ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 5, Issue 7, July 2016

A Study on online Handwritten Telugu Character Recognition

[1] C.V.Chakradhar, [2] B.Rajesh, [3] M.Raghavendra Reddy

ASST.Professor,G.Pulla Reddy Engg college ,Kurnool,Andhra Pradesh,India.

Abstract—**This paper presents a study and various methods to recognize the Online Handwritten Character Recognition in Telugu Language. Handwritten Character Recognition (HCR) is an inevitable part of Optical Character Recognition (OCR) and a very challenging field of pattern recognition. Various techniques have been proposed for a handwriting recognition system for Telugu. Here we primarily focus to analyse some of the existing systems for Online Handwritten Character recognition of Telugu Scripts their implementation, accuracy and comparisons, overfitting problem and why HMM and SVM are good methods for online handwritten character recognition**

Keywords—**Handwritten Character Recognition (HCR), Optical Character Recognition (OCR). Pattern Recognition (PR).**

I. INTRODUCTION

Handwriting character recognition is the process to recognize characters by the computer or any other electronic format machine or text based application which is written in our own handwriting .It obtains input of our handwriting from any pen-based computer screen surface, special digitizer or PDA or Digimemo or Kinwrite, touch screen devices, paper documents, or image readily available textual image, and converts it into American Standard Code for Information Interchange (ASCII) or other equivalent machine editable form so that the system recognize the characters .[1]

The image of the written text from a piece of paper is retrieved by optical scanning (Optical Character Recognition (OCR)) or intelligent word recognition .Optical character recognition (OCR) is the conversion of handwritten or typed text into an electronic format, which can be stored, interpreted and processed by a computer. It can be used as a direct data input method for a modern computer..

OCR is one of the most challenging and enrapturing areas of pattern recognition with various practical applications. It can contribute immensely to the advancement of an automated process and can improve the interface between man and machine in many applications.

There are many steps involved in the recognition of the handwritten character from input to electronic format and recognition procedure, is shown in the figure brief explanation regarding those phases are explained here

II Stages In Handwritten Character Recognition

Image acquisition that is input for recognition process we have already discussed how the input is taken .

Pre-processing: Pre-processing is the most import and first step in character recognition. After obtainting input. Here we correct the input by Binarization, smoothing, filtering, sampling, normalization, dancing, which will remove the deficiencies in the input which may have occurred due to the device error or scanning or limitations of the sensor. The input of a system may be taken in different environmental conditions. The same object may give different images when taken in different time and conditions. Hence, by doing pre-processing we get data that will be easy for the system to operate on, thereby producing accurate results.

Segmentation: this is the next step after preprocessing step.Segmentation is the process [2] [where we divide the 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 handwritten word written. [3]

Feature extraction: After the segmentation, we will use feature extraction to extract 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.

Classification: 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. Sometimes also a combination of algorithms is used for recognition and it is more effective than using a single classifier.

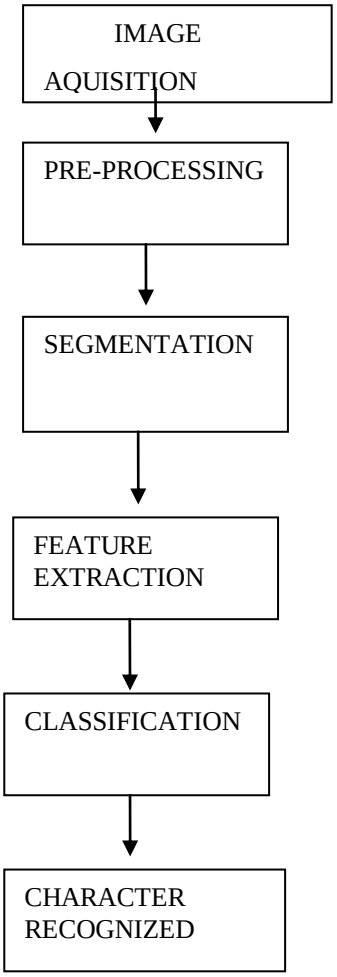

fig 1: Stages In Handwritten Character Recogntion

The accuracy and the probability of the recognized character depends on the sample set we have taken and training of the data so that we can get our desired Recognized Character

III Types of HCR

HCR is divided into two types

1) Offline Handwritten Character Recognition

2) Online. Handwritten Character Recognition [18]

Offline Handwritten character recognition is the process of the recognition of scanned handwritten document or image which was written already written in natural handwriting.

