International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.133 ISSN: 2581-4621

COMPREHENSIVE STUDY OF DEEP LEARNING BASED TELUGU OCR

Dr. M.V. Vijaya Saradhi[1], K. Rakesh[2], D. Ravi Prasanna[3] K. Swetha[4], B. Praveen[5]

[1]Professor and Head of the CSE Department of Ace Engineering College in Hyderabad, Telangana, India.
[2,3,4,5] Students, Department Of CSE, Ace Engineering College, Hyderabad, Telangana, India.

ABSTRACT

The aim of the project is to understand offline One of the most popular and difficult pattern recognition subjects is the use of optical character recognition (OCR) to read handwritten Telugu letters. This study suggests a three-stage OCR solution for Telugu documents that includes pre-processing, feature extraction, and classification. For the extraction of boundary edge pixel points during preprocessing, we used median filtering on the input characters as well as normalisation and skeletonization techniques. Each character is initially divided into three 3x3 grids during the feature extraction stage, and the associated centroid for each of the nine zones is assessed. This allows us to recognise characters in various styles. Following that, we drew the projection angel's horizontal and vertical symmetry to the character's closest pixel.

Keywords: *Image-based; Document processing; OCR; Text segmentation; Text Extraction;*

INTRODUCTION

In the digital world optical character recognition is one most interesting and challenging topic in pattern recognition. Additionally, recognition of various scripting languages is very much vital with respect to historical culture. This recognition system has evolved to accommodate offline characters in addition to printed, handwritten, and both. Every stage of a recognition system, from the beginning to the classification step, is crucial. A recognition system also needs a solid classifier and a well-defined feature extraction process in order to have a high success rate .Because of the wide range of writing styles, forms, and orientations, developing a good recognition system for handwritten formats is still a difficult task. Researchers have reported using a variety of methods to study different scripts, including Arabic, Chinese, and English. Basically Language written in Telugu is one the most popular languages which is derived from Devangiri scripts. It is one of the regional languages of India, mostly spoken at (Telugu) south, part of Indian. To achieve a good accuracy of recognition for handwritten characters of Telugu scripts is quite impressive. Though a good number works has done for Indian regional languages but a less in number related to Telugu script. In this paper we have used study of the character pictures using statistics We have set up 3x3 grids for each character and evaluated associated centroids for each of the nine zones in order to obtain the feature vector values. Draw the horizontal and vertical projection angles to the image's nearest pixel after that. Calculate the centroid of each zone's horizontal and vertical Euclidian distances for the same nearest pixel. The essential feature vector for the proposed recognition system is therefore further assumed to be these resultant values.

EXISTING SYSTEM

In the existing system, a decision tree algorithm is used Telugu characters written down are recognised using digital image processing. HNN(Hopfield neural network) is used to recognize the Telugu character by using multiresolution analysis and associative memory. sauvola’s technique is used to recognition Utilising the horizontal projection profile's variance to its fullest Telugu optical characters in a single font can be produced.

DRAWBACKS

- It only recognizes 50 simple features.
- The identification rate for Telugu characters is incredibly poor.
- It is only applicable to a single font.

PROPOSED SYSTEM

The Proposed system is to identify offline One of the most popular and difficult topics in pattern recognition is the use of optical character recognition (OCR) to read handwritten Telugu letters. This study suggests a three-stage OCR solution for Telugu documents that includes pre-feature extraction, processing, and classification. We used median filtering on the input characters during the preprocessing stage, together with the normalization and skeletonization methods, to extract the boundary edge pixel points from the characters. Each character is initially divided into three 3x3 grids during the feature extraction stage, and the associated centroid for each of the nine zones is assessed. This allows us to recognize characters in various styles.

COMPREHENSIVE STUDY OF DEEP LEARNING-BASED TELUGU OCR ARCHITECTURE

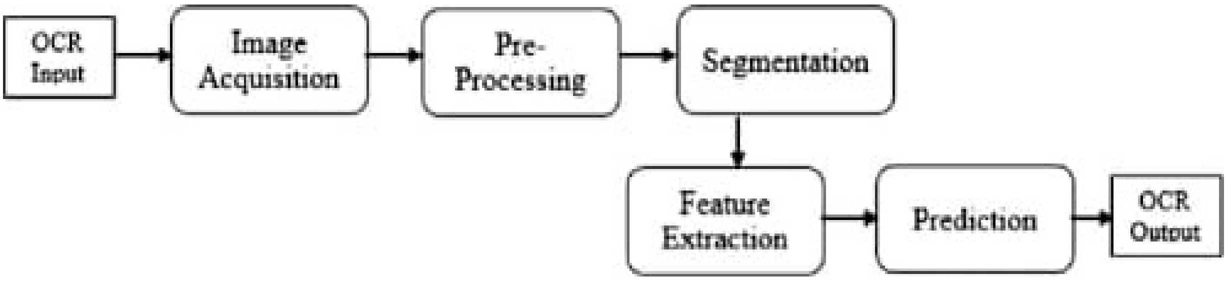


fig 1 : Architecture diagram

OCR:

The electronic or mechanical conversion of images of typed, handwritten, or printed text into machine-encoded text from a document that has been scanned or photographed. OCR, or an optical character reader, is a device that takes pictures of documents or scenes..

IMAGE ACQUISITION:

The picture-processing process begins with the acquisition of an image. Setup is another term for this stage of image processing. It entails getting the image from a source, typically a hardware-based source.

PRE-PROCESSING:

Preprocessing photos is a great way to enhance their quality and prepare them for analysis and further processing. Noise reduction, contrast improvement, image resizing, colour correction, Some efficient picture preprocessing methods include segmentation, feature extraction, etc.

SEGMENTATION:

Segmenting images is a common technique in digital image processing. relies on the properties of the pixels in the image, and analysis to divide an image into various segments or regions.

FEATURE EXTRACTION:

Feature extraction is a method of processing raw data so that it can be handled numerically while retaining the original information. Compared to using machine learning straight up, it produces better results to the raw data.

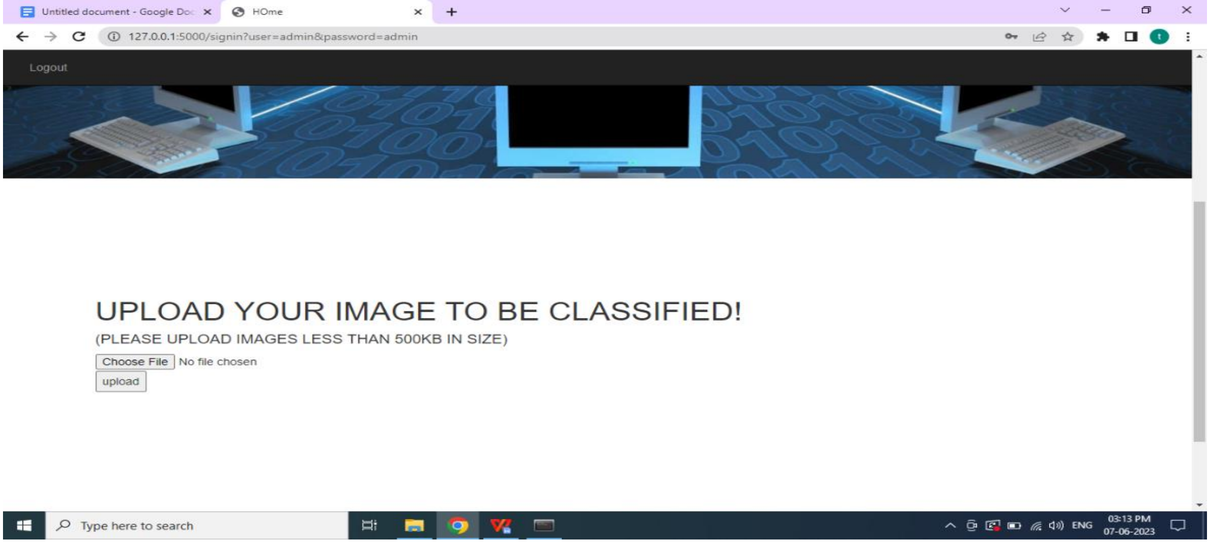
PREDICTION:

An analytical performance prediction method is presented for predicting the performance of transputer based low-level image processing operations.

OCR OUTPUT:

Optical Character Recognition (OCR) is the process of detecting and reading text in images through computer vision.

