Optical Character Reader for Blurry, Broken and Plain Text

Shivam Jaiswal[1], Arinjay Jain[2], Nallakaruppan M.K.[3]

[1,2,3,]School of Information Technology & Engineering, Vellore Institute of Technology, Vellore, 602014, India

Abstract- In a fast paced world that we live in today, there is a lot of scenarios that may arise where any text image we take is not up to par- it may be blurred. There may also be situations where the text that is physically present being photographed is not clear, it be physically damaged or defaced. That is where our product will be useful. We will use Optical character recognition (OCR) to give a clear text file of the unclear image.

OCR is a process of getting information from optical patterns present in a digital image. The text that is obtained from the image is done through a series of procedures, namely segmentation, feature extraction and classification. In this paper, we have taken broken and blurred text, applied OCR to it and classified the digital picture into words on a text file as an output. We have used OCR system techniques such as optical scanning, location segmentation, pre-processing, segmentation, representation, feature extraction, training and recognition and post-processing of the image and finally write it into a text file for the user to read.

Index Terms- OCR · Image Segmentation · Image Blurring · Image Sharpening · Image Acquisition

I. Introduction

In today's world, it is imperative to have machines that can recognize patterns and process them so that we can use it in an efficient manner. For example, optical character recognition, facial recognition, finger print recognition and more making great strides in different fields we can see the usefulness of pattern recognition by machines.

Optical character recognition is popular research field which attempts to develop programs and different software that can extract and classify text present in images. Nowadays there is a high need of storing information in a databases- both electronic as well as on computer disks- from sources that are present either in printed documents or in handwritten documents, and then later access and efficiently process and use the data through computers. Currently, inefficient methods of storage are utilized, such as taking images of the handwritten and printed material and storing it as image files. This is proper storage but later efficient access of the files will be very slow and will prove to be difficult to get information. Hence, the need for automatic processing of images, extracting the text from them and storing the text in a computer readable format arises. That is where optical character recognition comes in. However, this is not an easy task and it comes with its own share of challenges such as font characteristics, excessive blurriness of the image, presence of noise in the image – unwanted elements present in image that interfere with recognition- along with skewness. The presence of these challenges, the accuracy of character recognition may not be very high, as the computer

will recognize the characters incorrectly. Thus, we need to analyze the image correctly and produce an electronic document with the correct output. Optical character recognition is the process of transformation of any text present in handwritten, scanned or printed formats and images into the corresponding editable digital format that can be processed further and be stored in an electronic database. Optical character recognition allows computers to automatically recognize the text from images and replicate them into a computer readable format that can be processed and efficiently accessed later. We can compare the Optical character recognition to be similar to a human beings ability to view and process text that is seen in our everyday life. The ability of the brain to process the text is directly correlated to the quality of the image that is seen by the human eye and taken as input. During implementation of any OCR software there are challenges that are faced. For example, a very slight difference between some digits or characters may lead to the computer being unable to correctly identify and distinguish them accurately. For example, it is not very easy for the computer to easily distinguish and correctly recognize digit "1" from character "l", or it may have problems in correctly recognizing lower case characters and distinguishing them from upper case characters, such as "v" from "V", this becomes exponentially harder with the presence of noise in the input image. Another important focus of the OCR research being done is the recognition of font that is in cursive scripts, or handwritten text that is not very legible as they have a huge application of OCR as well. Normally an OCR system documents can be divided into 3 groups: Mono-font, Multifont, and Omni-font[2].

OCR research is an active and important pattern recognition field that requires comprehensive reviews at regular intervals to keep track of breakthroughs being made in the field. This paper attempts to further elaborate on earlier literature surveys by discussing the important and major challenges faced by OCR along with all of the phases that take place such as image preprocessing, image segmentation, text extraction, text detection and text recognition. With the implementation of these phases, we can get a computer readable document format that can be further processed and efficiently accessed from text present in various images. In the last section of the paper, different real world applications of OCR have been discussed.

