IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Handwritten Optical Character Recognition to Digital Text Conversion

Nagaraj M[1], B S Nanditha[2], Chaitanya G[3], Anusha Kumbar[4], Anusha B[5]
Associate Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4]
RYM Engineering College, Ballari, VTU Belagavi, Karnataka, India

Abstract: *Handwritten Optical Character Recognition (OCR) is a crucial technology for converting handwritten text into digital format. This process involves detecting and interpreting handwritten characters from images or scanned documents. In this project, we focus on OCRfor three languages: Kannada, Telugu, and Hindi. The system utilizes machine learning algorithms, specifically trained neural networks, to recognize and transcribe handwritten characters accurately. The OCR system preprocesses the input images, applies character segmentation, and then classifies each segment into the corresponding character class. Post- processing techniques may be applied to improve accuracy and handle noise. The converted digital text can then be further processed, analyzed, or stored as needed. This technology has various applications in digitizing historical documents, automating data entry tasks, and enabling accessibility for visually impaired individuals*

Keywords: Handwritten OCR, Optical Character Recognition, Kannada, Telugu, Hindi, Machine Learning, Neural Networks, Character Segmentation, Digital Text Conversion

I. INTRODUCTION

Handwritten Optical Character Recognition (OCR) is a transformative technology that bridges the gap between handwritten documents and digital text. With the advent of digitization and the increasing need for efficient data processing, OCR has emerged as a vital tool in various industries and applications. In this project, we focus on developing an OCR system tailored for three prominent Indian languages: Kannada, Telugu, and Hindi. These languages pose unique challenges due to their complex scripts and diverse character shapes. Unlike Latin-based scripts, such as English, Kannada, Telugu, and Hindi characters exhibit intricate strokes and ligatures, making automated recognition more challenging. The process of handwritten OCR involves multiple steps. Firstly, the input images containing handwritten text are preprocessed to enhance clarity and remove noise. Next, the preprocessed images are fed into a machine learning model, typically a neural network, trained specifically for recognizing characters in the target languages. During training, the model learns to extract features from the input images and map them to corresponding character classes. Character segmentation is a critical step in handwritten OCR, where the text regions are divided into individual characters to facilitate accurate recognition. Once segmented, each character segment is classified by the OCR model, and the recognized text is generated as output. The application of handwritten OCR extends across diverse domains, including digitization of historical documents, automation of data entry processes, and enabling accessibility features for visually impaired individuals. By converting handwritten text into digital format, OCR enables efficient searching, indexing, and analysis of textual data, thereby streamlining workflows and enhancing productivity.

II. REVIEW OF LITERATURE

This paper summarizes the top state-of-the-art contributions reported on the MNIST dataset for handwritten digit recognition. This dataset has been extensively used to validate novel techniques in computer vision, and in recent years, many authors have explored the performance of convolutional neural networks (CNNs) and other deep learning techniques over this dataset. To the best of our knowledge, this paper is the first exhaustive and updated review of this dataset; there are some online rankings, but they are outdated, and most published papers survey only

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

closely related works, omitting most of the literature. This paper makes a distinction between those works using some kind of data augmentation and works using theoriginal dataset out-of-the-box. Also, works using CNNs are reported separately; as they are becoming the state-of-the-art approach for solving this problem. Nowadays, a significantamount of works have attained a test error rate smaller than 1% on this dataset; which isbecoming non-challenging. By mid-2017, a new dataset was introduced: EMNIST, which involvesboth digits and letters, with a larger amount of data acquired from a database different than MNIST's. Inthis paper, EMNIST is explained and some results are surveyed.

Humans' control over technology is at an all-time high, with applications ranging from visual object recognition to the dubbing of dialogue into silent films. Using algorithms for deep learning and machine learning. Similarly, the most crucial technologies are text line recognition fields of study and development, with an increasing number of potential outcomes. Handwriting recognition (HWR), also identified as Handwriting Text Acknowledgment, is the capacity of acomputer to understand legibly handwritten input from bases such as paper documents, screens,and other devices. Evidently, we have performed handwritten digit recognition using MNISTdatasets and SVM, Multi-Layer Perceptron (MLP), and CNN models in this research. Ourprimary purpose is to compare the accuracy and execution times of the aforementioned modelsto determine the optimal model for digit recognition.

Recognizing Chinese handwriting may be a difficult topic within the space of character recognition. This paper planned a replacement offline system to acknowledge Chinese written characters. So as to avoid the difficulties in over-segmentation, this paper focuses on thepopularity of text lines, that are assumed to possess been segmental outwardly. We have atendency to evaluate the popularity performance on Chinese handwriting info CASIAHWDB of free Chinese written characters and texts, and incontestable superior performance by theplanned ways. The planned methodology is to implement high recognition rate and speed of written Chinese and written characters. Experiment result shows that our planned approachexpeditiously and effectively improved recognition speed.

Nowadays, deep learning methods are employed in a broad range of research fields. The analysis and recognition of historical documents, as we survey in this work, is not an exception.Our study analyzes the papers published in the last few years on this topic from different perspectives: we first provide a pragmatic definition of historical documents from the point of view of the research in the area, then we look at the various sub-tasks addressed in this research.Guided by these tasks, we go through the different input-output relations that are expected from the used deep learning approaches and therefore we accordingly describe the most used models.We also discuss research datasets published in the field and their applications. This analysis shows that the latest research is a leap forward since it is not the simple use of recently proposed algorithms to previous problems, but novel tasks and novel applications of state of theart methods are now considered. Rather than just providing a conclusive picture of the current research in the topic we lastly suggest some potential future trends that can represent a stimulus.

