Handwritten Text Recognition Using Image Processing By Machine Learning

Anuradha[1] Ms. Nidhi Sengar[2]

Student[1] Assistant Professor[2]

Department Of Information Technology

Maharaja Agrasen Institute Of Technology, Rohini Delhi India.

Abstract

Machine Learning is the field that is frequently evolved with the growing application in computer science. The handwritten text recognition is still one of the problems for the researcher because of its substantial variation in appearance. Still, there is a need to narrow down the gap between the reading capabilities of humans and machines. Handwritten text recognition is the process in which the handwritten text is converted into a machine readable form. Handwritten text recognition is a difficult task as every person has a different style of writing, and even the same person writes the same character with style, whether there is a difference in shape, size of the position of the character. Handwritten text recognition is a popular field of research because of its wide range of applications like digitizing ancient articles, postal address processing, bank cheque processing, application form processing, and many other fields. This paper introduces the recognition of handwritten text in the English language using Tesseract. Many approaches have been proposed for better recognition of text. In this paper, we have discussed in detail about handwritten text recognition using a recurrent neural network.

Keywords: -

Handwritten Character Recognition, Neural Network, Feature extraction, Image processing

I. Introduction

Machine learning is defined as the branch of artificial intelligence that deals with first learning from a dataset and then applied to solve the problems. The supervised machine learning model is given instances of data to a problem and an answer that solves the problem for those instances. When the machine completely learns the facts, the model is then able to provide answers to the data it has learned and also give answers to unseen data with high precision [4].

In this modern world of technology, Handwritten text recognition is one of the most challenging research areas in the field of image processing. It contributes to the advancement of an automation process and helps in improving the interface between man and machine in so many applications. There are several ongoing research that focuses on new techniques and methods that reduce the processing time and provide higher recognition accuracy [1].

In today's world, handwritten characters are increasingly used in daily life. Handwritten information comes in a variety of different forms, including bills, manuscripts, documents, forms, and photographed documents. Handwritten character recognition has wide application prospects, and there is a great demand for it in industrial fields such as image recognition systems and handwritten text input devices as society develops and progresses.

The main problem in the handwritten text recognition system is that there is a lot of variation in the handwriting styles that are completely different for different writers. The purpose of the handwritten text recognition system is to make a user friendly computer assisted character representation that will successfully extract characters from handwritten documents and to digitalize and translate the handwritten text into machine readable form.

Handwriting text recognition is classified into two types as off-line and on-line handwriting recognition methods.

- In off-line recognition, the written documents are usually captured optically by a scanner, and the whole document is available as an image.
- In on-line recognition, the written document's recognition is performed when text or characters are under creation.

Optical character recognition (OCR) is defined as an off-line character recognition process in which the system scans the documents and then recognizes static images of the characters of the text documents. It is referred to as the electronic translation of images of handwritten text or printed text into machine code without any variation in the text.

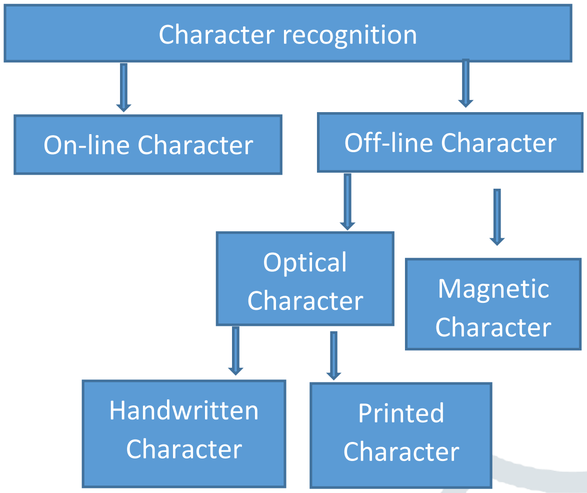


Fig 1. Classification of character recognition

II. Literature Review

This review was accepted by surveying the research in the last 5 years for collecting information about some major issues. The numerous research papers were reviewed to understand handwritten text recognition techniques.

Amir Bahador Bayat[2013][8] Recognition of handwritten characters automatically has been a goal of many research efforts in the pattern recognition field and text recognition field. This paper discussed the design of a highly efficient system for the recognition of handwritten numerals. Firstly, it proposes an efficient system that contains two modules: one is the feature extraction module, and the other is the classifier module. In the feature extraction module, the seven sets of discriminative features are discussed and implemented in the text recognition system. In the classifier module, the adaptive neuro-fuzzy inference system (ANFIS) is discussed. The results show that the proposed system has better accuracy for recognition.

In K. H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V. S. Chakravarthy, Sriganesh Madhvanath, [9], a technique was proposed to perceive the handwritten character in Tamil by utilizing the grouping in the strokes. A strokes' format or shape-based portrayal is utilized taken as a string of shape highlights. Utilizing this technique, the unrecognized stroke was perceived by contrasting it and a dataset of strokes by the string coordinating method in an adaptable mode. An individual character was perceived by distinguishing one of the strokes and its segments.

R. Bajaj, S. Chaudhari, L. Dey, et al. [10], In this paper for grouping the Devanagari digits, distinctive highlights like thickness, clear part, and minute highlights were utilized. Additionally, to increase the recognition capability of the system, the paper proposes multi classifier unwavering quality for handwritten Devanagari digits.

III. The Proposed Recognition System

In this section, we discussed the proposed recognition system. Handwritten Character Recognition System consists of the following stages:

1) Image Acquisition

2) Pre-processing.

3) Segmentation.

4) Feature extraction.

5) Classification and recognition.

6) Post-processing.

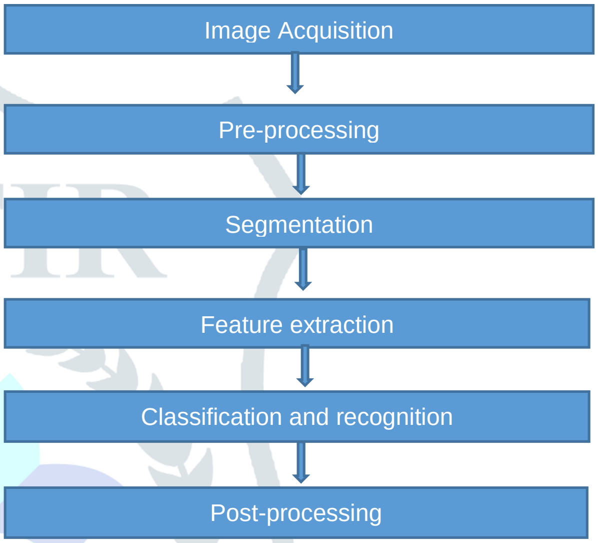


Fig 2. Block Diagram of the HCR system

Image Acquisition

In Image acquisition, the scanned image of handwritten text is acquired by the recognition system as input. The image must have a specific format like JPEG, BMT, etc. [1] This image is acquired through a scanner, digital camera, camscanner, or any other suitable digital input device, or one can draw on the canvas provided on the user interface.

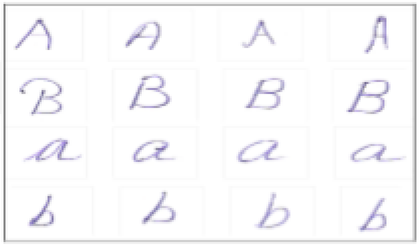


Fig 3. Sample Dataset

Pre- Processing

Preprocessing is defined as a series of operations performed on the scanned text image used as an input image. It enhances the scanned image and makes it suitable for further processing [2]. Preprocessing works to normalize the strokes and deletes all those variations that can reduce the rate of accuracy during the recognition of text. Preprocessing works on the different distortions like the irregular font size, points missed during the pen movement, jitters, left-right bend, and uneven spaces between the words and characters [3]. In our project, some of these methods are mainly used when we recognize the text from an image, but some of these methods, such as cropping the written text and scaling it to our specified input size, are also done in the touch mode.

The different tasks that are performed on the image in the pre-processing stage are noise removal, binarization, skew correction, Edge detection, Thresholding, Normalization, etc.

To preserve character strokes, we have used a non-linear operation in our project, which is called median filtering. In this operation, the median filter replaces the value of pixels with the median of the intensity of the surrounding pixels. The result is shown in Figure 4 [4].

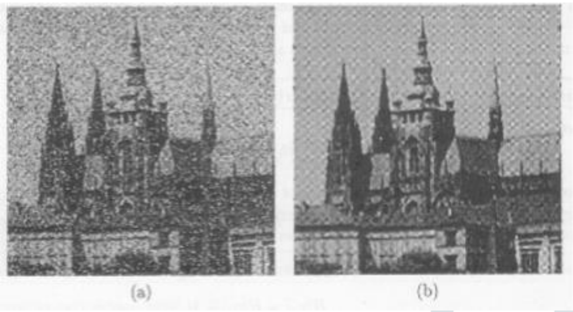


Fig 4. Median filtering: (a) image that is corrupted with noise (b) median filtering applied to the input image

Segmentation

The Segmentation stage is the most important stage of the whole process. Segmentation is the process in which the various characters of the scanned images get separated from each other. Handwritten character segmentation into different zones (upper, middle, and lower zone) is more difficult than the printed documents that are in the standard format. This happens because of the variations in the paragraph, characters of the word, slant, skew, size, and words of a line. Because of modified characters in the lower and upper zone, Overlapping and touching problems occur very frequently [5].

The Segmentation includes:

- Line segmentation: It is the separation of the line from the paragraph.
- Word segmentation: It is the separation of words from the line.
- Character segmentation: It is the separation of character from words [2].

Feature Extraction

Feature Extraction is defined as the process of retrieval of the most important data as per the user's need from the raw data. The main aim of feature extraction is to extract a set of features, which maximizes the rate of recognition with a small number of elements. It becomes a very difficult task due to the nature of handwriting with its high degree of variability and imprecision obtaining these features. Feature extraction methods are mainly based on 3 types of features- Statistical, Structural, Global transformations, and moments [6].

Classification And Recognition

The Classification is defined as the process of assigning labels (categories, classes) to unseen observations. In machine learning, it is done on the basis of firstly train an algorithm on a set of training examples [4]. The classification stage uses the features extracted in the last stages.

Classifiers are based on two types of learning methods: Supervised and Unsupervised Learning.

Post-Processing

Post-Processing is the last phase of character recognition. By using natural language, in this process, we can correct the misclassified output of the scanned documents [3]. In this stage, to perform syntax analysis and semantics analysis kind of higher level concepts, we connect the dictionary to the system to check the recognized characters and to increase the accuracy of the recognition [2]. Once the shape has been recognized, it processes the output. For different handwriting, Shape recognizers behave differently. The Post-processing stage is not compulsory in the HCR system.

IV. Result and Discussion

In our project, we recognize the handwritten text by using Tesseract and OpenCV EAST text detector model.

The results of testing are given in three different forms.

- BLOCK shows what the text extraction is done on the whole image.
- LINES show the different lines that are in the document vertically from top to bottom.
- WORDS shows all the different words that are used in the document.

By comparing the results with the original image, we can understand where our text detection, recognition, and extraction has failed.

The document to be scanned is

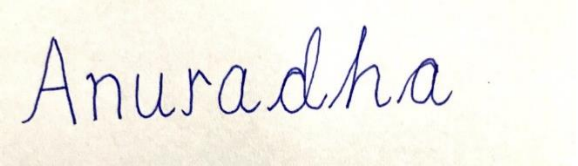


Fig 5. Test document 1

So the result is

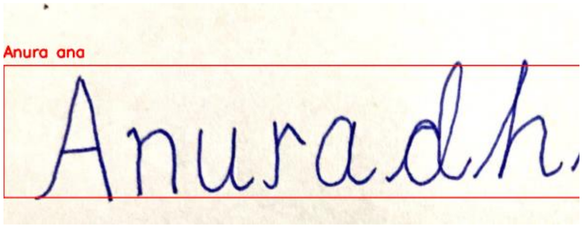


Fig 6. Result Image 1

It is evident from the results that text recognition is difficult in this type of condition.

From the resultant images, we can see that the text recognition failed to recognize “d” and “h” in the word.

In our next example, we show the importance of adding padding in certain situations:

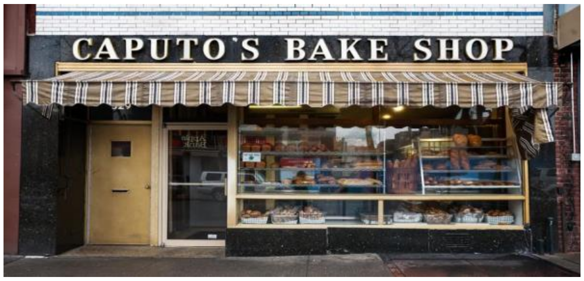
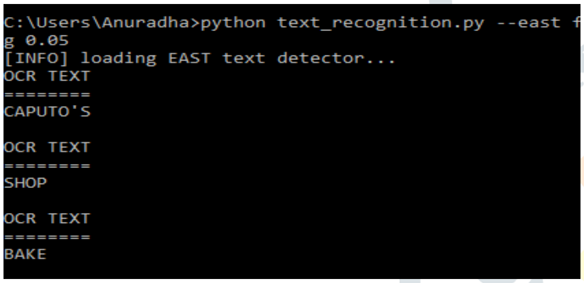


Fig 7. Test document 2

In the first attempt of OCR’ing, we see that “SHOP” is correctly OCR’d, but:

1. The letter “U” in “CAPUTO” is incorrectly recognized as “TI” in the output.
2. The apostrophe and the letter “S” is missing from the word “CAPUTO’S’.