II. CHALLENGES

Font format– The algorithm implemented might encounter problems while dealing with various fonts. For the algorithm to work properly it needs to be trained with numerous sets of fonts that are available. Some of the font styles might overlap each other that might result in difficulty during the word segmentation stage. Italic variants of the font might give us errors and some other characters will cause difficulty thus hindering accurate recognition of the text. Although the biggest challenge will be the enormous amount of new texts being created every day. It will take a lot of time to train the datasets for those (including their italic variants and numbers).

Excessive Bluriness-

Excessively blurry text with very high amount of blurriness results in difficulty when we apply sharpness to the text. The algorithm would reduce the blurriness to some extent, but getting the image of the text with full clarity is quite laborious. Characters which are similar to each other will generate different characters after cleaning or sharpening the image hence resulting in errors.

Noises -

In a text image of a real world scenario where text is to be detected, there will be a lot of unneeded elements in the image, called “noise”. Some of the most common noises present are salt and pepper noises. To avoid these kinds of noises we have applied median filter to the algorithm which would reduce the noises to some extent.

Ordering -

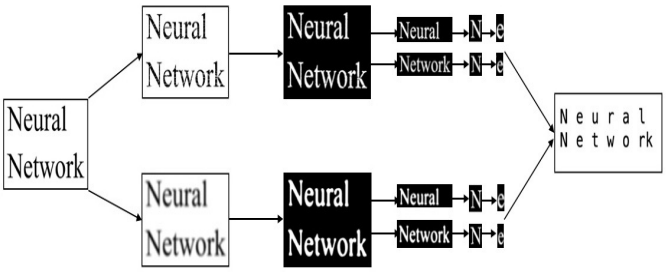
Usually when working with OCR, the text is arranged in a defined order by the process of association. In association, the text is broken into small rectangular segments by individual characters and then linked to their neighboring characters. In our algorithm the text is read line by line, followed by reading of each word of that particular line, then reading of every character of the word to process the information required.

Skewness -

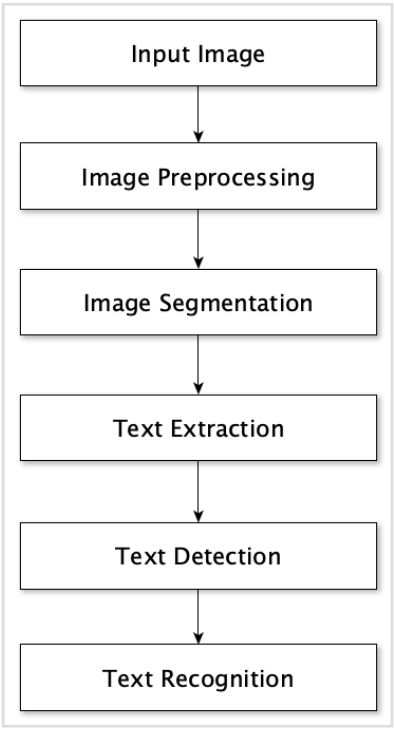
Skewness is the problem which occurs with the orientation of an image. There are two major types of skewness that are present and they can be classified as 2D skewness, which refers to the orientation of the text block in a 2D plane, and 3D skewness where the text block is oriented in a 3D plane. Only 2D skewness is rectified in our algorithm as dealing with 3D skewness is a tough and tedious process.

III. PROCEDURE

Most crucial part about this project is image segmentation, We are segmenting sentence from image then word from the sentences and at the last character from the word. After all this segmentation we are sending the image of segmented character to the neural network for the output.

