Intelligent Character Recognition- Character detection using Neural Networks

P. Giri Kishore[1], Athira Ajayakumar[2] , N Nitin[3] and Arun Natarajan[4]

[1]*Department of Computer Science and Engineering New Hoirzon College of Engineering, Bangalore, India,*

[2]*Department of Electronics and Communication Engineering, New Hoirzon College of Engineering Bangalore, India,*

[3]*Department of Mechanical Engineering, New Hoirzon College of Engineering Bangalore, India*

[4]*Chief Engineer QuEST Global, Bangalore, India*

Abstract *— This paper highlights how EMNIST database can be put to use, to create clean and synthetic images of texts in different handwriting styles. This paper presents a detailed review of Intelligent Character Recognition, the methods using which we can classify character by detecting and extracting the position of a character from synthetic images. With the help of the image processing, the raw image is cleaned, before it is sent for classification, so as to give a higher probability of successful recognition of characters.*

Keywords — *OCR, Feature Extraction, Segmentation and training.*

I. INTRODUCTION

Data storage has become an integral part of this era. There are a lot of documents that need to be stored for future reference, for instance, newspapers. One of the methods of storing data is by scanning the paper documents. However, we cannot reuse or reform the information present in the scanned document as these scanned documents are stored as images. This problem gave rise to software systems to recognize characters.

Just like how we recognize characters by size, shape, colour and other factors, this software system also recognizes the same factors and predicts the character as an output. OCR - Optical Character Recognition is one such technique that gives the power to the system to extract data from images and make it computer editable. It may seem simple, but it is a tedious task for the software to read the individual characters line-by-line or word-by-word or character-by- character. As the font features of the characters in paper documents are different to the font features of the characters in a computer system, it is difficult for the computer to recognize the character. The computer isn‟t rapid enough to predict as it reads the document. Instead, the document is scanned and stored. When required, the system reads the scanned documents to extract data. This process is called Data Processing and the software system required for it called Character Recognition System.

Although there is many existing software that can perform OCR, it still does not help much when it comes to deriving information from old documents or documents from constructions, old records etc. The tainted nature of the paper and the fading of the text make it very difficult.

Segmentation, Feature Extraction, Classifications and Recognition. Thus our need is to develop a character recognition system which does document analysis and picks the characters from an image and converts them into editable format. For this process we have chosen for the text to be read. Engineering drawings have various symbols or acronyms which are often predicted wrong by the system. This project aims to aid the digitization of documents, especially in the field of engineering.

II. PROJECT DEFINITION

Firstly, for the purpose of this project, we try to convert a text file into a synthetic image of handwritten letters and digits by using the EMNIST (extension of Modified Institute of Standard and Technology database), which is a database of 800,000, 28*28 pixel characters. This image is taken as an input image for recognizing the characters then it will recognize input character which is given in image. Recognition and classification of characters are done by Neural Network or a LSTM (Long short term memory) network. The ultimate aim of this project is to recognize the characters present in the documents such as engineering drawings efficiently using basic image processing techniques for recognition and neural network for classification. The texts in the scanned images are nothing but the tiny dots of pixels which are not editable. This system analyses the image and derives text from the images based on the pattern of pixels in the image.

III. PURPOSE

The main purpose is to perform document image analysis to process electronic documents into paper format. The documents we are emphasizing on are of that of the engineering domain. For instance, plans for a civil structure. The connotations used in

these drawings must be used to train our machine and predict them accurately. The primary objective is to speed up the process of character recognition to detect and classify characters in document processing. As a result the system can process a huge number of documents with-in less time and hence saves time.

It aims to recognize any character that belongs to different formats with different font properties and alignments. Intelligent character recognition (ICR) is usually referred to as an off-line character recognition process to mean that the system scans or accept an image as an input and recognizes images of the characters. It refers to the translation of images of printed, typewritten or handwritten text into machines- editable text. ICR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. Thus our need is to develop a character recognition system which does document analysis and picks the characters from an image and converts them into editable format. For this process we have chosen Optical Character Recognition as the main fundamental technique to recognize characters.