Online handwritten character is a convenient form of input for the person even not aware about technological aspects of about operating device, they simply have to use his own style of writing using any of the pointing object such as special digitizer or PDA or digimemo or Kinwrite 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 object.

Handwritten character written in particular script is classified and stored as Unicode or ASCII format for further processing or into letter codes which are usable within computer and text-processing applications.

If Some of the tools work only for a single user and those are called as writer dependent. Tools working for multiple users are called as writer independent.

Other than basic properties of character certain other properties such as:

(1) Lifting a pointing object is required to write a character

(2) Pointing object went from left to right, right to left, top to bottom or bottom to top – directional information.

(3) Identifying geometrical properties such as contour detection, line detection, loop detection can be done based on pointing object and coordinates.

(4) When the user writes on touch sensitive electronic device many real times Information such as coordinates, pressure given to write character can be obtained.

IV Brief History Of Telugu Language

Online handwriting recognition takes on a novel significance in the context of Indian languages. Here we are going to discuss about a survey on online recognition of Telugu script. [19]

Telugu language is mostly spoken in South India, it's the script is the most complex of all Indian scripts, the two reasons: it has the largest number of vowels and consonants and b) it has Complex Composition Rules The Telugu Script Consists Of 14 Vowels, 36 Consonants, and 3 Special Characters Figure 2, figure 3, and figure 4 shows regarding this information., besides this complication mentioned.

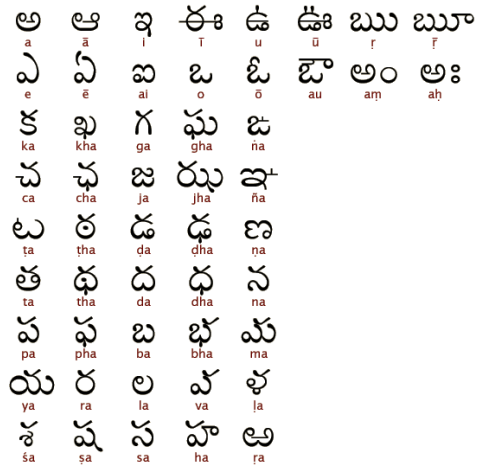

Fig 2:Telugu Letters set

If there is a variation of same character due to the change of fonts and sizes. To get an idea of similar shape of handwritten characters, we provide here the samples handwritten Figure 5 the differences in font types and sizes make the recognition task difficult and resulting the recognition of character process not accurate.

Fig 3: Telugu symbol set

. Here we discuss about the issues and techniques for online handwritten character recognition of Telugu script and research process that are involved up to now.

Fig 4: Telugu alphabets with symbol set

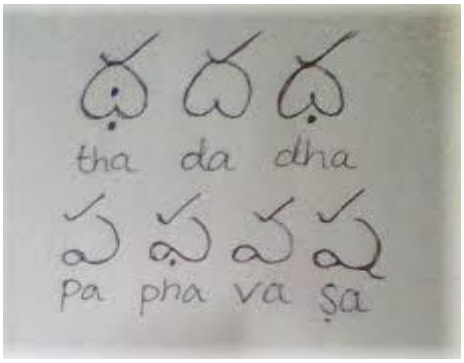

Fig 5: similar characters often difficulty in character recognition

V Online Handwritten Methods

AS discussed earlier a Handwritten character recognition is represented as a sequence of strokes whose features are extracted and classified. Each character has a baseline stroke and usually one or more attached strokes at the top, bottom and side of the base stroke, there can be a lot of size variation

in the strokes so we calculate Stroke density, stroke length and the number of strokes are employed as potential features to characterize the handwritten character.[20]

So as we discussed the after pre-processing step the data what we obtained, will be carried out and we recognize the strokes from left to right ,to down vice versa and further processing steps applied for recognition of steps, it is the basic recognition technique. So many methods were developed and extended along with this stroke recognition to get better results to get better results of character recognition Support vector machines (SVM), Hybrid Model, MLP, Hidden Markov Model (HMM), and EUCLEDIAN etc.

Main stroke, baseline auxiliary, top stroke and bottom stroke are components presented in single Telugu character. By using the strokes of stylus they constructed feature vector and classified with support vector machine.

The main challenge in online handwritten character Recognition for Indian languages is to build a system that is able to distinguish between variation in writing the same stroke (when the same stroke is written by different writers or the same writer at different times) and minor variation in similar characters in the script.

1 ELASTIC MATCHING:

Elastic matching focus on using local features. Dynamic time warping (DTW**)** has been used by four different feature sets: x-y features, shape context (SC), tangent angles (TA) features, generalized shape context feature (GSC) and the fourth set containing x-y, normalized first and second derivatives and curvature features.