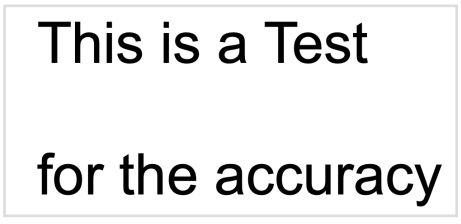


IV. METHODOLOGY

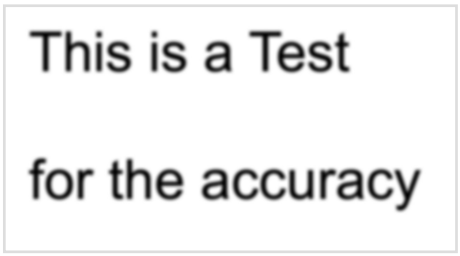


Here we divided the process in each step of the flowchart

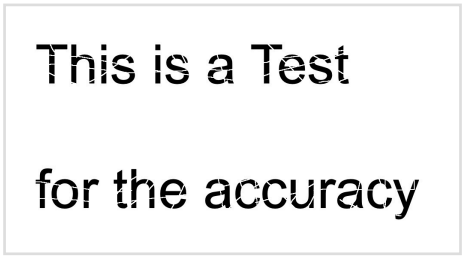
Input Image: The input image can be any one of the following- broken, blurred or normal plain text.



(a)Plain Image

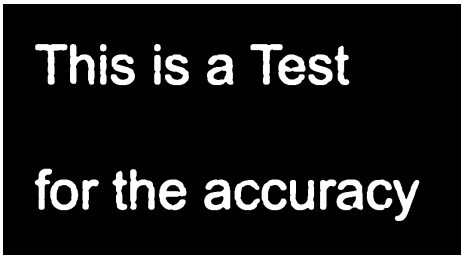


(b) Blurred Image

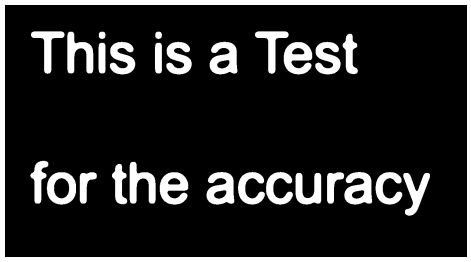


(c) Broken Image

Image Preprocessing: Image enhancement is a process that changes the pixel's intensity of the input image, so that the output image will look subjectively better[4].We will test whether the image that we took as input is blurred or broken. If the text broken, we will apply Gaussian blur to it so as to eliminate the spaces inside of every individual character and then apply sharpening to the image. In case it is not, we will sharpen the image directly. We will be using the convolution mask of 5x5.



(a) Broken Image

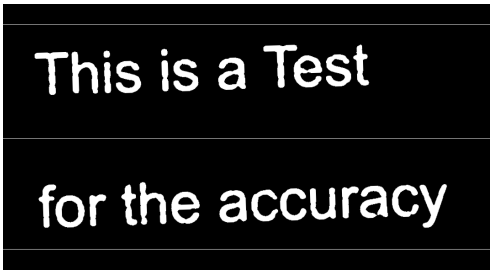


(b) Blurred Image

Image Segmentation: We will first segment each individual line of the image text. Following that we will apply segmentation and get each individual word from every line of the text.

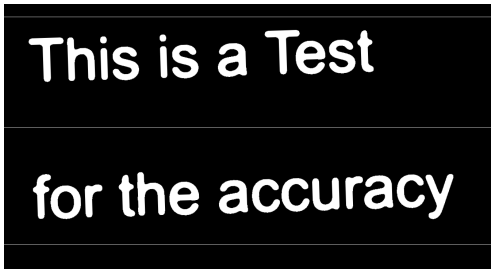
(a)Broken text

(a) Line segmentation



(b) Word segmentation

(b) Blurred Image

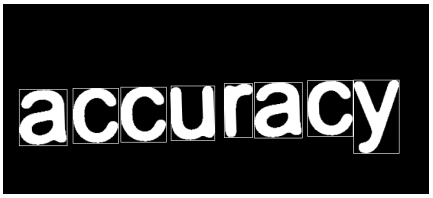


(a) Line segmentation

(b) Word segmentation

Text Extraction: We will perform further segmentation on every word to get individual characters thus extracting the text required from the image.

