International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Comprehensive study of deep learning based Telugu OCR: A survey

M. V. Vijaya Saradhi, K. Rakesh *, D. Ravi Prasanna, K. Swetha and B. Prawin

Department of Computer Science and Engineering, ACE Engineering College, Hyderabad, Telangana, India.

International Journal of Science and Research Archive, 2023, 08(01), 353–356

Publication history: Received on 03 December 2022; revised on 14 January 2023; accepted on 17 January 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.8.1.0050

Abstract

There will be no computer-editable text in the image file. The method of optical character recognition (OCR), which can read handwritten or printed text in images, turns that text into a file that can be edited on a computer. English has a well-developed OCR system. OCR is currently required for Indian languages in order to process application forms, categorise books in libraries, and preserve historical records that are mostly written in Indian languages. The Telugu language presents challenges for OCR since each character is made up of a single consonant, a single vowel, or a compound of vowels and consonants.

Keywords: Text segmentation; Text Extraction; Image-based; Document processing; Convolution Neural Network; OCR

1. Introduction

This project is “Comprehensive Study of Deep Learning based Telugu OCR”. Sometimes it is required to scan the documents and it is also required to edit those scanned documents as per our needs. The technology used to deal with this is an Optical Character Recognition System (OCR). An OCR converts any image containing texts into a format that can be edited with the help of MS word. It creates a digital file for the scanned document. For this the algorithm should be present in the administrator system. Once the user will upload the image it will be processed and convert it into a text. Then user will be edited with the help of MS word. The development of a full OCR tool for Telugu script has received little research. Although the need for an OCR system is justified by the availability of a sizable corpus of scanned data on the internet, the difficulty of the problem is exacerbated by the more sophisticated script and agglutinative grammar. Building a solution that functions well on real-world files that include sound and erasure is more difficult. The two main components of OCR are segmentation and recognition. The plan of each is guided by that of another. The segmentation will be stronger (to sound, erasure, skew, etc.), making this recognizer's work easier, and vice versa. Through all areas, segmentation procedures are comparable. This is due to the fact that typically, a related component (a nearby patch of ink).

2. Literature survey

Deekshatulu et al. and Rajasekaran (1977) The concept of "Recognition of written Telugu characters" using digital image processingwas put up by Rajasekaran and Deekshatulu. This is the initial investigation into Telugu character OCR. It offers atwo-stage syntax-aided character recognition approach and lists 50 essential characteristics. The rudimentary shapes are located and disregarded using apreliminary knowledge-based search. After the primitives are removed, the pattern thatis left over is traced along its points in the second step of coding. A decision tree isused to carry out the classification process. The definition of each letter by the correct joining and superimposition of primitives.

* Corresponding author: K. Rakesh

Sukhaswami et al. (1995) "Neural networks are used to recognise Telugu characters." a proposal by Sukhaswami et al. from 1995, provides an illustration of this. An associative memory neural network serves as the first recognition component of the (Hopfield neural network) The storage issue might be resolved by this multiple neural network associative memory (MNNAM), it was demonstrated. Limited storing capability of the Hopfield neural network. In the networks employ unrelated to one another. They offered proof that the storage shortage might be solved by this plan.

Rao and Ajith et al. (1995) Rao and Ajitha came up with Telugu Script Recognition—a Feature-Based Approach. Telugu characters are said to be composed of circular components with different radii, and this idea is used in the work. Separating Recognizing the characters requires breaking them down into their component parts and identifying each one. Because they maintain the traditional forms of Telugu characters, circular segments were chosen as the feature set. According to reports, when the reference and test sets were related to the same topic, the recognition rates increased from 91 to 95% to between 91 and 90% across a variety of subjects.

Negri et al. (2002) published a paper titled "Non-linear Normalization to Improve Telugu OCR." Scaling glyphic regions with minor curvature were chosen for application. The normalisation of dot density features serves as the foundation for this. Shape distortions were detected by the authors, although they also noted an improvement in OCR recognition accuracy. Not examined is how well different fonts perform.

Pujari et al. (2002) According to A Telugu character recognizer using multi-resolution analysis and associative memory was published by Pujari. is proposed. Line segmentation of the grey level input text pictures is accomplished horizontal projections, whereas vertical projections are used for word segmentation. Images are scaled evenly to 32x32 using zero-padding techniques. In order to create a group of four 8x8 images, a 32x32 image is down sampled three times using wavelet representation. Only the average image is then used for further processing... Using the mean value of the grey level as the threshold, character images measuring 8x8 are transformed to binary images. The generated 64-bit bit string is utilised as the input symbol's signature. For the goal of recognition, a dynamic neural network based on Hopfield is created.

Jawahar et al. (2003) " A Hindi-Telugu Document Bilingual OCR and Its Applications" was Jawahar's accepted proposal. It is based on support vector classification and principal component analysis. Around 96.7% of the total is thought to be the accuracy.