Nearest neighbourhood classifier with DTW distance was used as the classifier. Telugu data we obtained an accuracy of 90.6% with a speed of 0.166 symbols/Sec. To increase the speed we have proposed a 2-stage recognition scheme using which we obtained accuracy of 89.77%, but with a speed of 3.977 symbols/Sec:

EM is defined as the optimization problem of two dimensional warping (2DW) which specifies the pixel-to-pixel correspondence between two subjected character image patterns.

The image distance evaluated through 2Dw is called EM distance and invariant to a certain range of geometric deformations.

Thus, by using EM distance as a discriminate function, we can develop recognition systems robust to deformations of handwritten characters.

EM is defined as an optimization problem with respect to a linear or nonlinear pixel-to-pixel mapping called two dimensional warping.

From experiment we noted that the elastic matching recognition without PCA methods often suffers from the mis-recognitions due to over fitting. About a half of mis-recognitions were due to the over fitting. There are mainly two types of over fitting. The first type is the over fitting between two topological similar patterns (e.g., input ”M” → reference ”H”, ”T ” → ”Y ”, ”X” → ”K”, “ ”→” ”, ” ”→” ”). We have mentioned some of these pairs for English as well as for Devanagari characters in figure 5 and 6. The second type is more delicate and local one where a part of the reference pattern is skipped by the non-one-to-one mapping F. This is most common in asymmetric elastic matching where test image is compared with reference image while doing so some of the pixels of reference image are skipped. Suppose if is test image and is a reference image.

When we compare character with some of the pixels of is skipped and is misclassified as. That’s why, since we are getting a very poor recognition rate for asymmetric elastic matching due to high miss-recognition of characters shape.

Over fitting is the common problem that mostly occurs in the handwritten character recognition. The possibility of over fitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model. In particular, a model is typically trained by maximizing its performance on some set of training data. However, its efficacy is determined not by its performance on the training data, but by its ability to perform well on unseen data. Over fitting occurs when a model begins to "memorize" training data rather than "learning" to generalize from trend.

2 HIDDEN MARKOV MODEL (HMM):

A Hidden Markov model is a collection of finite states connected by transitions. Each state is characterized by two sets of probabilities: a transition probability and either a discrete output probability distribution or continuous output probability density function. This gives the condition probability of emitting each output symbol from a finite alphabet or a continuous random vector.

In online handwritten character recognition in Telugu The time domain and frequency domain features are used, although there are many algorithms and methods the reasons for this is explained below.

Hidden Markov Models (HMM) and uses a combination of Time-domain and Frequency-domain features. The system gives top-1 accuracy of 91.6% and top-5 accuracy of 98.7% on a dataset containing 29,158train samples and 9,235 test samples. We also introduce a cost-effective and natural data collection procedure based on ACECAD® Digimemo® and describe its usage in building a Telugu handwriting dataset.

The HMM needs to be trained on a set of seed sequences and generally requires a larger seed than the simple Markov models. The training involves repeated iterations of the Viterbi algorithm which can be quite slow.

The Viterbi algorithm is expensive, both in terms of memory and compute time. For a sequence of length *n*, the dynamic programming for finding the best path through a model with *s* states and *e* edges takes memory proportional to *sn* and time proportional to *en*. For the REP searches, doing a search with a Hidden Markov Hodel is about 10 times slower than using a simple Markov model--for larger HMMs (needed for longer target sequences) the penalty would grow.

Other algorithms for hidden Markov models, such as the forward-backward algorithm, are even more expensive. For a given set of seed sequences, there are many possible HMMs, and choosing one can be difficult. Smaller models are easier to understand, but l arger models can fit the data better. This shows the compression efficiency (in bits per base).

There are some general problems in HMM and they are as follows.

. 1) Requires training using an annotated data

2) Not completely automatic

3) May require manual markup

4) Size of training data may be an issue

For a number of different HMMs compressing the same set of REP sequences. Note that the compression continues to improve with larger models, and so deciding which model to use is somewhat arbitrary.

3 SUPPORT VECTOR MACHINES:

SVM is generalized as kernel machines and are maximum margin methods for classification. SVM is based on finding maximum separability or margin between the different classes with the help of training data features. The main idea during training phase is to find the parameters for the hyper plane, which maximally separates the different classes involved in the problem. The general SVM formulation for linear binary classification is given in1. minw; 1kwk2 subject to dk[wT xk]

Where k = 1, 2, 3…. m. wT x is the equation for hyper plane separating the m input samples, x2R. The decision function for new sample x is evaluated using f (x).

f(x) = sign (wT x) [9]

An SVM based stroke recognition module has been considered because of its generalization capability in large dimensional data. [11]

The principle of an SVM is to map the input data [12] Onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyperplane with a maximum margin between the two classes in the feature space [5]. This results in a nonlinear boundary in the input space. The optimal separating hyperplane can be determined without any computations in the high dimensional feature space by using kernel functions in the input space.