IV. LITERATURE SURVEY

A. Offline Handwritten English Numeral Recognition using Correlation Method:

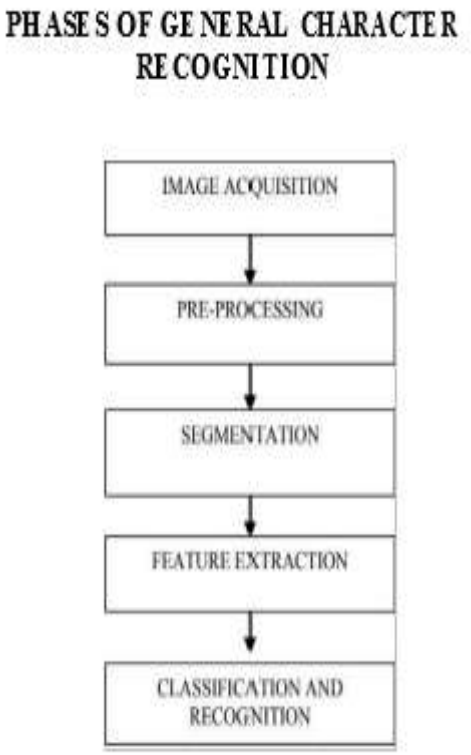
The author of this paper has proposed a system which efficiently recognizes offline handwritten digits which much higher precision than others. Also the previous handwritten digit recognition algorithms are concerned with single digits. This paper is focused on clear segmentation for isolating digits.

B. INTELLIGENT SYSTEMS FOR OFF-LINE HANDWRITTEN CHARACTER RECOGNITION:

Handwritten character recognition is always a frontier area of research in the field of pattern recognition and image processing and there is a large demand for Optical Character Recognition on handwritten documents. This paper provides a comprehensive review of existing works in handwritten character recognition based on soft computing technique during the past decade.

C. Recognition for Handwritten English Letters:

Character recognition is the most challenging and interesting research areas in the field of Image processing. English character recognition has been extensively studied in the last two decades or so. Nowadays there are different methods for character recognition. Document verification, digital library, reading bank deposit slips, reading postal addresses, extracting information from cheques, data entry, applications for credit cards, health insurance, loans, tax forms etc. are application areas of digital document processing. This paper gives an overview of research work carried out for recognition of handwritten English letters. In Hand written text there is no constraint on the writing style. Handwritten letters are difficult to recognize due to diverse human handwriting style, variation in angle, size and shape of letters. Various approaches of handwritten character recognition are discussed here and comparison of its performance is also made.



A. Image Acquisition

In this phase, the input image taken through the camera or scanner. The image should have a specific format, such as JPEG, TIF, and BMT etc. The input captured may be in gray, color or binary from scanner or digital camera. Undoubtedly, it is a tedious process to write in different handwriting styles on paper and scan them for training. Instead, synthetic images for training can be produced. For this purpose, we make use of a readymade dataset called the EMNIST Database that consists of different handwriting styles (More than 3000 styles). It has over 800,000 images with hand checked classifications. All the characters/images are of 28x28 pixels, which is usually the input for all standard neural networks for classification. The dataset is called in the program in which the input is a text file that consists of words/ sentences. The output of the program is the image (in JPEG format) of the characters, written in different handwriting styles taken from the EMNIST Database. The program is written in Python. We make use of OpenCV (Open Source Computer Vision), which is a highly optimized library for programming functions.

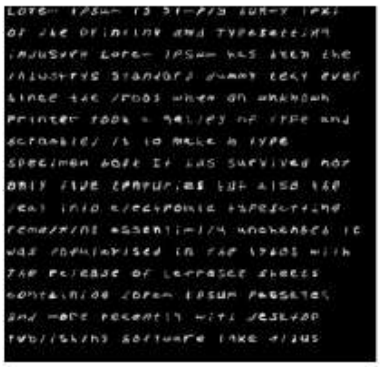


FIG 1.1 - OUTPUT IMAGE OF THE PROGRAM

FIG 2.1.1- BLACK TEXT ON WHITE BACKGROUND

B. Pre-processing

Before training, it is always better to process the image, so that the OpenCV functions can be put to use in an efficient manner. The function arguments states how the image should be before giving it as an input to the training model. Hence, pre-processing is a series of operations performed on the image. It essentially enhances the image rendering it suitable for segmentation.

