A Survey On OCR For Telugu Language

Buddaraju Revathi, G.Naveen Kishore, V Dheeraj

Abstract: Text in the image file will not be in editable format on computer. Optical Character Recognition (OCR) is the process to understand the text in the image, either printed or handwritten and creates a file with the text in the image file that can be editable on the computer. OCR for English language is well developed. At present day there is a need of OCR for Indian languages to preserve historical documents which are written mostly in Indian languages, to organize books in library and for application form processing etc. OCR for Telugu language is difficult as a consonant or single vowel forms a single character or it can be a combination of vowels and consonants that can form a compound character. This paper presents survey on methodologies used in OCR system for Telugu Language till now.

Index Terms: OCR-Optical Character Recognition, SVM- Support Vector Machine, PCA- Principal Component Analysis, KNN -k-Nearest Neighbors QDA- Quadratic Discriminant Analysis, CNN- Convolutional Neural Network, RNN- Recurrent Neural Network.

1. INTRODUCTION

Telugu is a popular and primarily spoken language in Andhra and Telangana. In Telugu every word ends with a vowel. It mainly consists of Sanskrit words. Telugu has 56 characters. Each character is a mixture of consonants and vowels which represents syllables. When the scanned document is of poor quality or if the script has different styles then most of the OCR systems will fail. OCR system employs the following processes

1) Image file loading.
2) Quality enhancement of image which involves the processes like removing noise and skew correction.
3) Removing lines especially for detecting text inside tables.
4) Detection of text position and spaces.
5) Identifying words.
6) Broken character correction.
7) Character recognition.
8) File saving.

OCR for English language is well developed, but OCR for South Indian languages is in developing stage due to their complexity in recognition. Telugu is one among the most highly spoken languages in South India. Developing an OCR for Telugu language is needed to preserve historical documents, automatic sorting of books in library, postal services etc. When compared to printed text, handling handwritten Telugu characters is very difficult due to gunintham and vattus in handwritten fonts are difficult to identify. To completely automate the banking services, postal services and tax forms, OCR for handwritten languages must be developed. We have mainly focused on OCR for Telugu language and written a review of techniques developed till now for handling printed Telugu text, Handwritten Telugu text, printed or handwritten Telugu characters and numerals.

- Buddaraju Revathi, ,Research Scholar, ,ECE Department, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
- G.Naveen Kishore, ,Associate Professor, ECE Department, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
- V Dheeraj, student, CSE Department,,Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502

2 STUDY ON TELUGU OCR

For Telugu text in printed form, Arun K Pujari et al. [1] in 2002 proposed an OCR system. Text is scanned in the form of gray scale image. Horizontal and vertical projection techniques are used for line and word segmentation. Zero padding technique is used to convert characters into fixed size. Wavelet analysis is used for obtaining information of images at different scales like 32x32. Performed 2-dimensional filtering so that 32x32 image is converted into 4, 8x8 images, which gives average image. Then by using thresholding, convert images to binary which gives 64 bits and these are referred to as signature of input symbol. For recognizing symbols Dynamic Neural Network is used in which every node in the network is Hopfield network. This method does not depend on font and shape. Some symbols dha, dhaa, na and sa are not correctly recognized by this technique. C. Vasantha Lakshmi et al. [2] proposed an OCR system in 2003 for printed text which is in Telugu. Scanned image is converted to binary scale and noise is removed through rectification. Skew is corrected and then lines, words and symbols are extracted from text segmentation. Pre-Classification of each symbol by size property to compute direction features which are real valued. Neural recognizers are used for classification and finally information associations of basic symbols for a word are outputted. Testing is performed on one lakh symbols which resulted in 99% accuracy for DeskJet prints and laser prints using additional logic. OCR system for printed characters in Telugu was proposed by Negi, et al. [3] in 2003. Non Linear normalization is performed using modified crossing count, which enhances the features of input image. In different zones, pixel densities are used for searching initial candidate of input glyph. If the candidates are found in-conclusive, they are passed through another stage where input image cavities are analyzed. Template matching is done on the basis of Euclidean distance on normalized characters for nonlinear shapes which are controllable. This technique obtained correct results for 1463 glyphs out of 1500 glyphs which are collected from magazine. This method needs improvisation while dealing with different fonts. C. V. Jawahar et al. [4] proposed an OCR for recognition of Telugu printed characters in 2003. The document which is scanned is filtered, converted to binary and skew is corrected by employing projection profiles. For the text blocks which are extracted from the pages, word and line segmentations are performed. Sirorekha was employed to differentiate between Hindi and Telugu languages. Removal of Sirorekha must be done in order to perform segmentation for

