random kernel thus performs a localweighted-averaging over the input image. More useful kernels tend to act as various kinds of edge/eaturedetectors.

Convolution with a 2l + 1 kernel reduces the side of the image by l, which might leads to loss of information at borders. To avoid this, the original image can be zero-padded by l pixels. Fig 4.2 shows a few such kernels applied to a binary image.

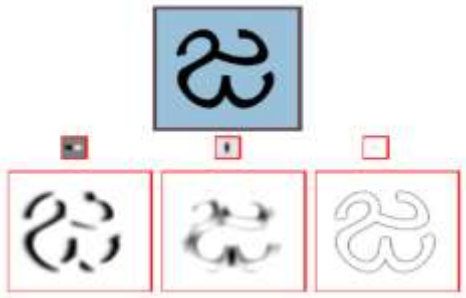


Fig 4.2 kernels applied to binary image

The original binary image of the letter (ja) is shown at the top, the three sample convolution kernels at the center, and the corresponding outputs at the bottom.

Pool Layer: A typical convolution layer outputs many more maps than it takes in. There is also high correlation between the adjacent output values in a map. Hence it makes sense to somehow 'scale' the maps down, generally by a factor of two in each co-ordinate. This operation is called pooling (subsampling).

Typically pooling is done over a 2 X 2 grid. Most of the subsampling on CNN uses Max pooling. The maximum of the four pixels in the grid is extracted as output. Taking maximum gives the neural network translational invariance, which is key to good classification.

Iṭs function is to reduce the spatial size of the representation or reduce the amount of parameters and computation int the network and to also control over fitting. This layer operates independently on every depth slice of the input and resizes it spatially, using the MAX operation.

Max-pooling: A technique used to reduce the dimensions of an image by taking the maximum pixel value of a grid. This also helps reduce overfitting and makes the model more generic. The fig 4.3 show how 2 X 2 max pooling works.

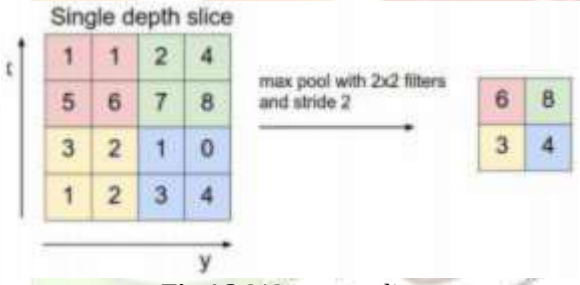


Fig 4.3 2*2 max pooling

Dense Layer: A dense layer is just a regular layer of neurons in a neural network. Each neuron receives input from all the neurons in the previous layer, thus densely connected. The layer has a weight matrix W, a bias vector **b,** and the activations of previous layer **a.** A dense layer represents a matrix vector multiplication. The values in the matrix are trainable parameters which get updated during back propagation.

$u^t .w, w € R^{nxm}$

So you get a **m** dimensional vector as output. A dense layer thus is used to change the dimensions of your vector.

Dropout Layer: A dropout layer is used for regularization where you randomly set some of the dimensions of your input vector to be zero with probability **Keep-prob** at each update during trainng time, which helps prevent overfitting**.** A dropout layer does not have any trainable parameters i.e. nothing gets updated during backward pass od backpropagation. To ensure that expected sum of vectors fed to this layer remains the same if no dropout was applied, the remaining dimensions which are not set to zero are scaled by $1/keep\_prob.$

Fully-connected Layer: CNN's can have anywhere between two to twenty convolutional and pooling layers. Their final output, which is a 3D image, is flattened into a vector. One or more fully-connected layers follow. A fully-connected layer is basically a simple matrix multiplication followed by a non-linearity.

Flatten Layer: Flatten layer converts the data into a 1-dimensional array for inputting it to the next layer. This is called flattening. We flatten the output of the convolutional layers to create a single long feature vector. And it is connected to the final classification model, as shown in fig 4.4, which is called a fully-connected layer.