SVMs are trained by quadratic programming (QP), and the training time is generally proportional to the square of the

number of samples. Some fast SVM training algorithms with nearly linear complexity are available.

Both HMM and SVM are the most widely used methods for character recognition. Even in language they were used together for better accuracy of character recognition.

Why HMM and SVM and are more usefull methods and important methods because of their basic functionality like using hidden networks and mapping to high dimensional data respectively the determining of probability of telugu characters or any character can be caluculated so easily so that we can use the OCR techniques to determine the character recognition.

4 MULTILAYER PERCEPTION:

Multilayer Perceptron (MLP) [14] is used to classify different numerals uniquely. In recognition of Handwritten Telugu Character Recognition (HTCR) by MLP. Every multilayer Perceptron network uses two-layer feed forward network with nonlinear sigmoidal functions. We present an automatic HTCR system with MLP [17] classifiers are constructed and then these classifiers are integrated i.e. Hidden layer. Every class has some features and based on these features, classifiers are integrated. For this each MLP classifier is trained using the Error Back Propagation algorithm.

5 RASPBERRY PI:

Character Recognizer (CR) is developed by Embedded handheld device consists of one electronic board which consists of Broadcom BCM 2835 system on chip (SOC) which has ARM11 processor using a Raspbian operating system which is used to recognize individual characters. LCD screen acts as an input and output device. [10]

It is implemented for both writer dependent and independent the detailed description of online and offline recognition has been explained.

CR is implemented on embedded development board by using Raspbian operating systems.

Online system uses the position of the pen as a function of time directly from the interface.

This character recognition tool, is one step forward in the journey to make devices more accessible to people who wish to interact with mobile devices in their local languages. Even letters which, different users write completely different from the other users, are recognized with high efficiency.

Table 1.Online Telugu character Recognition methods and their accuracies.

S.NO	TITLE	ACCURACY
1	Real Time Implementation of Telugu Character Recognition using Raspberry Pi [10]	more number of samples should be trained
2	HMM-based Online Handwriting Recognition System for Telugu Symbols[7]	For top 1-91.6% For top 5-98.7
3	Elastic matching of online handwritten Tamil and Telugu scripts using local features[8]	90.6%
4	Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine[13]	71%.
5	Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines[14]	Dependent On the accuracy of stroke recognition
6	Two schemas for online character recognition of Telugu script based on Support Vector Machines [9]	90.55% 96.42%
7	Online Handwritten Character Recognition for Telugu Language Using Support Vector Machines[6]	96.69%
8	Preprocessing of HP Data Set Telugu strokes in Online	270 samples of each of 166 Telugu

	Handwritten Telugu Character Recognition[13]	"characters" written by native Telugu writers
9	Handwritten Character Recognition for English and Telugu Scripts Using Multi Layer Perceptions (MLP)[14]	Classification and identification. 0.1 Seconds.
10	Improvement in Efficiency of Recognition of Handwritten Telugu Script[5]	Recognition accuracy of Telugu character improves by using more training samples
11	A Hybrid Model for Recognition of Online Handwriting in Indian Scripts[15]	93.10%

VI CONCLUSION

. In this paper, we presented a review of Online Handwritten Character Recognition(OHCR) work done on Telugu language scripts. Here, we briefly discussed different steps used in OHCR development and work done on Telugu text recognition. This can be used as a starting point to develop the OHCR system for Telugu text. During the review of OHCR systems, we have discussed about the most import techniques like HMM and SVM and problems in those techniques and problems in other techniques they accuracies and their recognition methods and their approach. [16]

We have identified that there are no complete, accurate OCR systems because of more similarity of characters and compound characters in telugu language . The major problems in the pre-processing stage and segmentation stage because of overfitting problems. [18] There is a scope to develop an OHCR system with more accuracy on font, size independent and handwritten Telugu text. Figures, Tables and Examples given in this paper are based on the authors' work on developing character recognition systems for Telugu, a south Indian language.

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 online Telugu characters.

REFERENCESES

[1] Surya Nath R S*, Afseena S," Handwritten Character Recognition – A Review" International Journal of Scientific and Research Publications, Volume 5, Issue 3, March 2015 1 ISSN 2250-3153.