Hindi, while for Telugu, the components that are connected has to be separated. All the components needed to be scaled to fixed scale to do feature extraction. To enhance results entire image has to be used as feature vector. To reduce the dimensionality of the feature vector Principal components are used, which is not dependent on font and can be used for different languages and may be applied on handwritten documents. SVM is used for classification. They have achieved accuracies of 97.5% and 96.8% for Telugu and Hindi. For further improvement, geometric features of characters need to be considered. For printed text in Telugu, C. Vasantha Lakshmi et al. [5] in 2006 developed an OCR system. To avoid confusion between similar symbols they have used the features of edge histogram and in addition confusion table, instead of matching the words with dictionary which increases computational complexity. Thresholding is used for conversion of gray scale image to binary image. Modified Hough Transforms are used for Skew detection as well as for skew correction. To segment the text in the image into words and lines, Profiling was used. Recognizer makes use of preliminary classification scheme and a classifier scheme called Nearest Neighbour (NN) for identification of symbol. Due to scanning noise or paper defects, symbols looking similar may be incorrectly recognized, so confusion tables are used to avoid false interpretations. It is able to recognize all sizes of font and recognition of all fonts is 98.5%. For the symbols which are confusing, the logic for correction is executed only when the symbols are recognized by the recognizer to be present in the set. For printed Telugu characters, Rinki Singh et al. [6] proposed an OCR in 2010.The process is done in 3 phases. Phase 1: Firstly train the system by collecting data, then scan the documents. Detect the line. By considering the vertical gaps words are separated, later characters are separated from words and then extract the feature vector for each character. Phase-2: Using Adaptive sampling, each character is scaled to fixed size and text image is converted to binary by Otsu's threshold algorithm. Hilditch algorithm is used to obtain image matrix from boarder pixelsPhase-3: Back propagation based classifier is used in which artificial neural network is trained by using supervised learning for classification. It was designed to handle standardized mass documents. An OCR system was proposed by P. Pavan Kumar et al. [7] in 2011 for Telugu text in printed form. In order to handle the broken characters, classifier uses distance measure by employing feedback and character segmentation takes the advantage of orthographic properties in Telugu Language to improve OCR system accuracy. Adaptive binarization and skew detection mechanisms are used. Based on the projection profiles line segmentation, word segmentation and character segmentation are performed. Neural Networks, SVM classifiers can be used for classification. A test was conducted on 969 pages which yielded satisfactory results. To recognize the characters in Telugu and Kannada which are handwritten Dhandra B.V et al. [8] proposed an OCR system in 2011. Dataset sets are not adequately available for south Indian languages, so they had developed their own dataset set by collecting documents from schools, lawyers etc. Otus's method was used for binarization. By the operation of morphological opening, noise is removed. After segmenting manually the digits in images, they are converted into 32x32 pixels. For obtaining features, normalized image will be used by dividing it into 8x8 sizes to obtain 64 zones. For each zone by calculating the densities of pixels features are recognized. Classification is done through SVM and KNN classifiers. Recognition accuracies achieved for bilingual samples are 96.18%, 97.81%. Improvising of classifier is needed to further improve the rates of recognition. For indentifying the size and font for printed Telugu text, Ram Mohan Rao, et al. [9] in 2013 proposed an OCR system. First, image is converted into binary from gray scale, later image is cleared. Add 20 rows on top and bottom which are white pixels and 20 columns left and right which are also white pixels. Line segmentation is done by calculating horizontal profile of each row. Calculate the horizontal profiles for the head line, top line, base line and bottom line. Connected components are separated. Each Component is classified as core and non-core, based on Zonal information to identify tick mark on the character among non core components. Calculate pixel ratio, aspect ratio of components to identify size and font by comparing the above two ratios. It is applicable for only sizes of 19, 16, 14 and for fonts like Priyanka, Brahma, Anupama , Pallavi. For Telugu numeral printed characters Ramalingeswara Rao K V et al. [10], proposed an OCR in 2014. This paper focuses mainly on feature extraction. For a given character the topological features are extracted based on counting the Holes (Number of closed regions) after selective morphological unification. The character images are exposed to morphological unification by joining (selective bridging) to generate unique and distinguishable features. Character images are subjected to different types of unifications to follow the attributes that are specific to target characters. Depending on the requirements, increase the type and number of morphological unifications. This technique is implemented only for numeral characters. This method is neither for complete sentences nor for non numeral Telugu characters and doesn't consider broken characters into account. For Handwritten English characters and Telugu characters, P.V.Manoj et al. [11] in 2014 developed an OCR system. Initially, description of input page which is later followed by preprocessing stage. Data set is generated. For obtaining feature vector, measure the distribution of pixel density in various zones and compare it with training set which is pre-computed and placed in code book to obtain neighbors which are nearer. In the phase of image acquisition the coordinates of pixel are identified which are measured using Digital Measuroscope. Accuracy of recognition is in between 85% to 95%. Data collection should be automated mostly regarding depth informationFor recognition of Telugu characters in handwritten style Panyam Narahari Sastry et al. [12] proposed an OCR in 2014. Binarize the image using thresholding with a value of 0.7 which is optimum. Normalize each character into a fixed size image. Divide each image into 100 zones of fixed size 5x5. Feature vector is obtained by adding all the pixels of zone. Later add all zones and concatenate one below the other. So the feature vector has 100 rows for the column vector. Find the Euclidean distance and select the one with minimum distance between training image and test image i.e., Nearest neighborhood classifier is used. Recognition accuracy is 78%. For Telugu printed characters J. Jyothi et al. [13] proposed an OCR system in 2015. In this characters which are segmented will be converted into corresponding character codes. Based on Singular Value Decomposition (SVD), Projection Profile (PP) and Discrete Wavelet Transform (DWT) features are extracted and are evaluated with SVM and k-Nearest Neighbor Classifier This method should be improved for working on all fonts. N. Shoba Rani et al. [14] proposed an OCR for Telugu