(a)Broken Image

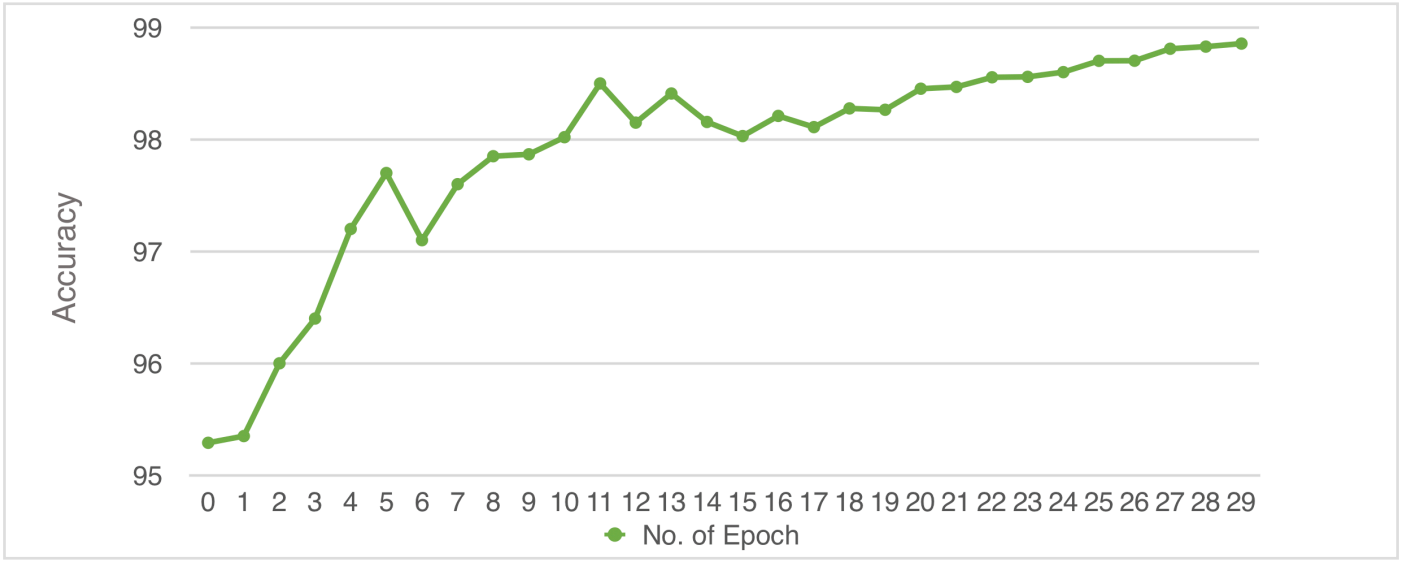


(b) Blurred Image

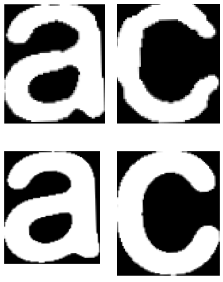
Text Detection: We will take every segmented character as input and use it in the neural network. The input format will be 32x32 pixels

Neural Network: This project uses back propagation Neural Network that has 1024 input layer neurons (each of 32x32 image pixel), 30 hidden layer neurons and 66 neurons in the output layer. Each of the individual neurons in the output layer represent 66 different characters. We are using back propagation network due to its ability to improve itself after we feed it the input vector at the time of training. With this neural network we have used Stochastic gradient descent learning algorithm for higher accuracy.

