2023 4th International Conference for Emerging Technology (INCET)
Belgaum, India. May 26-28, 2023

A Novel Approach to Recognize Handwritten Telugu Words Using Character Level CNN

Ashlin Deepa R N
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
rndeepa.pradeep@gmail.com

Guru Sai Jayanth Kalluri
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
gurujayanth31@gmail.com

Zeeshan Mohammed
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
zeeshan144144@gmail.com

Mantri Pramod Sai Sushank
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
saisushank.mantri@gmail.com

Abhirama Raju V
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vegesnaarr6666@gmail.com

Atul Negi
School of Computer and Information Sciences
University of Hyderabad
Hyderabad, India
atul.negi@uohyd.ac.in

2023 4th International Conference for Emerging Technology (INCET) | 979-8-3503-3575-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/INCET57972.2023.10170305

***Abstract*— A computer can interpret and receive handwritten input from various sources, such as documents and photos. This is an area of great interest in document text recognition and image analysis, especially in the field of Indic languages. The language of Telugu is a part of the Dravidian language family, and it has primary status in multiple states. It has 52 core characters, consisting of 36 consonants and 16 vowels. This paper aims to provide a model to recognize convert handwritten Telugu documents at word level. The dataset consists of words of varying length between one letter word to seven letter words. The recognition accuracy varies according to the length of the word.**

***Keywords*— Handwritten Text Recognition, Telugu Script, Deep Learning, Optical Character Recognition for Indian language, word level recognition.**

I. Introduction

Handwritten text recognition is a complex problem that has been extensively researched in the field of computer vision and pattern recognition. It is a crucial task in digitalization of handwritten documents and has applications in the field of document text analysis, information retrieval, and digital archiving. Telugu language is spoken over 80 million people in India, presents its own unique challenges in recognition of handwritten text. The complexity of the script, combined with variability of individual handwriting styles, presents significant difficulties for recognition algorithms. Despite these challenges, there have been numerous attempts to address the issue, including the development of specialized handwriting recognition systems, the integration of language-specific models, and the use of deep learning techniques.

The English alphabet only has 26 letters. In Telugu language, the letter structure consists of 36 consonants, 3 vowel modifiers, and 16 symbols. The combination of the matras (vowel diacritic) with vowels and consonants allows for the creation of over 18324 unique characters. It can be hard to recognize handwriting characters due to the complexity of their design and variations in strokes as well as style.

TDIL is a government program that aims to promote the standardization of technology in Indian languages. Along with that, it seeks to produce multilingual information resources and make them available to people. E-Aksharayan helps people change printed or scanned documents into editable text. Text recognition can be used in rural schools to improve the students' communication skills and teaching abilities. Since most of the time, the learners prefer to keep their teachers' notes.

To share handwritten notes using mobile scanners, students usually take images of them, but the text in these photos is not clear because of the varying handwriting. This paper proposes a method to convert these notes into editable text.

This research paper aims to explore the latest advances in handwritten text recognition specifically for the Telugu language, highlighting the challenges and solutions to improve the accuracy of recognition algorithms. Through a comprehensive review of the literature and experimental analysis, this paper provides insights into the current state of the field and future directions for further research. The findings of this study will have implications for the development of digital systems that can effectively process and analyse handwritten text in Telugu language.

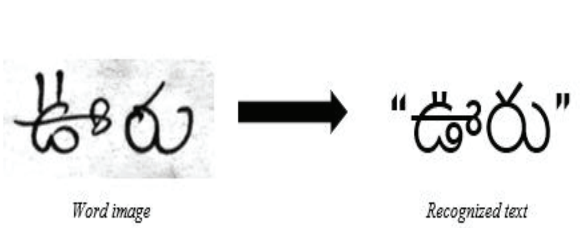

Fig. 1. Handwritten word converting to text

The following paper is structured in the following manner. Section II discusses previous studies that focus on handwritten text recognition in various languages. In Section III, the methodology is briefly outlined, with Sub-section III A contains a detailed explanation of the preprocessing steps and segmentation algorithm, Sub-section III B consists of details about the character classifier and Sub-section III C presents the overall model's architectural details and framework for handwritten word classification. The results are presented in Section IV, and concluding remarks are made in Section V.

