THE DIFFERENT PRE-PROCESSING TECHNIQUES USED IN HANDWRITTEN TELUGU CHARACTER RECOGNITION SYSTEM

Neerugatti Varipally Vishwanath[1], G.Manaswithareddy[2], K.Navya Durga Pavani[3], M.Ramya sri[4]
[1]Assi.Professor, Department of ECE, St. Peter's Engineering College, Maisammaguda, Medchal, Hyderabad, Telangana, India
vishwanath@stpetershyd.com
[2,3,4]UG student, Department of ECE, St. Peter's Engineering College, Maisammaguda, Medchal, Hyderabad, Telangana, India

Abstract

In India there are many popular and oldest languages, in that Telugu is one which is used by Telangana, Andhra Pradesh and neighbouring people of the state, in abroad also so many people speak and write Telugu language.There is lot of research required in this area(i.e., offline handwritten character recognition of Telugu).In OCR, pre-processing Technique play very important role for segmenting, Feature selecting, Feature extraction, classification and post processing of characters. The accuracy of the system depends on these processes. So, at pre-processing level much concentration required. This article gives the overview of pre-processing techniques which are useful for researchers and we are discussing in detail about pre-processing techniques. Handwriting character recognition is a popular area in patent recognition. As compared to the online character recognition, offline character recognition is now also demanding job due to following reasons; writing of styles individual people, writing speed, letters size, and letters over lapping and wrier physical and mental conditions. so this article gives the basic methods used in pre-processing level for researchers.

Keywords: Pre-processing technique, offline Telugu Hand-writer characters.

I. Introduction

Optical Character Recognition (**OCR**) is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. There is plenty of research work required in Telugu OCR system. The Telugu is one of the official communications in the state of Telangana and Andhra Pradesh in India.The million people speak Telugu in Telangana, Andhra Pradesh and neighbouring states.Telugu is treated as a regional script.An attempt is made in this article to discuss the important results.With the influence of Christian missionaries written carry Telugu script was standardized at the beginning of the 19th century.The script is syllabic in nature .In the scheme proposed by Sitamahalakshmiot. Al (2010) for the recognition of handwritten Telugu characters, the probability of identifying the given input character was obtained using five distance measurement methods. The result obtained is then combined using the Dempster-shafer theory(DST). [1]One of the challenging works in Telugu scripts is this unconstrained handwritten word recognition because of the segmentation problem due to its position of compound characters and modifiers.To the best of our knowledge, no work has been reported so far toward the recognition of handwritten compound characters handwritten words in Telugu script.Researchers should works extensively on this problem in the near future.Sizes of some of the database for Telugu are not very large and these are a need to develop large data base for this script.The 3rd important popular script in India is Telugu which id Dravidian language.The Telugu language consists of sixteen vowels and thirty-five consonants that can combine to form more than five thousand compound characters.OCR is a important sub area of patent recognition which means handwritten characters are identified modified into machine readable text. The figure 1 gives step by step process in Telugu OCR.

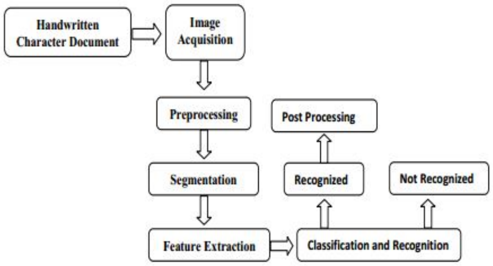

Figure 1: Planning of a offline handwritten character recognition system.

Application of OCR is data entry for business documents, e.g. check, passport, invoice, bank statement and receipt. Automatic number plate recognition, Automatic insurance documents key information extraction, Extracting business card information into a contact list, More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg, Make electronic images of printed documents searchable, e.g. Google Books, Converting handwriting in real time to control a computer (pen computing).

II. Literature survey

Dhardra and Hangrage used nearest neighbour and KNN algorithm to classify word images belonging to Kannada and Telugu scriptsTelugu Data base available at ISI ,Kolkata is 10,870 characters.In IIIT and University of Hyderabad, The major research centre in India HWTCRPal [2] proposed a quadratic classified based scheme for the recognition of the handwritten charactersVasantha lakshmiet.al[3,4,5,]have reported the development of a Telugu OCR system for printed text based on identify from a symbols.Sastry et al[6] proposed a method for classification of Telugu characters extracted from the palm leaves , having DaTA Since thirty years onwards research going on Telugu OCR. The first reported Research on Telugu OCR was completed by Rajasekharan and Prekshafater[7]Sukhaswamyetal[8] provided a neural network based system to find Telugu script .Recently , Negi et .al [9] proposed robust Telugu OCR,has used Template matching using fringe distance metric. Figure 2 is Telugu character set.

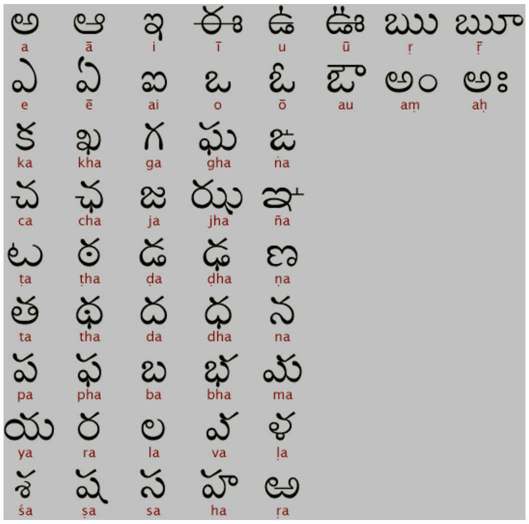
అ a
ఆ ā
ఇ i
ఈ ī
ఉ u
ఊ ū
ఋ ṛ
ౠ ṝ
ఎ e
ఏ ē
ఐ ai
ఒ o
ఓ ō
ఔ au
అం aṃ
అః aḥ
క ka
ఖ kha
గ ga
ఘ gha
ఙ ṅa
చ ca
ఛ cha
జ ja
ఝ jha
ఞ ña
ట ṭa
ఠ ṭha
డ ḍa
ఢ ḍha
ణ ṇa
త ta
థ tha
ద da
ధ dha
న na
ప pa
ఫ pha
బ ba
భ bha
మ ma
య ya
ర ra
ల la
వ va
ళ ḷa
శ śa
ష ṣa
స sa
హ ha
ఱ ṟa

Figure 2.complete telugu character set

OCR reduces time for processing for processing data from large number of forms. If done manually, may lead to human error and takes up much of the time.Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text.Higher rates of recognition of general cursive script will likely not be possible without the use of contextual or grammatical information.Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text. The figure 3 specifies the evolution of Indian scripts. Higher rates of recognition of general cursive script will likely not be possible without the use of contextual or grammatical information.[redacted]