There are many applications of the handwritten character recognition (HCR) approach still exist. Reading postal addresses in various states contains different languages in any union government like India. Bank check amounts and signature verification is one of the important application of HCR in the automatic banking system in all developed countries. The optical character recognition of the documents is comparing with handwriting documents by a human. This OCR is used for translation purposes of characters from various types of files such as image, word document files.The main aim of this research article is to provide the solution for various handwriting recognition approaches such as touch input from the mobile screen and picture file. The recognition approaches performing with various methods that we have chosen in artificial neural networks and statistical methods so on and to address nonlinearly divisible issues. optimization strategies.

III. METHODOLOGY AND IMPLEMENTATION

3.1 System Design

The methodology involves several key steps. First, we preprocess input images to enhance clarity and remove noise. Then, we apply character segmentation techniques to isolate individual characters. Next, we employ machine learning algorithms, particularly neural networks, trained on labelled datasets of handwritten characters for Kannada, Telugu, and Hindi. These networks classify each segmented character into the appropriate class. Post-processing techniques such as error correction and noise reduction. We fine-tune the model iteratively, incorporating feedback to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

enhance accuracy and robustness. Finally, we integrate the OCR system into a user-friendly interface for practical usage. This methodology ensures the accurate conversion of handwritten text into digital format across multiple languages, catering to various applications effectively.

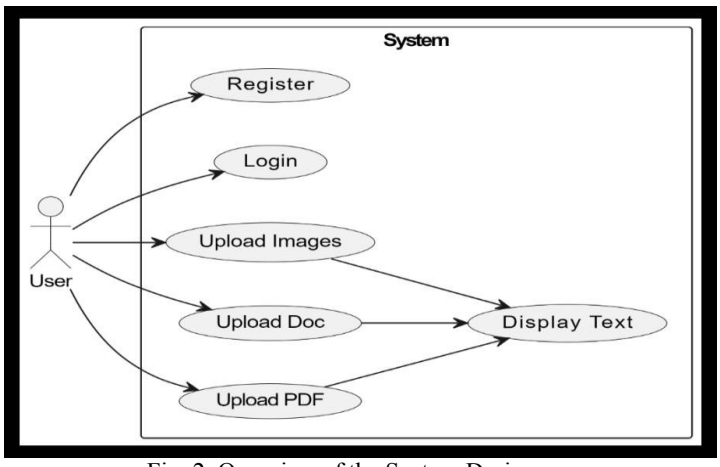

Fig. 2. Overview of the System Design

3.2 Software Design

The software design for handwritten OCR to digital text conversion entails a comprehensive architecture composed of distinct modules addressing various stages of the process. It begins with the Input Module, responsible for capturing handwritten text, followed by Pre-processing techniques aimed at enhancing image quality through noise reduction and normalization. Subsequently, Feature Extraction identifies key characteristics like edges and shapes, leading to Character Segmentation, which breaks down text into individual components. The Recognition Model, often leveraging deep learning algorithms, maps these features to corresponding characters. Post-processing steps refine the recognized text, incorporating language modeling and context analysis. An intuitive User Interface facilitates interaction, while Integration and Deployment ensure seamless incorporation into existing systems

IV. CONCLUSION

The proposed Handwritten Optical Character Recognition (OCR) system represents a significant advancement in digitizing handwritten texts in Kannada, Telugu, and Hindi. By leveraging machine learning and specifically trained neural networks, the system offers a robust solution for accurately transcribing handwritten characters into digital format. Its integration of preprocessing, character segmentation, and post-processing techniques enhances accuracy and efficiency, addressing challenges like noise in the input images. This technology holds immense potential for various applications, including the digitization of historical documents, automation of data entry tasks, and improving accessibility for the visually impaired, marking a pivotal step towards bridging the gap between traditional handwriting and digital data management.

REFERENCES

[1]. TensorFlow. MNIST for ML Beginners. 2017. Available online: https://www.tensorflow.org/get_started/mnist/beginners (accessed on 20 April 2018).

[2]. LeCun, Y.; Cortes, C.; Burges, C.J.C. The MNIST Database of Handwritten Digits. 2012. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 25 April 2018).

[3]. Benenson, R. Classification Datasets Results. 2016. Available online: http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_resu lts.html (accessed on 21 May 2018).

[4]. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324. [Google Scholar] [CrossRef] [Green Version]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

[5]. Belongie, S.; Malik, J.; Puzicha, J. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 509–522. [Google Scholar] [CrossRef] [Green Version]

[6]. Keysers, D.; Deselaers, T.; Gollan, C.; Ney, H. Deformation models for image recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 1422–1435. [Google Scholar] [CrossRef] [PubMed]

[7]. Kégl, B.; Busa-Fekete, R. Boosting products of base classifiers. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18

[8]. June 2009; pp. 497–504. [Google Scholar]

[9]. Decoste, D.; Schölkopf, B. Training invariant support vector machines. Mach. Learn. 2002, 46, 161–190. [Google Scholar] [CrossRef]

[10]. Simard, P.; Steinkraus, D.; Platt, J.C. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK, 3–6 August 2003; Volume 2, pp. 958–963. [Google Scholar]

[11]. Deng, L.; Yu, D. Deep Convex Net: A Scalable Architecture for Speech Pattern Classification. In Proceedings of the 12th Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011; pp. 2285–2288. [Google Scholar]