characters. Segmented characters from the scanned image are given as input. After preprocessing, feature computation is done by dividing each character into 9 zones. Compute Hu-moments as well as Statistical Features through which we obtain 81 features for every character. SVM and KNN classifiers are used for indentifying character. Accuracy of KNN classifier is 80% and SVM accuracy on test image is 76.47%. It is applicable for characters only. Dr. B. Rama et al. [15] proposed an OCR for both printed and handwritten in 2016. Based on Quality ration, process the image. Characters from image are extracted based on quality, base ration, print ration, character thickness and alignment are checked. Characters are classified as normal consonants and conjunct consonants. The first and third layers are conjunct consonants and second layer is normal consonants. Process of recognition starts from second layer and then it checks first and third layer. Second layer detects base character and first and third layers detect symbolic characters. The quadrants location in the layer and three layer differentiation of quadrants method for conjunct consonants and consonants are inputs to fuzzy logic controller. Performance is good for low quality written text in the case of mixed consonant text. For Telugu handwritten characters N Prameela et al. [16] in 2017 proposed an OCR for Offline recognition. Preprocessing of image is done with the help of median filter. Pixel points which are at the edges of boundaries are obtained by performing skeletonization and normalization on the input characters. Divide each character into 3X3 grids. For the 9 zones evaluate the centroid to identify style of characters. Draw binary external symmetry axis for the character which is unconstrained to calculate vertical and horizontal Euclidean distance from centroid of every zone to the same pixel which is nearest. Calculate Euclidean distance, mean and the zone mean angular values. For recognition of characters, SVM and quadratic discriminate classifiers are employed. Recognition accuracy achieved by employing SVM classifier is 80.6% and QDA classifier achieves 87.6% accuracy. This method needs improvement in preprocessing stage and feature extraction stage. Dependency of the technique on shape and font is major disadvantage. For standard recognition of Scene images which in text telugu or Malayalam text, Minesh Mathew et al. [17] proposed an OCR in 2017. This method can be used to recognize cropped images that consist of words. For transcription of word in the images to text a hybrid CNN-RNN is used, which initially consists of convolutional layers for generation of features vectors in column wise. The output from convolutional layers is passed through recurrent layers for making predictions by using deep BLSTM nets. Finally these predictions are passed through the transcription layers which produce label sequences based upon the predictions of RNN layers. New dataset is created for training by collection of images from various boarding's in markets, banners, traffic signs etc. By using RNN, Word Recognition Rate (WRR) for Telugu is 33.6%, while the Character Recognition Rate is 61%. Hybrid CNN-RNN achieved WRR of 57.2% and CRR of 86.2% for 1211 input images

CRR=(C1-C2)/C1

C1-Number of caharacters

C2-∑LevenshteinDistance(RT,GT)

Where RT stands for Recognized Text and GT stands for Ground Truth

WRR = Number of words Recognized correctly/total number of words

Fine tuning should be incorporated to improve the performance. For Telugu printed text Devarapalli Koteswara Rao et al. [18] proposed an OCR in 2017. For binarization, Otsu's method is used and for segmentation, projection techniques i.e., both horizontal and vertical are used. To find the zones, method based on fringe map is used depending on orthographic rules. Sliding window mechanism is used for obtaining the feature extraction which is matched with akshara HMM sequences based on bi-grams of akshara with the help of HVite tool to match sequences of feature vector. The word error rate is 26% and character error rate is 15%.For Telugu Handwritten characters Neerugatti Varipally Vishwanath et al. [19] in 2018 proposed an OCR. Using MATLAB command rgb2gray the image is converted into gray scale from RGB. Thresholding is used to separate character from background. Noise removal is done through filtering. To handle broken characters, image dilation process is used. Techniques used in image processing such as image filling, Blobs analysis approaches are used. For detecting discontinuities in gray level, Edge detection is used. Segmentation is performed through the command in MATLAB. Calculated confusion matrix and used morphological features to recognize character. They have collected 250 samples for one character. This method is applicable for characters not for words. Accuracy over tested samples is 98.1%For Telugu printed text, Kesana Mohana Lakshmi et al. [20] proposed an OCR system in 2011.This paper mainly focuses on feature extraction. By making the image consisting of word into16x16 patches, kernel features which are of higher dimension are operated by a fix. For recognition AKD algorithm is performed on the patches and SVM classifier is employed. Improvisation is needed in the algorithm for handling different scales of text images and needs comparison of this method on large databases. For Telugu printed text Konkimalla Chandra Prakash et al. [21] proposed an OCR in 2011. Skew correction is done by using Hough transform for straight line. For binarization modified Otsu's Threshold is used. Morphological closing algorithm is used for noise removal. Later perform logical OR operation between thresholding result and denoised image. Further thresholding based on mode is applied. Modified MSER is used for obtaining vattus and dheergas. Connected components algorithm is used for segmentation at character level. CNN are used for classification of characters. Improvement in segmentation is needed so that each character is segmented along with vattu and gunintham

3 CONCLUSION

OCR for Telugu Language is the current area of research due to its enormous applications. OCR for printed text is developing now but needs improvisation in processing stage, as well as for handling broken characters, and segmentation as none of the above techniques are able to achieve 99% accuracy. OCR for Handwritten text is highly challenging and has a very low recognition rates. Due to lack of dataset for Telugu words, end to end recognition of handwritten text still remained unattempted.

4 REFERENCES

[1] Arun K Pujari, C Dhanunjaya Naidu, B C Jinaga , "An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory ", ICVGIP, Ahmedabad,2002.

[2] C. Vasantha Lakshmi, C. Patvardhan, “High accuracy OCR system for printed Telugu text”,TENCON 2003,Conference on Convergent Technologies for Asia-Pacific region, Volume 4,15-17 October.

[3] A.Negi , C.K.Chereddi, ”Candidate search and elimination approach for Telugu OCR”, TENCON 2003,Conference on Convergent Technologies for Asia-Pacific region, Volume 2.

[4] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran,”A bilingual OCR for Telugu-Hindi documents and its applications”, Seventh International Conference on Document Analysis and Recognition(ICDAR’03), 2003.