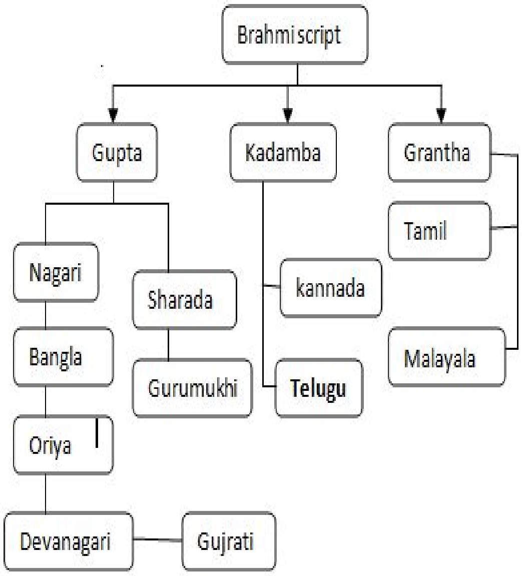

Figure 3.Evolution of various scripts

III. METHODOLOGY

Telugu is one among the prehistoric languages of India. The essential alphabet set of Telugu consist of sixteen vowels and thirty-six consonants. The Telugu language consists of straightforward and compound character shaped from basic alphabet set. Some characters in Telugu square measure created quite one connected symbol. Compound characters' square measure shaped by associate modifiers with consonants, of import in an exceedingly vast variety of attainable mixtures square measure there in an exceedingly Telugu script. Telugu includes a varied writing with a sizable amount of distinct character shapes composed of straightforward and compound characters evolved alphabet set. Telugu could be a phonetic language and written from left to right like West Germanic and additionally, in the Telugu language, every and each character represents a linguistic unit. Not a lot of work has been according to on the event of OCR systems for Telugu. Therefore, development of associated OCR system for Telugu is a very important space of current analysis. The identification method is extremely tough for Telugu as a result of it consists massive and numerous teams. In offline Telugu handwritten character recognition system pre-processing very important step.The OCR systemsdeal with improving quality of the Image for better recognition by the system. OCR software often "pre-processes" images to improve the chances of successful recognition.Techniques include:Line and Word Detection, Script Recognition, Segmentation, Normalize Aspect Ratio and Scale, De-Skew, DE speckle, Binarization ,Line Removal, and Zoning.

Figure 4. Is pre-processing step by step process:

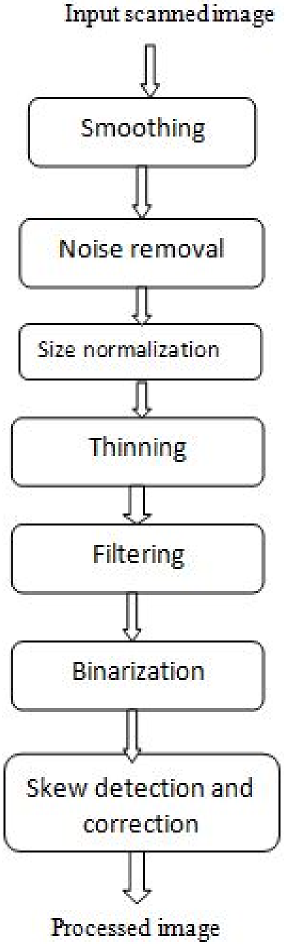

Figure 4. Pre processing steps

A.SMOOTHING

It is a process where noise within the image is reduced (or) developing a less pixel rated image. Useful for reducing noise and unimportant details. The popular method used for smoothing are low pass Filters. The smoothing operation mainly applied to images for clearing unnecessary pixels. To prepare image data set for experiments, first images should be smoothened.

B. SIZE NORMALIZATION

Normalization is the process of converting a random sized image into a standard size. It is also defined as process that changes the range of pixel intensity values. To bring all characters into a common size platform in order to extract features on the same footing, a minimum bounding box is fitted to the character and the element is cropped and then resized to fit into 32x64 windows.

C.THINNING

It is a process where the range of pixel intensity values are changed to improve the poor contrast due to glare, simple operation is applied on

Image that is [10] thinning. It is operation performed on digital image which simplified morphological operation like dilation is used for thinning. The thinning operation as shown in figure 5.

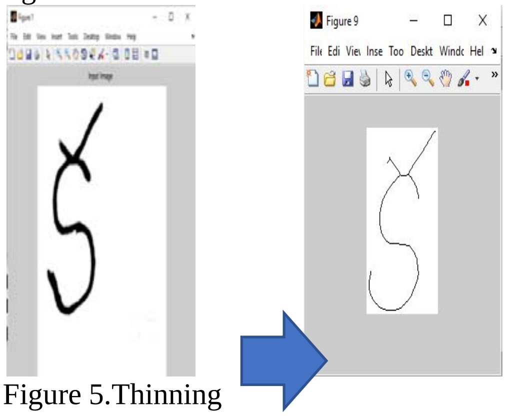

Figure 5.Thinning

D.FILTERING

It is a process where an image is enhanced or modified to obtain exact features simply emphasizes certain feature or removes other features [10]. Sharpening method and smoothing methods more popular in the research. Figure 6 is Filtered output:

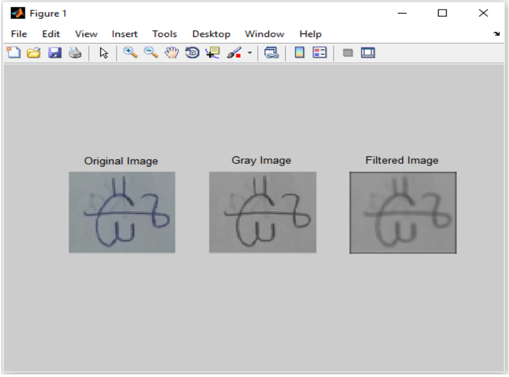

Figure 6.filtered output.