II. Related work

Several attempts to recognize handwritten text in various languages have been attempted in recent years. For offline Handwritten Text recognition, Lecun et al. utilized a Neural Networks approach to recognize single digits [1]. Kang et al. built a model based on CNN and Transformers [2]. In another study, Graves and Schmid-Huber utilized a combination of MDLSTM and loss and decoding methods to identify the town names in Tunisia [3]. In a study, Roy et al. [4] analysed six common Indian scripts using various techniques, such as circularity-based features, component-based features, and the Fractal Dimension-Based Feature. Hangarge et al. proposed a recognition method at word level based on the various Indian scripts, such as Devanagari [5]. Barua et al. suggested a method for identifying handwritten city names in Bangladesh using the HOG feature descriptor [6]. In a study, Singhal et al. was able to identify the handwritten characters from various Indian scripts, such as Devanagari, Roman, and Telugu [7].

Ashlin Deepa R N presented various techniques for recognition and classification of offline handwritten character recognition in Tamil language [8-9-10-11]. An end-to-end segmentation free paragraph text recognition neural network was proposed by Denis Coquenet, Clément Chatelain, and Thierry Paquet [12]. A new method for character segmentation and word recognition was proposed by Bhowmick, Tan, and Pal [13].

Several works have been published in the literature about the recognition of machine-printed text in the Telugu script [14-15].A method for offline and online recognition of handwritten Telugu characters has been proposed using the ResNet framework by Bindu Madhuri Cheekati [16]. The proposed method for extracting Telugu documents by Prameela et al. [17] consists of two phases: preprocessing and feature extraction. The first one involves using a median filtering technique to extract the character's boundary edge pixel points, while the second one involves normalizing and skeletonization methods. Rani [18] proposed a method for extracting and classification Telugu documents using a customized template matching method. The method is carried out through a caching method that stores the frequently used character template in a main database. The class representation of the documents is then generated by the various XML databases used in the project. Rajeti and Cheekati [19] present a reliable and fast ResNet for identifying and classifying Telugu documents online and offline. Sarika, Sirisala's paper [20] talks about various types of classification and preprocessing techniques, such as segmentation, digitization, and preprocessing. It also covers HWCR with different ML techniques, such as SVM, Bayesian decision theory, and Bayesian classifier. The paper looks into their native language features and how they work.

Madhavi [21] proposed a model that can detect and correct slant angles of MTW in Telugu.

III. Methodology

In order for Optical Character Recognition (OCR) to function effectively, it is necessary to have both a dataset and a character classifier. These components are critical to the accuracy and efficiency of the OCR process. Without a sufficient dataset, the OCR system may struggle to recognize characters or words accurately, leading to errors or misinterpretations. Similarly, without a reliable character classifier, the OCR system may not be able to distinguish between different characters or fonts, resulting in inaccurate recognition. Therefore, having a high-quality dataset and a robust character classifier are crucial to the success of any OCR system.

A. Preprocessing and Segmentation

When an image is fed to the model, it is processed by thresholding, removing noise, and resulting in a binarized image. In this stage, the noise in the image is reduced using a gaussian filter of size (5,5), and the input is produced by thresholding using the Otsu algorithm [23] and binary thresholding to separate foreground and background pixels.

The segmentation phase allows the model to extract significant details from the images for additional analyses.

Algorithm:

Step - 1: The image is transformed into a numerical format suitable for computer analysis.
Step - 2: The image is transposed to enable processing one row at a time.
Step - 3: Each row is checked for black pixels to detect the presence of a letter.
Step - 4: Rows with letters are grouped together until the end of a character is reached.
Step - 5: The rows of each character are saved as a separate image and added to a list of characters.
Step - 6: The process in steps 3 to 5 is repeated until all characters in the image have been identified and added to the list.