[2] 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

[3] M. Swamy Das, C.R.K. Reddy, A. Govardhan , and G. Saikrishna "Segmentation of Overlapping Text lines, Characters in printed Telugu text document images." *International Journal of Engineering Science and Technology*, *2*(11), 6606-6610, 2010

. [4] Rinki Singh and Mandeep Kaur, "OCR for Telugu Script Using Back-Propagation Based Classifier", *International Journal of Information Technology and Knowledge Management*, Volume 2, No. 2, pp. 639-643, 2010.

[5] K. Vijay Kumar and R.Rajeshwara Rao,"Improvement in Efficiency of Recognition of Handwritten Telugu Script*", International Journal of Inventive Engineering and Sciences*, Vol. 2, Issue 1, pp. 1-4, 2013

[6] K. Vijay Kumar, R.Rajeshwara Rao, "Online Handwritten Character Recognition for Telugu Language Using Support Vector Machines", *International Journal of Engineering and Advanced Technology*, Vol.3, Issue.2, pp. 189-192, 2013.

[7] Jagadeesh Babu. V, Prasanth. L Raghunath Sharma. R, Prabhakara Rao G.V., Bharath. A, "HMM-based Online Handwriting Recognition System for Telugu Symbols", In *Document Analysis and Recognition*, ICDAR, Ninth International Conference,Vol.1, pp. 63-67, IEEE, 2007.

[8] Prasanth, L., Jagadeesh Babu, V., Raghunath Sharma, R., Prabhakara Rao, G. V., and Dinesh, M., "Elastic matching of online handwritten Tamil and Telugu scripts using local features." *In Document Analysis and Recognition, ICDAR 2007*, Ninth International Conference, Vol. 2, pp. 1028-1032, IEEE, 2007.

[9] 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.

[10] Naga Deepa. Ch., C. Sri Divya, Dr. N. Balaji, Dr. V. Padmaja "Real Time Implementation of Telugu Character Recognition using Raspberry Pi "International Journal of Science and Research (IJSR) Volume 3 Issue 12, December 2014 .

[11] Raju Dara, Urmila Panduga," Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine" *International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 5, April 2015.*

[12] H. Swethalakshmi, Anitha Jayaraman, V. Srinivasa Chakravarthy, C. Chandra Sekhar Amit Arora and Anoop M. Namboodiri," Online Handwritten

Character Recognition of Devanagari and Telugu Characters using Support Vector Machines", HAL Id: inria-00104402, 6 Oct 2006

[13] srilakshmi inuganti and r. rajeshwara rao," preprocessing of hp data set telugu strokes in online handwritten telugu character recognition", i j c t a, 8(5), 2015, pp. 1939-1945

[14] P.V.Manoj, A.K.Sahoo, Samudra Gupt Maurya, Rohit Kumar "Handwritten Character Recognition for English and Telugu Scripts Using Multi Layer Perceptions (MLP)",International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)Volume No.3 Issue No.6, pp : 730-733 1 June 2014 IJSET@2014 Page 730

[15] Amit Arora and Anoop M. Namboodiri," A Hybrid Model for Recognition of Online Handwriting in Indian Scripts"`

[16] Ch. N. Manisha[1], E. Sreenivasa Reddy[2], Y.K. Sundara Krishna," A Study on Recognition Methods of Telugu Numerals and Characters" *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 11 Issue 5 –NOVEMBER 2014

[17] ARUN K PUJARI,C DHANUNJAYA NAIDU,B C JINAGA,"

An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory, 978-1-4673-5952-8/13/$31.00 ©2013 IEEE

[18] 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

[19] Jomy John, Pramod K. V, Kannan Balakrishnan," Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011

[20] S.V. Rajashekararadhya, and P. Vanaja Ranjan. "Efficient Zone Based Feature Extraction algorithm for handwritten numeral recognition of four popular south Indian Scripts", *Journal of Theoretical and Applied Information Technology*, Vol. 4 Issue 12, pp. 1171-1181, 2008.

Author's Profiles

C.V.Chakradhar working as an Assistant Professor in G.pulla Reddy Engg college and having 1 year of teaching experience and my area of interest are Pattern Recognition, Cryptography and Network security

B.Rajesh working as an Assistant Professor in G.Pulla Reddy Engg college and having 5 years of teaching experience and my area of interest are cloud computing and networks.

M.Raghavendra Reddy working as an Assistant Professor in G.Pulla Reddy Engg college and having 4.8 years of teaching experience and my area of interest is Datawarehousing