[5] C. Vasantha Lakshmi, Ritu Jain, and C. Patvardhan,” OCR for Printed Telugu Text with High Recognition Accuracies”, Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006, Madurai, India, December 13-16, 2006. Proceedings (pp.786-795)

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

[7] P. Pavan Kumar, Chakravarthy Bhagvati, Atul Negi, Arun Agarwal,B. L. Deeekshatulu, “Towards Improving the Accuracy of Telugu OCR Systems”, International Conference on Document Analysis and Recognition ,2011.

[8] Dhandra B.V, Gururaj Mukarambi, Mallikarjun Hangarg, “A Script Independent Approach for Handwritten Bilingual KANNADA AND TELUGU Digits Recognition”, International Journal of Machine Intelligence, ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 3, 2011, pp-155-159

[9] K.Ram Mohan Rao,, B.Ramesh, , G.Indrasena Reddy, “Font and Size Identification in Telugu Printed Document”, International Journal of Engineering Research and Development, Volume 6, Issue 11 (April 2013), PP. 92-110, April 2013.

[10] Ramalingeswara Rao K V, Bhaskara Rao N, Ramesh Babu D R, “Telugu character recognition based on topological feature alterations after selective Morphological unification of the target “, Proceedings of 8th IRF International Conference, 04th May-2014, Pune, India, ISBN: 978-93-84209-12-4.

[11] 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.

[12] Panyam Narahari Sastry ,T.R. Vijaya Lakshmi , N.V. Koteswara Rao T.V. Rajinikanth , Abdul Wahab, “Telugu Handwritten Character Recognition Using Zoning Features”, International Conference on IT Convergence and Security (ICITCS), 2014, 28-30 October.

[13] J. Jyothi, K. Manjusha, M. Anand Kumar and K. P. Soman, “Innovative Feature Sets for Machine Learning based Telugu Character Recognition”, ISSN (Print) : 0974-6n846,ISSN (Online) : 0974-5645, Indian Journal of Science and Technology, Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/79996, September 2015.

[14] N. Shoba Rani, Sanjay Kumar Verma, Anitta Joseph, “A Zone Based Approach for Classification and Recognition of Telugu Handwritten Characters”, International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, No. 4, August 2016, pp. 1647~1653, ISSN: 2088-8708, DOI: 10.11591/ijece.v6i4.10553.

[15] Dr.B.Rama and Santosh Kumar Henge, “OCR-The 3 layered approach for classification and identification of telugu hand written mixed consonants and conjunct consonants by using advanced fuzzy logic controller”, pp. 75–88, 2016. Computer Science & Information Technology (CS & IT), CSCP 2016, DOI : 10.5121/csit.2016.60407.

[16] N Prameela , P Anjusha , R Karthik, “Off-line Telugu handwritten characters recognition using optical character recognition”, 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Volume 2,20-22 April 2017.

[17] Minesh Mathew, Mohit Jain , C.V. Jawahar,” Benchmarking Scene Text Recognition in Devanagari, Telugu and Malayalam”, ”, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),2017, Volume 9,November 9-15.

[18] Devarapalli Koteswara Rao , Atul Negi, “Orthographic Properties Based Telugu Text Recognition Using Hidden Markov Models”, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),2017, volume:9 November 9-15.

[19] Neerugatti Varipally Vishwanath, K. Manjunathachari and K. Satyaprasad, “Handwritten Telugu Composite Character Recognition Using Morphological Analysis”, International Journal of Pure and Applied Mathematics ,Volume 119 No. 18, 2018,pp: 667-676.

[20] Kesana Mohana Lakshmi, Tummala Ranga Babu”, “Telugu Script Recognition Approach Using Kernel Features”, International Journal of Engineering Technology Science and Research (IJETSR), ISSN 2394 – 3386, Volume 5, Issue 5 May 2018.

[21] Konkimalla Chandra Prakash, Y. M. Srikar, Gayam Trishal, Souraj Mandal, Sumohana S. Channappayya, “Optical Character Recognition (OCR) for Telugu: Database, Algorithm and Application”, arXiv:1711.07245v2 [cs.CV] 25 Dec 2018.