Lakshmi & Patvardhan et al (2003) Lakshmi and Patvardhan came up with the idea for " Fundamental Symbols in Printed Telugu Text: Optical Character Recognition." Each character (basic sign) is given its smallest enclosing rectangle with the same aspect ratio, which is then scaled to 36 columns.All the symbols are categorised into height groups in order to create a preliminary classification (rows). From the second and third order moments, a total of seven invariant moments are extracted to create the feature vector. These feature vectors can be identified using K-nearest neighbour technology. Both the training and test sets of data were written in the same font. Gaussian, salt-and-pepper, and speckle noise are also introduced to the test datasets' noisy character pictures. This does not apply to pre-processing like line, word, and other types.

Anuradha Srinivas et al. (2007) Anuradha Srinivas have suggested a single font Telugu optical character recognition system. By maximising the variation in the horizontal projection profile, Sauvola's technique is used to accomplish binarization when skew detection and correction is being carried out Lines, words, and characters are separated from the text using horizontal and vertical projection profiles. Telugu characters are quantified as zero-crossing features, and 11 groups of characters are created based on these zero-crossing attributes. The test character's group number is identified in the first stage of a two-stage classifier, and the test character itself is identified in the second stage of the classifier using a minimum-distance algorithm. Accurate recognition is stated to be 93.2%.

Table 1 Literature Survey

Sl.No	Year	Author	Technology used	Advantages	Disadvantages	Remarks
1.	1977	Rajasekaran & Deekshatulu	Decision Tree	Pattern will be obtained when primitives are removed, and it is coded by following the points on it.	It only recognises 50 simple features.	After the primitives have been eliminated, the pattern is obtained, and points on it are traced to create the code.
2.	1995	Sukhaswami	HNN (Hopfield neural network)	It was proven that this multiple neural network associative memory (MNNAM) could solve the storage problem.	The Hopfield neural network's limited capacity for storage.	The networks utilise sets of training patterns that are mutually exclusive.
3.	1995	Rao & Ajitha	Feature-Based approach	When the reference and test sets were from the same subject, the recognition scores were reported as 91 to 95%.	Higher computation costs were needed.	According to reports, the recognition ratings for the various subjects range from 78 to 90%.
4.	2002	Negi	SVM Classification.	OCR recognition accuracy has reportedly increased.	Simple characters are divided into their constituent glyphs.	The performance of various fonts is not examined.
5.	2002	Pujari	HNN (Hopfield neural network)	It has been noted that the performance varies from 93% to 95% across fonts and sizes.	The identification rate for English characters using this method was incredibly poor.	According to reports, the performance varies from 93% to 95% across fonts and sizes.
6.	2003	Jawahar	SVM(Support vector machine)	SVM classification reports the highest accuracy.	It only applies to Hindi-Telugu documents.	It is reported that the overall accuracy is about 96.7%.
7.	2003	Lakshmi & Patvardhan	k-nearest neighbor algorithm	Both training and test data use the same font style.	Works only with a specific sort of input.	This study doesn't handle pre-processing like line, word, or character segmentation.

8.	2007	Anuradha Srinivas	Sauvola's algorithm	To identify Telugu characters, crossing features are used.	It only applies to a single font.	The test character's group number is identified by a classifier in the first stage; the character itself is identified in the second stage by a minimum-distance classifier. It is reported that the recognition accuracy is 93.2%.

3. Conclusion

This suggested system offers typeface independence. With this method, CNN was used, which has a high accuracy rate. This algorithm's output was created with the intention of recognising the Telugu character set accurately. The telugu characters that are written or printed on an image will be detected and turned into digitally editable texts using the suggested methods. In this project, a tool that displays each Unicode character from a document image is created. The associated Telugu text can also be displayed using this Unicode.

Compliance with ethical standards

Acknowledgments

We would like to express our gratitude to the project coordinator Mrs. Soppari Kavitha and our guide Dr. M. V. VIJAYA SARADHI for their assistance in completing the survey. We are very appreciative of Dr. M. V. VIJAYA SARADHI, Head of the Computer Science and Engineering Department at Ace Engineering College, for his invaluable and ongoing help.

Disclosure of conflict of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

References

[1] Recognition of written Telugu characters, Rajasekaran S.N.S. and Deekshatulu B.L. 1977. Graphics in computers, image processing

[2] Sukhaswami R. 1995, "Recognition of Telugu Characters Using Neural Networks,"

[3] Rao P. V. S. & T. M. Ajitha 1995 Telugu Script Recognition - a Feature Based Approach. Proce.of ICDAR, IEEE.

[4] Atul Negi, 2002 Indo-European Conference Telugu OCR Non-linear Normalization to Improve Multilingual Communication Technologies Proceedings, Tata McGraw Hill Book Co., New Delhi.`

[5] Scripts by C. Dhanunjaya Naidu, B. C. Jinaga, and Others: Multiresolution Analysis and Associative Memory for Adaptive Telugu Character Recognition and Pujari Arun K. The Ahmedabad, India, ICVGIP, 2002.

[6] M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, and Jawahar C. V. 2003. An application that provides Telugu and Hindi documents with bilingual OCR. International Symposium on Document Recognition and Analysis.

[7] Lakshmi C.V. et al., "Optical Character Recognition of Basic Symbols in Printed Telugu Text,"2003.

[8] Anuradha Srinivas, 2007 - Arun Agarwal, C.R. Rao Character recognition for Telugu. The Hyderabad International Symposium on Systemics, Cybernetics, and Informatics.