E.BINARIZATION

It is a process where analog image is converted into Digital image which is obtained by scanning a document and stored as an electronic representation of the original in the form of Tiff or BMP or Jpeg. Binarization plays very important role in pre-processing. Thresholding is process where foreground is obtained from the background is called thresholding. Binarization is performed in the pre-processing stage for document analysis. Its aim is that a picture containing text can be edited. Image binarization is the process of separation of pixel values into two groups white as background and black as foreground. Threshold plays a major role in binarization of images. The images which are scanned copies of these degraded documents are provided as an input to the system. Thresholding plays major role in binarization of images. No single method can work for all types of images. As shown in figure 7. locally adapted binarization[11] and heuristic algorithm[12] compared to are suitable for binarization Othsu method are suitable for binarization.

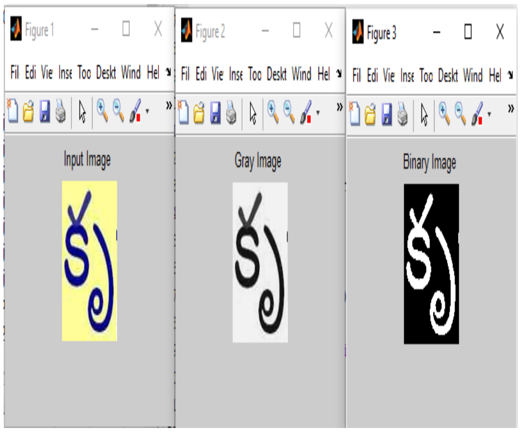

Figure 7.Binarization

F.NOISE REMOVAL AND NOISE DELETION

Many handwriting documents will have salt pepper noise removal of noise discussed by Nair [10]Main by 2 peaks comprise the histogram gray scale values of a document image: first one is a high peak analogous to the background and the smaller peak corresponding to the foreground. Putting the threshold value is finding the one of the optimal value the peak gray scale values. [13] each value of the threshold is tried and the one that maximize the criterionchosen from the class as the foreground and background outputs given in the figure 8.

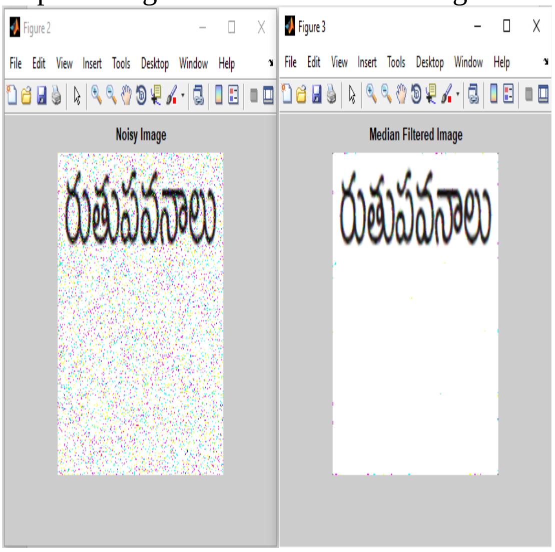

Figure 8.Noise Removal

Accuracy or efficiency of the character recognition system decreases if noise is present in image. Noise may have occurred while scanning or poor quality of the document, but it has to be removed for further processing the best methods for noise removal in an image are wiener filter method and median filter method [14][15].

G.SKEW CORRECTION

It is process where the paper document aligning is done accordance with coordinate system there are many methods for skew correction, they are projection method, profile method, correlation method. Hough transform method so on. Hough transform method is so popular to calculate skew angle.as shown in figure 9.