No of epoch vs Accuracy graph (Learning rate 0.5)



Back Propagation Neural Network

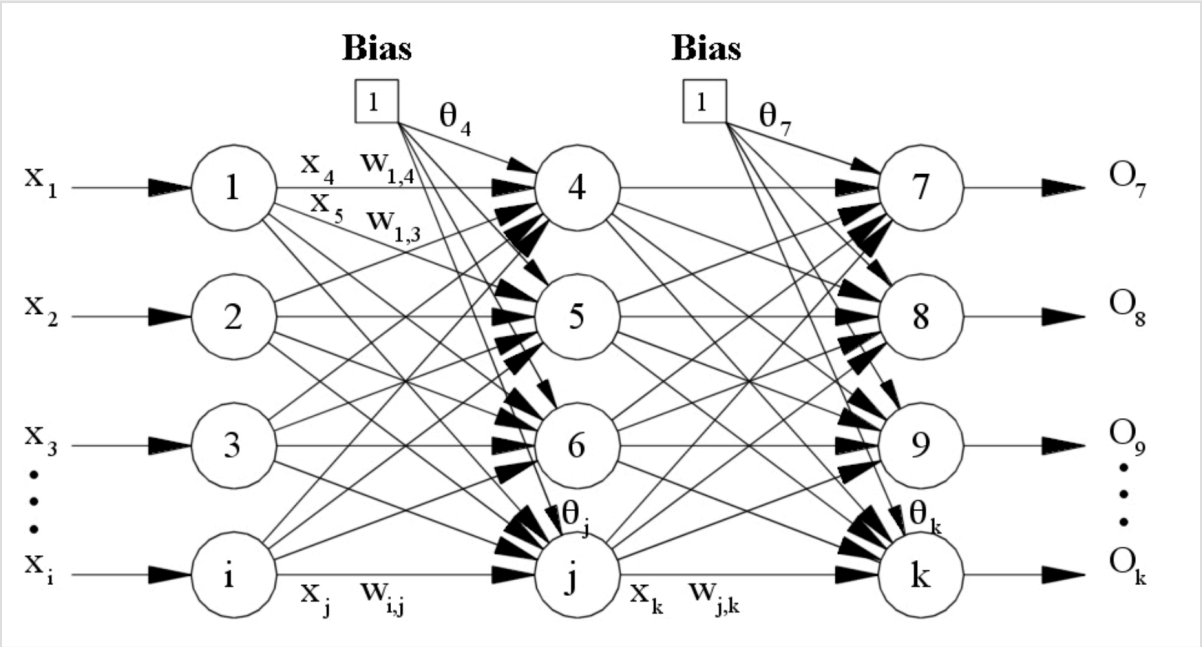


(a)Broken Image

(b) Blurred Image

Text Recognition: The neural network will compare it to the dataset it was trained with to predict the input text as its output.

Dictionary: In the project, we have a dictionary of words where there are characters that are similar and may cause error in output. If a word matches a word present in the dictionary, it is selected as output else the word that algorithm detects is given as output.



Final Output

This is a Test
for the accuracy

(a) Broken Text

This is a Test
for the accuracy

(b) Blurred image

III. Three Main Phases of process

Removal of Blur: We utilize Gaussian sharpening on the image and a 5x5 convolution mask to remove the blurriness from an image to get a sharp image.

Removal of Broken text: We use the Gaussian blur to remove the spaces that are present inside every individual character.

Removal of Skewness: We create a rectangle around the minimum enclosed area and then straighten the rectangle to remove the skewness and get image.

IV. OCR Applications

Banking: One of the most useful and important applications of OCR technology is in the banking sector, wherein it can be used to process documents such as cheques without requiring any human interaction. We can embed a cheque with a machine where the framework automatically filters the amount that is being issued and the corresponding amount is then exchanged. It has been optimized for the printed

cheque but still has a decent accuracy when dealing with handwritten cheques, which results in lowering the hold-up time in banks.

Vehicular identification using number plate: Automatic vehicular identification through recognition of the number plate using mass observation and OCR techniques at various public places such as crossroads, intersections etc. This will also result in the storage of the pictures that the cameras capture, including the number plate digits which can be utilized to identify the toll that a vehicle electronically accumulates on tolled streets. The information can also be used by the local police force as and when they require.

Healthcare: In the field of healthcare, doctors and nurses are required to maintain extensive medical records for a large number of patients. