IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HANDWRITTEN TEXT RECOGNITION

[1]Krutika Bobade, [2] Pratiksha Fusate,[3]Awani Karkade, [4] Vrushali Awale

Assistant Professor

[1, 2, 3, 4] Department of Computer Science Engineering, Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India

Abstract

The study of handwritten Text recognition has gained popularity. The handwritten characters that were scanned as input in the proposed technique were identified using a variety of machine learning algorithms.. It segments each character in the image and recognises the letters after receiving the handwritten document as input in the form of a high resolution image. Additionally, it recognises the letters before going on to find the words in the image. Based on the training it received from the training data, this is accomplished with the help of machine learning algorithms. The specified input image will be provided in word document format as the intended output. Large data sets of images that display the various writing styles and shapes can be used to train the system. When training the system with vast amounts of data, machine learning is crucial. This can also be used to businesses and organisations that only keep critical records in writing form. Through this review and implementation, we aim to provide researchers and practitioners with insights into the current landscape of HTR techniques and offer a practical guide for building accurate and efficient handwritten text recognition systems.

Keywords :– Handwritten Text Recognition, OCR, CNN and RNN Network.

I. INTRODUCTION

The process of turning a handwritten text image into a text file that a computer can read and utilize for a variety of applications is known as handwritten recognition.

The goal of this method is to create software that can comprehend handwritten documents. The technique extracts contour-based information to identify the character or word. One flaw in handwritten papers is that they are challenging to read through. In recent years, one of the difficult study areas in the realm of image processing and pattern recognition has been handwriting recognition. The task of handwriting recognition is difficult for a variety of reasons. The main explanation is that different writers have distinctive writing styles. The abundance of characters, including capital letters, small letters, digits, and special symbols, is a secondary factor. As a result, to train the system, a sizable dataset is needed. Since the system scans and detects static images of the characters, optical character recognition (OCR) is frequently referred to as an off-line character recognition method. We use the term "handwriting" to describe manuscript and cursive written writings. Because the characters are separated and written in block letters, manuscript-style texts are simpler to recognise. Cursive handwriting, on the other hand, joins the characters as they are written. To correctly perceive and recognise each individual character, handwriting recognition software is required. To create various text recognition algorithms that can be converted from paper format to electronics. The writing style is not constrained in a handwritten manuscript. The various human writing styles, variations in letter size and shape, and angles make it difficult to read handwritten alphabets. A branch of OCR technology called handwriting recognition, often known as handwriting OCR or cursive OCR, converts handwritten characters

into corresponding digital text or commands in real-time. These systems use pattern matching to recognise diverse handwriting styles to accomplish this goal. According to Wikipedia, handwriting recognition is: a computer's capacity to read and comprehend legible handwritten input from sources like paper documents, photos, touch-screens, and other devices.

II. LITERATURE SURVEY

An early notable attempt in the area of character recognition research is by Grimsdale in 1959. The origin of a great deal of research work in the early sixties was based on an approach known as analysis-by-synthesis method suggested by Eden in 1968. The great importance of Eden's work was that he formally proved that all handwritten characters are formed by a finite number of schematic features, a point that was implicitly included in previous works. This notion was later used in all methods in syntactic (structural) approaches of character recognition.

Hidden Markov Model (HMM) model is proposed for recognizing unconstrained offline handwritten texts. In this, the structural part of the optical model has been modelled with Markov chains, and a Multilayer Perceptron is used to estimate the emission probabilities. In this paper, different techniques are applied

However, even after applying all the said techniques might not possible to achieve the full accuracy in a preprocessing system.

III. METHODOLOGY

Pre-processing takes input image is to perform cleaning tasks. It effectively enhances the image by noise removal. Furthermore, images may be required to be in greyscale or binary formats which are done in this stage.

Segmentation

After the input images are pre-processed, individual characters are separated using a segmentation technique. These characters are then stored into a sequence of images. Then boarders in each character image are eliminated if the boarder is available. Next, individual characters are scaled to specific size.