   And at last, “BAKE” is incorrectly recognized as a vertical bar/pipe (“|”) with a period (“.”) [7].



```
C:\Users\Anuradha>python text_recognition.py --east f
g 0.05
[INFO] loading EAST text detector...
OCR TEXT
========
CAPUTO'S

OCR TEXT
========
SHOP

OCR TEXT
========
BAKE
```

Fig 8. Result Image 2

By adding padding of just 5% surrounding each corner of the bounding box, we’re not only able to correctly OCR the “BAKE” word but also able to recognize the letters “U” and “’ S” in “CAPUTO’S”.

V. Conclusion

This paper gives a detailed review of text recognition in the English language using Tesseract and studies so many algorithms for recognition. The accuracy of text recognition fully depends on the nature of the image to be read and by its quality. Current research does not deal with the cursive handwriting and also not able to recognize child handwriting because it requires a high supervised system. This review implements a system that converts scanned images of handwritten text documents to digital text documents. In this paper, we have studied numerous papers with different algorithms to increase the accuracy of the result.

Apart from these, we have concluded that OpenCV OCR performs both text detection and text recognition. To complete the task of recognition of text, firstly, we utilize the OpenCV EAST text detector that enables us to localize the region of text in the scanned input image. From there, we extract the text ROI and apply text recognition using Tesseract and OpenCV.

After going through several test results, we concluded that for better text recognition results using OpenCV, our input ROIs must be cleaned and preprocessed as much as possible. Also, our text has been captured at a 90-degree angle from the camera.

VI. Future Work

In the future, we need to improve the current performance of our project. We can use some other techniques to the most confusing characters, to increase the accuracy of recognition of text. For higher accuracy, we need to try one of the “big 3” computer vision API services like Google Vision API OCR Engine, Amazon Rekoginition, and Microsoft Cognitive Services, which uses more advanced OCR methodology running on very powerful machines in the cloud.

References

1 J.Pradeep, E.Srinivasan and S.Himavathi, , Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network,Department of ECE, Pondicherry College Engineering, Pondicherry, India

2. Monica Patel, Shital P. Thakkar, Handwritten Character Recognition in English: A Survey, Department of Electronics and Communication, Dharmsinh Desai University, Nadiad, Gujarat, India

3. Megha Agarwal, Shalika, Vinam Tomar, Priyanka Gupta, Handwritten Character Recognition using Neural Network and Tensor Flow, Computer Science and Engineering, SRM IST Ghaziabad, India.

4. Mgr. Ľudovít Malinovský,” Handwritten Character Recognition Using Machine Learning Methods”, Comenius University In Bratislava Faculty Of Mathematics, Physics And Informatics. 2013

5. Er. Neetu Bhatia, Optical Character Recognition Techniques: A Review, Kurukshetra institute of Technology & Management Kurukshetra, India

6. Vijay Laxmi Sahu, Babita Kubde, Offline Handwritten Character Recognition Techniques using Neural Network: A Review, 1Rungta College of Engineering & Technology Bhilai, Chhattisgarh, India – 490021

7. Adrian Rosebrock on September 17, 2018 in Deep Learning, Optical Character Recognition (OCR), Tutorials.

8. Amir Bahador Bayat Recognition of Handwritten Digits Using Optimized Adaptive Neuro-Fuzzy Inference Systems and Effective Features Journal of Pattern Recognition and Intelligent Systems Aug. 2013, Vol. 1

9. K. H. Aparna,, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V. S. Chakravarthy, Sriganesh Madhvanath, "Online Handwriting Recognition for Tamil", IWFHR, 2004, Proceedings. Ninth International Workshop on Frontiers in Handwriting Recognition, Proceedings. Ninth International Workshop on Frontiers in Handwriting Recognition 2004, pp. 438-443, doi:10.1109/IWFHR.2004

10. Reena Bajaj, Lipika Dey, and S. Chaudhury, “Devnagari numeral recognition by combining decision of multiple connectionist classifiers”, Sadhana, Vol.27, part. 1, pp.-59-72, 2002.