B. Character Classifier

The classifier of an OCR system plays a critical role in determining its performance. However, despite the availability of multiple Telugu character classifiers, none of them cover the complete range of character set. Furthermore, previous studies on Telugu handwritten character classifiers have not included the entire character set. Additionally, there is no standard Telugu character dataset that encompasses all the character combinations. Although certain datasets, like HPL, contain iso-characters, they are insufficient. Only 90 characters, a minuscule portion of all characters are present in the HPL dataset. Thus, the dataset used in this study, created by [22], is unique, covering 17387 categories of classes, with each category having 560 samples.

In general, a Telugu character is composed of two key components - the primary character and the vattu/gunintham. This paper employs a classifier consisting of two convolutional neural networks (CNNs), with the first CNN being responsible for identifying the main character, and the second CNN being utilized for recognizing the vattu and/or gunintham associated with the main character. Architecture diagram for both the CNN models is given in Fig. 2, Fig. 3. To improve accuracy, the model employs a straight-line

Hough transform-based method to correct skewing, which can detect and correct up to 90 degrees of skew, as well as a modified version of Otsu's thresholding for binarization and a morphological closing algorithm for noise reduction. This character classifier boasts an accuracy rate of 98.74% for identifying the main character and 96.09% for identifying the vattu/gunintham. It is important to note that the accuracy rates mentioned above (98.74% for the main character and 96.09% for the vattu/gunintham) are specific to printed Telugu characters.

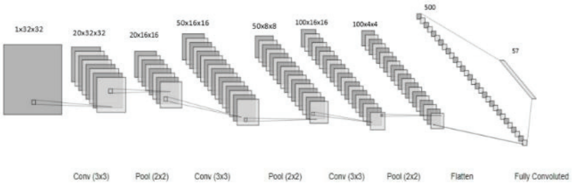

Fig. 2. Architecture for main character

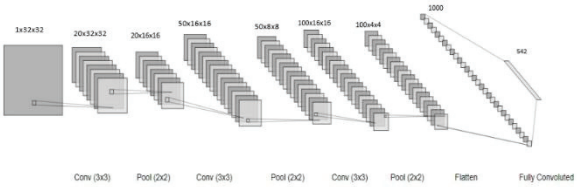

Fig. 3. Architecture for vattu/gunintham

C. Recognition

A character in Telugu is generally divided into two main components: the main character and a gunintham or vattu. Since there are so many classes arising from the different permutations of these two characters, a 2 CNN architecture character recognition model is utilized to classify them, where the first aims to identify the primary character, while the second identifies the variants of the character along with the main one. Thus, a character list is represented by the character recognition model [22], and a predicted string is generated.

The Obershelp/Ratcliff algorithm [24], returns the best match between the given predicted string and the predefined words in the corpus. The algorithm uses a mathematical method to find the closest match between a given candidate string and a target string. This method is very useful in finding the longest continuous sequences between two sequences. This algorithm is used for comparing the similarity of two strings. It has been developed to be both fast and precise, even when the strings being compared have vastly different lengths. The method involves dividing the strings into sequences of matching characters, and then determining the longest sequence of matching characters between the two strings.

This process is then repeated with the remaining substrings until either all characters match, or no further matches can be found. The result of the comparison is a score that reflects the degree of similarity between the two strings, with a value ranging from 0 (no match) to 1 (perfect match). This algorithm is an efficient and effective tool for string matching, widely used in various applications. Architecture diagram for handwritten word recognition is given in Fig. 4. The result returned by this method is stored as recognized word.

The obtained results are compared with result obtained using Levenshtein distance algorithm [25]. It is a prominent metric for determining the dissimilarity of two strings. It is calculated by determining the smallest number of single-character modifications required to transform one string into the other, with suitable operations such as character insertion, deletion, and substitution.