a) Thresholding

Since usually a lot of characters need to be predicted, processing speed is an important factor when it comes to text recognition. Now that we use OpenCV library for machine learning, it is favorable to convert an RGB image as binary image by fixing a threshold value. If the pixel value is above the threshold, value is set to 1, otherwise 0. One type of thresholding in Global Thresholding, which fixes a threshold value for the entire image, according to the background from the intensity histogram of the image. The other thresholding is called Adaptive Thresholding, in which different threshold values are used in different regions of the image. Based on the type of image, you can choose the type of thresholding. The latter thresholding is preferred as color images are usually set as an input to the model. First, the image (Fig 1.1) is inverted, such that the font is black and background is white, as shown in Fig 2.1.1.

Then, the image undergoes a process of dilation. **Dilation** is essentially adding pixels to the characters, such that the characters look thicker/ **Contours detection** is a **process**, which can be explained simply as a curve joining all the continuous points (along with the boundary), having same colour or intensity. The **contours** are a useful tool for shape analysis and object **detection** and recognition. This process makes it easier for Line Detection. After contour detection, dilation of the image takes place.

C. Segmentation

When it comes to text recognition, Segmentation is one of the important techniques to detect characters, words and lines. It is generally used to distinctly separate characters in a word. With the help of dilation and contour detection, segmentation can be easily done on handwritten text and printed text. Although it does get difficult to segment, when the characters touch each other or gets overlapped. Segmentation is of two types: External and Internal. The image after Dilation is shown in Fig 3.1.

After dilating the image, it is easy to detect lines. Detection of lines is represented by a box, enclosing the characters that are in the same line, as shown in Fig 3.2. Dilation can also be done on words in order to detect words. The dilated image for words is shown in Fig 3.3. Thus, words can also be detected as depicted in Fig 3.4.

In the same fashion, characters are detected using dilation as depicted in Fig 3.5 and Fig 3.6.