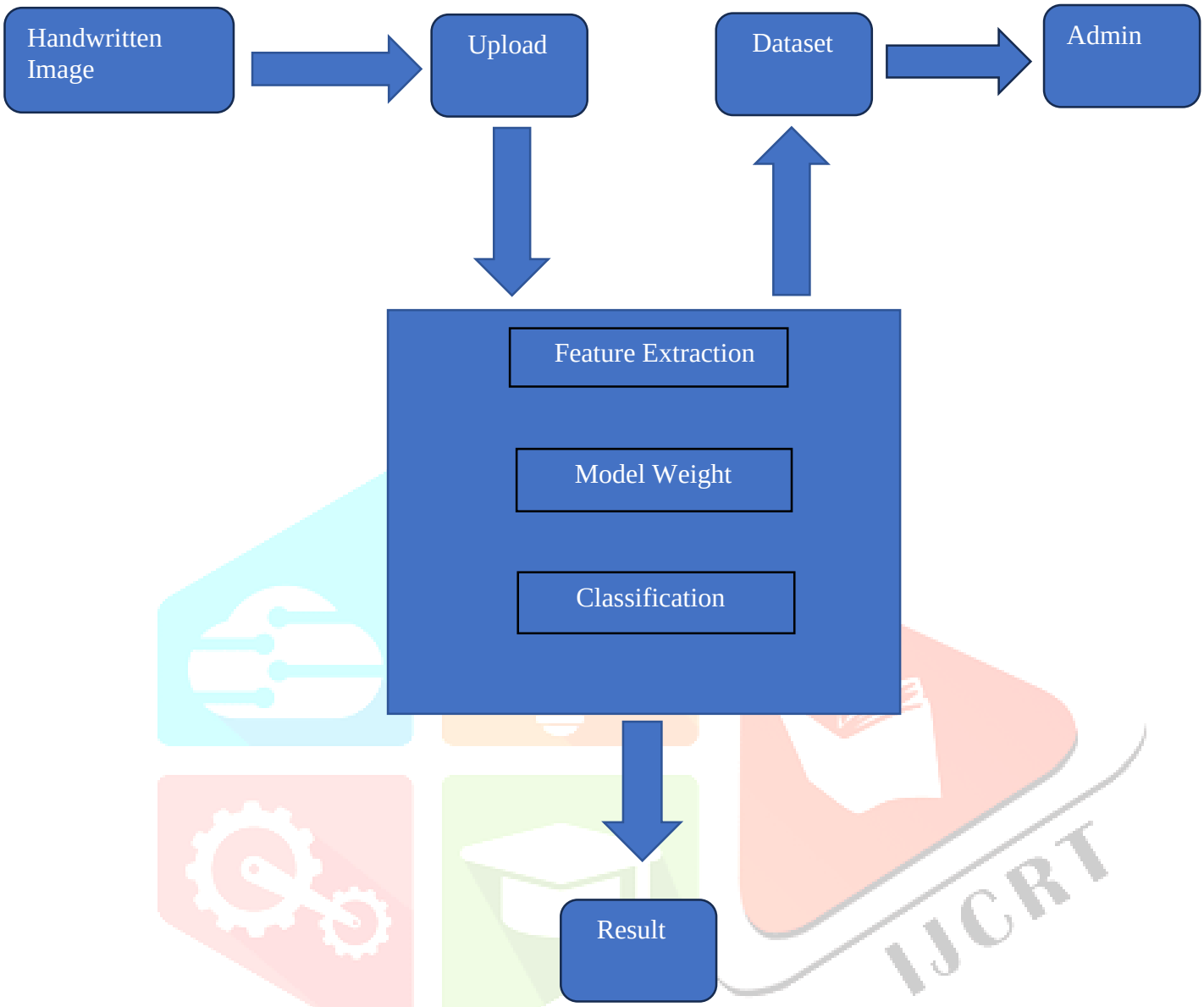
Feature Extraction

Feature extraction is made on the segmented characters. In our case, the features are extracted using CNN with ReLU activation function as shown in Figure 1. CNN works on each character image to form a matrix of reduced size using convolution and pooling. Finally, the reduced matrix is compacted to a vector form using the ReLU function. This vector is regarded as feature vector [5].

Classification and Recognition

The derived feature vector is used as individual input to formulate corresponding class. During the training phase, the parameters, biases, and weights are calculated. The calculated parameters, biases, and weights are used in the testing phase for classification and recognition purpose.

IV. SYSTEM DESIGN



Software requirements:-

- Image Processing Libraries*: Libraries for image processing are essential for preprocessing handwritten documents before recognition. Popular choices include:
  - OpenCV (Open Source Computer Vision Library): Provides functions for image processing, such as noise reduction, binarization, and deskewing.
  - PIL (Python Imaging Library): Useful for basic image manipulation tasks in Python.
- Machine Learning Frameworks*: Frameworks for building and training machine learning models are necessary for developing the recognition models. Common options include:
  - TensorFlow: TensorFlow offers a comprehensive platform for building deep learning models, including convolutional and recurrent neural networks for HTR.
  - PyTorch: PyTorch is another popular deep learning framework with extensive support for building neural network models.
- Deep Learning Model Libraries*: Libraries that provide pre-trained models or architectures for deep

learning networks can accelerate development. Examples include:

- TensorFlow/Keras: Offers pre-trained models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for various tasks, including image recognition and sequence modeling.
- PyTorch: Provides pre-trained models and architectures for tasks like image classification and sequence modeling, which can be adapted for HTR.