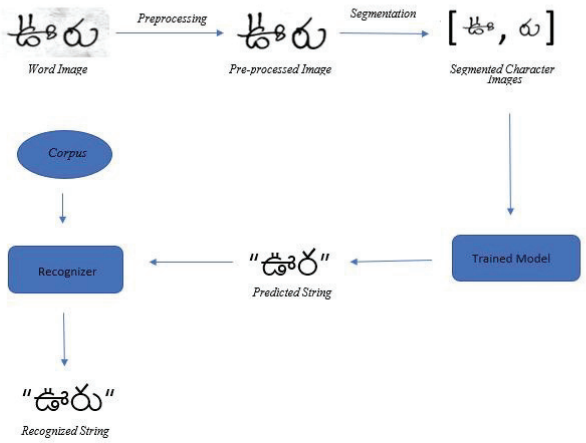

Fig. 4. Architecture diagram to recognise handwritten words.

IV. Analysis

The dataset was collected from various sources, such as students from Telugu medium schools and also individuals who know how to write in Telugu. There were about 18,000 unique characters in the language, which is a bit complex for machine learning computations. The dataset consists of 2,980 words. The dataset upon testing with various methods (Ratcliff/Obershelp, Levenshtein) obtained the word accuracies, word error rates, character accuracies and character error rates as shown in Table I and Table II. The word accuracy graph is shown in Fig. 5. where P denotes prediction of word without using any method, GC denotes the result obtained using Ratcliff/Obershelp algorithm, L1, L2, L3 denotes the result obtained using Levenshtein distance method).

TABLE I. Word Accuracy Scores v/s Method

	Obershelp /Ratcliff algorithm	***Levenshtein 1***	***Levenshtein 2***	***Levenshtein 3***
Word error Count	1668	2322	2567	2730
WER	55.9732	77.9196	86.1409	91.6107
Word Accuracy	44.0604	22.0805	13.8590	8.3892

TABLE II. CHARACTERS ACCURACY SCORES V/S METHOD

	Obershelp /Ratcliff algorithm	***Levenshtein 1***	***Levenshtein 2***	***Levenshtein 3***
character error count	7790	8347	9465	11405
CER	56.9735	61.0473	69.2240	83.4125
Character Accuracy	43.0264	38.9526	30.7759	16.5874

Out of all the algorithms, the Obershelp/Ratcliff method was fed with 2980 words which consisted of 13,673 characters. This method obtained a Character Error Rate (CER) of 56.9% and Word Error Rate (WER) of 55.9%. The methods v/s error rate graph is shown in Fig. 6. The word accuracy obtained was 44.1% along with Character Accuracy of 43.4%. The error count of word and characters were 1667 and 7793 respectively.