Fig 3.1

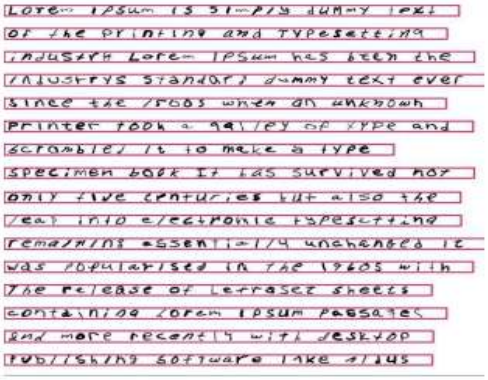


Fig 3.2

Fig 3.3

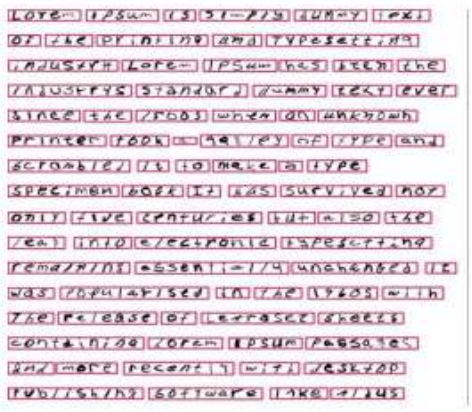


Fig 3.4

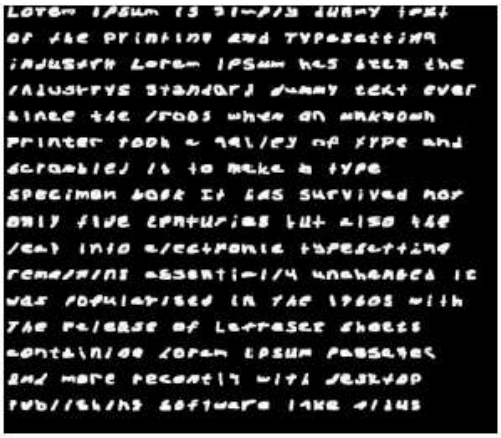


Fig3.5

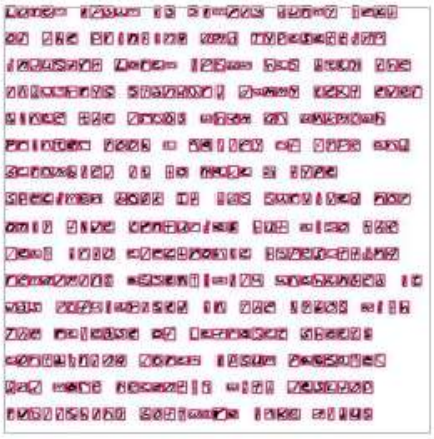


Fig 3.6

Therefore, by making use of dilation and contours, we can extract sentences, words and characters, One disadvantage that was noticed while extracting the characters was that the images of the characters weren't of the standard 28x28 pixel size. It is necessary that the images are of the standard size as the input of the training model takes 28x28 images. Hence, the program also makes sure that.

D. Feature Extraction

Feature extraction means retrieving the most important feature from raw data. In our case, the important features are the characters. We achieve this by masking all other characters from the images. We convert the entire image except the character into a contrast image and then extract this character as a single image. This way we get multiple images are all the characters present in the image. The major problem is that these images are not in order. So, we have to explicitly sort all the images based on y coordinate first then by x coordinate. Sorting by y coordinate gives us all the lines from top to down. On each line, we sort the words from left to right using the x coordinate. Thus we obtain words from left to right.

Once that is done, we can use the sort function provided by OpenCV to get individual images of each character in order.

E. Classification and Recognition

The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage to classify into one of the classes. The individual images are given as input into the Convolutional Neural Network (CNN). We classify the images by using K nearest neighbor algorithm or Support Vector Machine algorithm. We have 62 (A-Z, a-z, 0-9) classes in our classifier. Though this system only detects alphanumeric characters, we can extend it by producing a dataset with such symbols and special character. The input to this model must be a 28*28 pixel alphanumeric character, as the EMNIST dataset has images that are 28*28 pixels.

An alternative approach is to use Recurrent Neural Network (RNN) as their paths are directed cycles and this makes the network learn dynamically and store what's learnt to predict.

V. CONCLUSION

In this paper we have had a detailed review of how to form synthetic images from available EMNIST dataset containing over 800,000 images. We use this synthetic image to extract and classify characters. It is hoped that this paper was of benefit to advance studies in this area. Though this paper detects characters from the synthetic images created from EMNIST dataset, it works equally well on scanned images of handwritten digits and alphabets. The use of Recurrent Neural Network (RNN) over Convolutional Neural Network (CNN) as the special ability of RNN is vital when dealing with sequential data.

VI. FUTURE WORK

This topic still needs a lot of research for exploiting ways to eliminate noises from noisy images. We also need to find ways to improve current performance, by predicting the confusing characters based on the previous and the next character and increase the recognition rate. We can also try to detect the symbols and special characters by creating a dataset and following the mentioned steps. We can also extend this work in converting engineering drawings from an image to an editable format.

REFERENCES

[1] ISha Vats, Shamandeep Singh, "*Offline Handwritten English Numerals Recognition using Correlation Method*", International Journal of Engineering Research and Technology (IJERT): ISSN: 2278-0181 Vol. 3 Issue 6, June 2014. Access Date: 09/07/2015.

[2] Shabana Mehfuz, Gauri katiyar, „*Intelligent Systems for Off-Line Handwritten Character Recognition: A Review*", International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 4, April 2012. Access Date: 09/07/2015.

[3] Nisha Sharma et al, "Recognition for handwritten English letters: A Re-view "International Journal of Engineering and Innovative Technology (IJEIT) Volume2, Issue 7, January 2013. Access Date: 09/07/2015. Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: an extension of MNIST to handwritten letters. Retrieved from http://arxiv.org/abs/1702.05373.