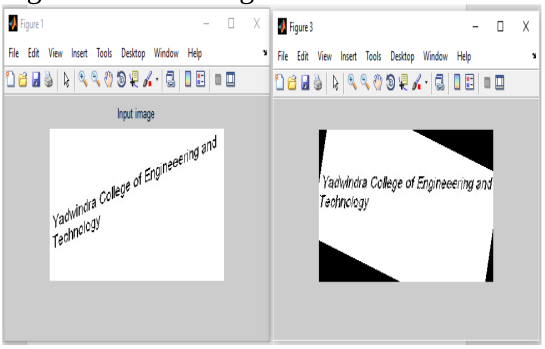
Figure 1
File Edit View Insert Tools Desktop Window Help
Input image
Yadwindra College of Engineeering and Technology
Figure 3
File Edit View Insert Tools Desktop Window Help
Yadwindra College of Engineeering and Technology

FIGURE 9.Skew detection and correction

IV. Conclusion

In this article, offline handwritten character recognition system pre-processing methods for Telugu discussed in detail.The different pre-processing technique like, binarization, skew detection, noise removal, size normalization, smoothing, thinning, filtering are suggested.The difficulties in this research area is the not availability of data base for handwritten characters of Telugu language. Further we will try to develop data basefor Telugu hand written character recognition. This article will be helpful for researchers to use particular pre-processing method.

REFERENCES

[1] Sitamahalakshmi,T.,Debu, U., and Jagadeesh M. 2010. "Character recognition using Demprtershafter theory combining different distance measurement methods",Int. J.enginSciTechnol 2,5;1177-1184.

[2] U.PalN.Sharma, T.wakabayashi and F.Kimura, "Handwritten character recognition of popular south Indian script", proceeding &ACH'06 proceeding of the 2006 conference on Arabic and Chinese handwriting recognition springer-verlag Berlin, Heidelberg.

[3]C.VasanthaLakshmi ,C.Patvardhan Ranjit Singh "A novel basic symbols approach for Telugu OCR with neural networks" Journal of the computer society of India, march 2003, pp31-39.

[4] C.VasanthaLakshmi ,C.Patvardhan "Recognition of basic symbols in Telugu by neural network" STRANS-2002 march15-17,2002, IIT Kanpur, Kanpur.

[5] C.VasanthaLakshmi,C.Patvardhan "An OCR international journal on pattern analysis and applications" July 2004,pp 190-204.

[6]Sastry, P.N Krishna R., and Ram B.V.S, "classification and identification of Telugu handwritten characters extracted from palm leaves using decision Tree approach", J. Applied Engn.Sci b, 3 ,2010,pp2L-3L.

[7]S.N.S Raja sekharan, and B.C. Deekshatulu "Generation and recognition of printed Telugu characters" computer graphics and Image Processing vol 6, pp335-360,1977.

[8] M.B.Sukhaswami, P.Seetharamulu, A.K Pujari "Recognition of Telugu characters using neural networks" Int-journal of neural system vol.6 No. 3 pp 317-357,1995.

[9]Atul Negi, N. Shankar and C. Chandrakant, "localization, extraction and recognition of text in Telugu Documents Image, processing" of seventh ICDAR, IEEE comp. society press 2003.

[10] Rafael Gonzalez,Richard Eugene woods (2007). Digital Image processing,prentice hall p.85 ISBN 0-13-168728-X.

[11] Øivind Due Trier and Torfinn Taxt "Evaluation of Binarization Methods for Document Images"IEEE,1995

[12] George D. C. Cavalcanti, Eduardo F A. Silva, Cleber Zanchettin , Byron L. D. Bezerral ,Rodrigo C. Do'rial and Juliano C. B. Rabelo "A HEURISTIC BINARIZATION ALGORITHM FOR DOCUMENTS WITH COMPLEX BACKGROUND". ICIP 2006.

[13] Nair ,A.S.., Ravathy K., and Tatavarti R Remael of "salt pepper noise in images: a new decision based algorithm", process of the int. maulticont of engineers and comp scientiests, march 2008 vol.1.

[14] .K.khatatnah, "probabilistic artificialneural network for recognition the ArabicHandwritten characters "Journal of comp Science 3(12) 881-886,2006.

[15] R.Plamondon and Srihari. Online and offline Handwriting recognition:A comprehensive survey. IEEE Trans.on pattern analysis and machine intelligent 22(1):63-84, 2000