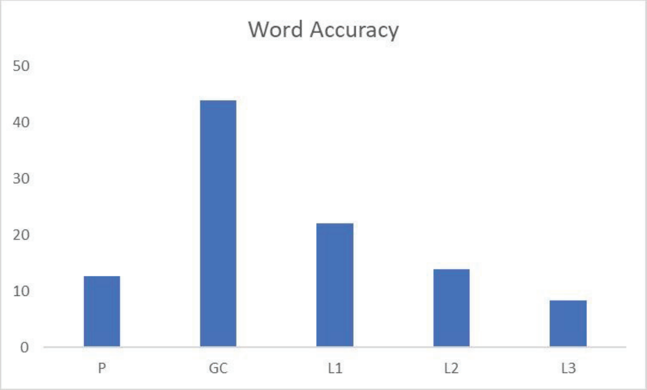

Fig. 5. Graph representing Methods vs Word Accuracies

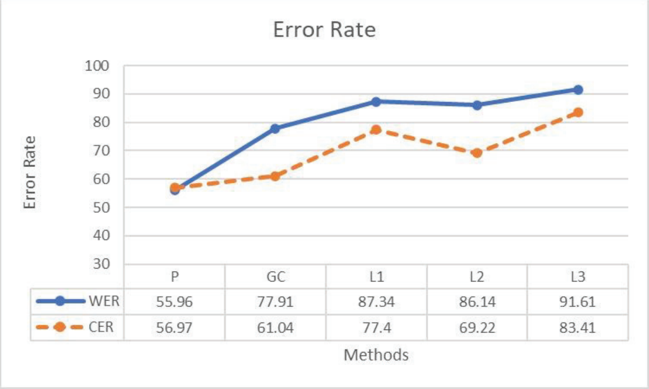

Fig. 6. Graph representing Methods vs Error Rates (Word, Character) (where P represents Direct prediction (without auto-correct), GC represents Ratcliff/Obershelp algorithm, L1-L2-L3 represent Leveshtien1-Levenshtien2-Levenshtien3 respectively)

About 50% of the text in the dataset of 3000 words contains two-letter words. The model achieves an accuracy score of 49.78%. 28% of the dataset contains three-letter words, where the model achieves an accuracy of 51.43%. 13% of the text contains four-letter words, where the model achieves an accuracy of 31.01%. 8.3% of the text contains five-letter words, where the model achieves an accuracy of 63.19%. 0.7% of the text contains six and seven-letter words, where the model achieves an accuracy of 23.80%.

V. CONCLUSION

The recognition of handwritten text in Telugu language is a complex problem that requires the integration of multiple techniques and approaches. The Handwritten Text Recognition for Telugu Language is a promising area of research that has a lot of potential for real-world applications. The successful development of a handwritten text recognition system for Telugu would not only provide a useful tool for the Telugu speaking population but also contribute to the advancement of the field of OCR. This work can be extended to paragraph level recognition after applying post OCR error correction methods.

It is important to note that the implementation of the Handwritten Text Recognition system for Telugu language requires a robust and scalable architecture to handle the complexity of the language. The development of a Handwritten Text Recognition system for Telugu has the potential to revolutionize the way Telugu language is processed and used in various industries such as education and healthcare.

The study findings suggest that there is potential for improving the accuracy and robustness of the model through fine-tuning with a larger dataset. To further enhance the performance and accuracy of the model, it is crucial to have a comprehensive Telugu Handwritten dataset. With such a dataset, we could fine-tune the model and optimize its performance for Telugu language, ultimately improving its accuracy and robustness. Future studies could focus on enhancing the algorithms by expanding the dataset, which helps in improving the system's overall performance. Additionally, researchers may prioritize improving the accuracy, robustness, and scalability of the system.

ACKNOWLEDGMENT

This work has been supported by the Department of Science & Technology, Scheme for Young Scientists and Technologists, under Innovation Technology Development and Deployment.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[2] Kang, Riba P, Rusin˜ol M, Forn´es A., and Villegas, M. “Pay attention to what you read : Non-re current handwritten text-line recognition”, 2020.

[3] Graves, A. and Schmidhuber: “Offline hand-writing recognition with multi -dimensional re current neural networks”: In NIPS, 2008.

[4] K. Roy, S. Das, S. Obaidullah, Script Identification from Handwritten Document: In Proceedings of The third National Conference on Computer Vision, Pattern Recognition, ImageProcessing and Graphics (NCVPRIPG)., Hubli, Karnataka., pp. 66-69, 2011.

[5] Hangarge, M., Santosh,K., Pardeshi,R., :Directional Discrete Cosine Transform for Hand-written Script Identification,: In Proceedings of 12th International Conference on Document Analysis and Recognition, pp. 344-348, 2013.

[6] S.Barua, Malakar, Bhowmik, S. : Bangla hand written city-name recognition using gradientbased feature.: Proc. 5th Int. Conf. on Frontiers in Intelligent Computing: Theory and Applications, Singapore, 2017, pp. 343– 352

[7] Singhal, V., Navin, D. Ghosh, :Script-based Classification of Hand written Text Document in a Multi-lingual Environment., Research Issues in Data Engineering, pp.47, 2003.