- Text Processing Libraries*: Libraries for text processing are useful for post-processing recognized text and performing tasks like spell checking and language modeling. Some options include:
- NLTK (Natural Language Toolkit): NLTK offers tools and libraries for natural language processing tasks such as tokenization, part-of-speech tagging, and syntactic parsing.
- SpaCy: SpaCy is another popular library for natural language processing, offering fast and efficient tools for tasks like named entity recognition and dependency parsing.

- OCR Libraries*: While not always necessary if using deep learning approaches, Optical Character Recognition (OCR) libraries can still be useful, especially for integrating legacy systems or working with specific document formats. Tesseract OCR is a widely-used open-source OCR engine that can recognize text from images.
- Data Management Tools*: Tools for managing and preprocessing datasets are crucial, especially when working with large volumes of handwritten text data. This may include tools for data cleaning, augmentation, and annotation.
- Version Control*: Version control systems like Git are highly recommended for managing codebase versions, collaborating with team members, and tracking changes over time.

V. TECHNOLOGIES

Handwritten text recognition (HTR) technologies encompass a variety of methods and tools aimed at converting handwritten text into digital format. Here are some of the key technologies commonly used in HTR:

- Optical Character Recognition (OCR)*: OCR is a fundamental technology that converts images of handwritten or printed text into machine-readable text. It involves segmenting the text into individual characters or words and then matching these patterns to a database of known characters or words.
- Feature Extraction Techniques*: Various feature extraction techniques are used to represent handwritten text images in a format suitable for machine learning algorithms. These techniques may include Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), or more modern techniques like deep feature extraction using pretrained CNNs.
- Handwriting Synthesis*: In some cases, handwriting synthesis techniques are employed to generate realistic-looking handwritten text from digital text inputs. This can be useful for tasks like creating handwritten documents or enhancing the visual appeal of digital interfaces.
- Hybrid Approaches*: Many HTR systems employ hybrid approaches that combine multiple techniques mentioned above to leverage their respective strengths and overcome their weaknesses.

These technologies are constantly evolving, driven by advances in machine learning, computer vision, and natural language processing, leading to continual improvements in the accuracy and efficiency of handwritten text recognition systems.

VI. WORKING

Handwritten text recognition (HTR) involves the process of converting handwritten text into machine-readable text. Here's a simplified overview of how it works:

- Preprocessing*: The handwritten document is scanned or photographed to create a digital image. Preprocessing techniques such as noise reduction, binarization (converting the image into black and white), and deskewing (straightening tilted text) may be applied to enhance the quality of the image.
- Feature Extraction*: Next, features are extracted from the preprocessed image to represent the handwritten text. Common features include line and word segmentation, which involve identifying the boundaries of lines and words within the text.
- Feature Representation*: Once the features are extracted, they are represented in a format suitable for machine learning algorithms. This representation typically involves converting the image data into a numerical format that can be understood by the algorithms.
- Training*: The feature representations are used to train a machine learning model. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or a combination of both are commonly used for this purpose. During training, the model learns to recognize patterns in the handwritten text data.
- *Recognition*: Once the model is trained, it can be used to recognize handwritten text in new images. This involves feeding the preprocessed image into the trained model, which then predicts the corresponding text.
- *Postprocessing*: Finally, postprocessing techniques may be applied to improve the accuracy of the recognized text. These techniques may include language modeling, spell checking, and context analysis.

The performance of a handwritten text recognition system depends on various factors including the quality of the input images, the effectiveness of preprocessing and feature extraction techniques, the choice of machine learning model, and the size and quality of the training data.

VII. FUTURE SCOPE

Education and Accessibility:

HTR can play a significant role in education by providing tools for digitizing handwritten notes, making them searchable and accessible. This can be especially beneficial for students with different learning preferences and accessibility needs.

Healthcare Applications:
Handwritten text recognition can be applied in healthcare settings for digitizing patient records, prescriptions, and other handwritten documents. This could streamline healthcare workflows and improve data accessibility.

Security and Forensics:
HTR can find applications in forensic analysis, examining handwritten documents for authenticity and identifying potential forgeries. This can be valuable in legal and investigative processes.

VIII. CONCLUSION

This project will be helpful in converting the handwritten text to a digital format which will be very helpful in converting old handwritten documents and notes to text files. These files are easy to store, edit, share and read easily. This project will also help in preserving old important documents which are difficult to store physically.

IX. REFERENCE

[1] Plamondon R and Srihari S N 2000 IEEE Trans. on Patt. Anal. and Mach. Intelligence 22 63.

[2] Kato N, Suzuki M, Omachi S I, Aso H and Nemoto Y 1999 IEEE Trans. on Patt. Anal. and Mach. Intelligence 21 258.

[3] Islam M M, Tayan O, Islam M R, Islam M S, Nooruddin S, Kabir M N and Islam M R 2020 IEEE Access 8 16611.

[4] Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta "Offline Handwritten Character Recognition Using Neural Network" International Conference on Computer Application and Industrial Electronics 2011 (ICCAIE-2011)