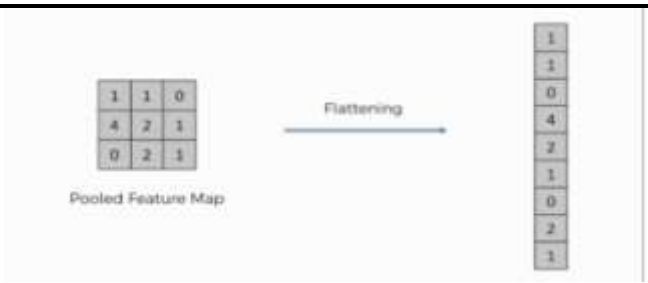


Fig 4.4 flattening model

V. EXPERIMENTAL ANALYSIS

In our process work, we first tried to manually crop the dataset characters. Then we have developed a code to crop the characters. After cropping we have preprocessed the dataset by applying grey scale binarization. Segmentation is done using filter threshold-ostu method. The table 5.1 shows online telugu character recognition using different methods and their accuracies.

Table 5.1 Online Telugu character Recognition methods and their accuracies**.**

S.NO	TITLE	ACCURACY
1	Real Time Implementation of Telugu Character Recognition using Raspberry Pi	more number of samples should be trained
2	HMM-based Online Handwriting Recognition System for Telugu Symbols	For top 1-91.6% For top 5-98.7
3	Elastic matching of online handwritten Tamil and Telugu scripts using local features	90.6%
4	Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine	71%
5	Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines	Dependent On the accuracy of stroke recognition
6	Two schemas for online character recognition of Telugu script based on Support Vector Machines	90.55% 96.42%
7	HAND WRITTEN TELUGU CHARACTER RECOGNITION USING DEEP LEARNING	74%

VI. CONCLUSION

In our process work, we have developed a new benchmark database for telugu hand written recognition. The database is collected from 200 datasheets. The objective is to capture a wide variation of writing styles by individuals and to address a potentially wide range of telugu OCR applications with varying constraints on the writing space. Many hand written samples had to be discarded because of writing errors like use of inconsistent glyphs, over writing etc. All the extracted characters are then divided into train and test sets. The final training set consisting of 5371 characters, where as the test set consists of 3070 characters.

Here we briefly discussed different steps used in HCR development and work done on telugu text recognition. This can be used as a starting point to develop the HCR systems. We have discussed about the most important techniques like HMM and SVM and problems in those techniques and problems in other techniques their accuracies and their recognition methods.

In our work, HCR is implemented using Convolution Neural Networks which is aimed at recognizing the handwritten characters with high accuracy. The input image after preprocessing and segmentation passes through multiple convolution and pool layers and finally through fully connected layer of CNN to produce output.

To recognize online telugu characters many recognition systems are still developing. Recognition of telugu language is a challenging task for the researchers. This provides a comparative analysis on various methods implemented on telugu characters.

Further scope exists to improve the feature set, especially by introduction of global and local features or multi stage recognition strategy.

In a nutshell, the current work is an effort to facilitate researchers in the domain of handwritten OCR to work on large and character database of one of the widely used script of the world.

VII. REFERENCES

[1] Nibaran Das, Kallol Acharya, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, "A benchmark image database of isolated *Bangla* handwritten compound characters" (2014).

[2] C.V.Chakradhar, B.Rajesh, M.Raghavendra Reddy, ASST.Proffessor, G.Pulla Reddy Engg college,Kurnool,Andhra Pradesh,India, "A Study on online Handwritten Telugu Character Recognition" (2016).

[3] Surya Nath R S, Afseena S," Handwritten Character Recognition – A Review" International Journal of Scientific and Research Publications", Volume 5, Issue 3, March 2015

[4] J. Bharathi and P. Chandrasekhar Reddy, "Improvement of Telugu OCR by segmentation of Touching Characters.", International Journal of Research in Engineering and Technology, Vol.3 Issue. 10, 2014

[5] Su, T.-H., Zhang, T.-W., Guan, D.-J., Huang, H.-J.: "Off-line recognition of realistic chinese handwriting using segmentation-free strategy." (2009)

[6] Cheriet, M., El Yacoubi, M., Fujisawa, H., Lopresti, D., Lorette, "G.:Handwriting recognition research: twenty years of achievement · · · and beyond." (2009)