[8] Ashlin Deepa R N, Rajeswara Rao, R. A novel nearest interest point classifier for offline Tamil handwritten character recognition. Pattern Anal Applic 23, 199–212 (2020).

[9] Ashlin Deepa R N & Rajeswara Rao, Ramisetty. (2016). An efficient offline Tamil handwritten character recognition system using zernike moments and diagonal-based features. 11. 2607-2610.

[10] Ashlin Deepa R N & Rajeswara Rao, Ramisetty. (2017). A modified GA classifier for offline Tamil handwritten character recognition. International Journal of Applied Pattern Recognition. 4. 89. 10.1504/IJAPR.2017.10003582

[11] Ashlin Deepa R N & Rajeswara Rao, Ramisetty. (2017). An Eigencharacter Technique for Offline-Tamil Handwritten Character Recognition. 10.1007/978-981-10-2035-3_51.

[12] Coquenet, Denis & Chatelain, Clement & Paquet, Thierry. (2020). “End to end Hand written Paragraph Text Recognition Using a Vertical Attention Network.”

[13] Shiva kumara,P., Bhowmick, B. Su, C. Tan and U. Pal, A New Gradientbased character Segmentation Method for Video Text Recognition,: 2011 International Conference on Document Analysis and Recognition, Beijing, China, 2011, pp. 126-130, doi: 10.1109/ICDAR.2011.34

[14] Chinnuswamy, P., Krishnamoorthy, : “Recognition of Handprinted Tamil Characters.”: Pattern Recognition 12, 141–152 (1980)

[15] N.Shanthi, K.Duraiswamy: “Preprocessing Algorithms for Recognition of Tamil Hand written Characters.”: In:3rd Int.CALIBER (2005).

[16] B. M. Cheekati and R. S. Rajeti, Telugu hand written character recognition using deep- residual learning,: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 788-796, doi: 10.1109/I-SMAC49090.2020.9243348.

[17] N.Prameela ,P.Anjusha, andR.Karthik, (2017), “Offline Telugu hand written characters recognition using optical-character recognition.” In2017 International conference of Electronics, Communication and Aerospace Technology (Vol. 2, pp. 223- 226). IEEE.

[18] Rani, N.S., Vasudev, T. and Pradeep, C.H., (2017). “An Enhanced Template-Matching-Technique for Recognition of Telugu Script”.: International Journal of Signal Processing.

[19] B.M. Cheekati, and Rajeti, (2020 October). “Telugu handwritten character recognition using deepresidual learning.”: In2020 Fourth International Conference on IoT in Social, Mobile, Analytics and Cloud (pp. 788-796). IEEE.

[20] Sarika, and Sirisala, (2021). “Deep Learning Techniques for Optical Character Recognition.” InSustainable Communication Networks and Application (pp. 339-349). Springer, Singapore.

[21] G.B .Madhavi, Kumar, and V.K.,Vakula, (2021). “An Effective Slant Detection and Correction Method Based on the TiltedRectangle Method for Telugu Manuscript Terms.”:International Journal of Information Technology Project Management 12(4), pp.25-37.

[22] Chandra Prakash, K. , Srikar, Y.,M., Trishal, G., Mandal, S., and Channappayya, S., Optical-Character Recognition (OCR) for Telugu: Database, Algorithm and Application,: 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018, pp. 3963-3967, doi: 10.1109/ICIP.2018.8451438.

[23] Otsu, N., A Thresholdselection Method from Gray Level Histograms, in IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[24] Ratcliff, J. W. and Metzener, D. E. (1988). Pattern-matching-the gestalt approach. Dr Dobbs Journal, 13(7):46.

[25] Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." In Soviet physics doklady, vol. 10, no. 8, pp. 707-710. 1966