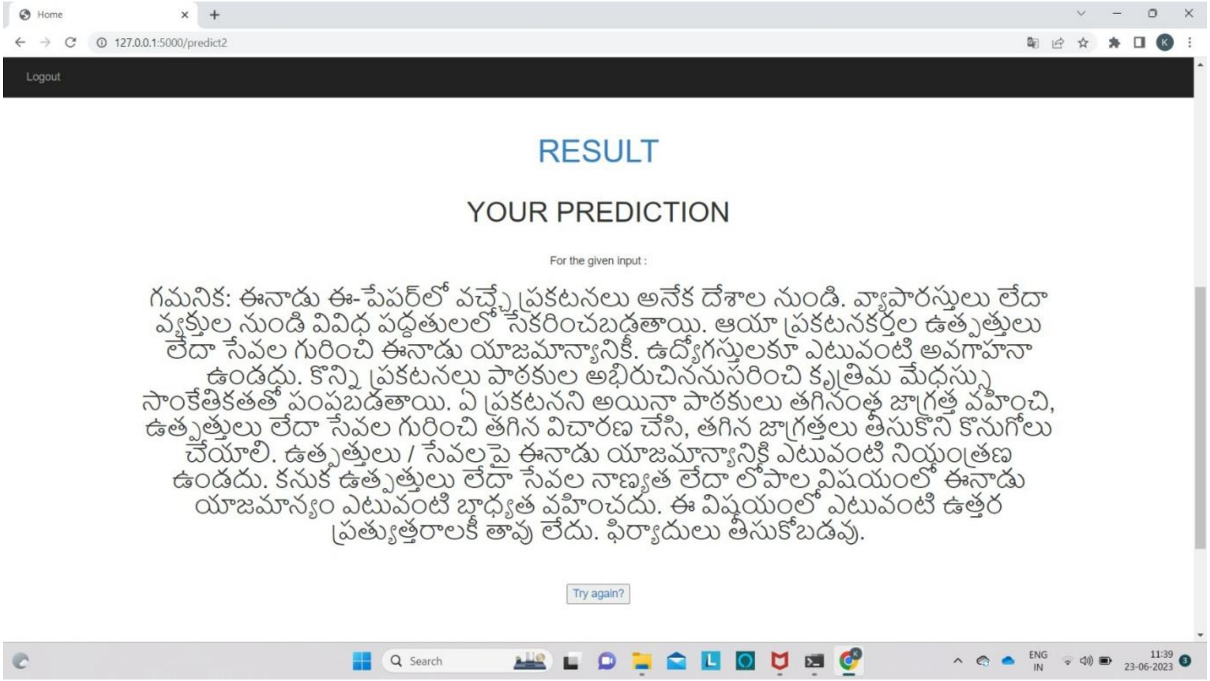
RESULT



UPLOAD YOUR IMAGE TO BE CLASSIFIED!

(PLEASE UPLOAD IMAGES LESS THAN 500KB IN SIZE)

fig 2 : upload image



RESULT

YOUR PREDICTION

For the given input :

గమనిక: ఈనాడు ఈ-పేపర్‌లో వచ్చే ప్రకటనలు అనేక దేశాల నుండి. వ్యాపారస్తులు లేదా వ్యక్తుల నుండి వివిధ పద్ధతులలో సేకరించబడతాయి. ఆయా ప్రకటనకర్తల ఉత్పత్తులు లేదా సేవల గురించి ఈనాడు యాజమాన్యానికీ. ఉద్యోగస్తులకూ ఎటువంటి అవగాహనా ఉండదు. కొన్ని ప్రకటనలు పాఠకుల అభిరుచిననుసరించి కృత్రిమ మేధస్సు సాంకేతికతతో పంపబడతాయి. ఏ ప్రకటనని అయినా పాఠకులు తగినంత జాగ్రత్త వహించి, ఉత్పత్తులు లేదా సేవల గురించి తగిన విచారణ చేసి, తగిన జాగ్రత్తలు తీసుకొని కొనుగోలు చేయాలి. ఉత్పత్తులు / సేవలపై ఈనాడు యాజమాన్యానికి ఎటువంటి నియంత్రణ ఉండదు. కనుక ఉత్పత్తులు లేదా సేవల నాణ్యత లేదా లోపాల విషయంలో ఈనాడు యాజమాన్యం ఎటువంటి బాధ్యత వహించదు. ఈ విషయంలో ఎటువంటి ఉత్తర ప్రత్యుత్తరాలకి తావు లేదు. ఫిర్యాదులు తీసుకోబడవు.

Fig 3 :Output

CONCLUSION

We presented a very comprehensive and practical OCR system for Telugu language. The proposed system is shape and font dependent and requires pre-processing and feature extraction. It may be seen that the two shapes of characters resemble each other with difference in the region at the bottom and at the top. To conclude, the design, approach and implementation were driven by the need for a practical OCR system for Telugu character recognition.

ACKNOWLEDGEMENT

The survey's research would not have been effective without the guidance of our mentor, Dr. M. V. VIJAYA SARADHI, and the project's organiser, Mr. V. Chandra Sekhar Reddy. We sincerely thank to The head of the computer science and engineering department at Ace Engineering College, Dr. M. V. VIJAYA SARADHI, for his important and unwavering support in getting this project done.

REFERENCES

[1] V. K. Govindan, and A. P. Shivaprasad, “Character recognition –a review”, Pattern Recognition, vol. 23, pp. 671-683, 1990.

[2] J. Mantas, "An overview of character recognition methodologies", Pattern Recognition, vol. 19, no. 4, pp. 425–430, 1986.

[3] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwritten Recognition: A Comprehensive,IEEE Trans on PAMI , vol. 22, pp. 62-84, 2000.

[4] U. Pal and B. B. Chaudhuri, “Indian script character recognition: a survey”, Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[5] Recognition of Telugu letters using neural networks, Sukhaswami, P Seetharamulu, International Journal of Neural Systems, 6(3):317-57 · October 1995

[6] S.V.Rajashekararadhya, and Dr.P.Vanaja Ranjan ,Handwritten numeral/mixed numerals recognition of south-indian scripts: the zonebased feature extraction method”, Journal of Theoretical and Applied Information Technology

[7] C.Vikram and C.Shoba Bindhu,”Hand written character Recognition for Telugu Script using Multilayer Perceptrons”,IJARCET-VOL2

[8] N. Anupama, Ch. Rupa & Prof. E. Sreenivasa Reddy “Character Segmentation for Telugu Image Document using Multiple Histogram Projections”

[9] C. Vasantha Lakshmi,Ritu Jain and C. Patvardhan “OCR High Recognition Accuracy of Printed Telugu Text," Springer.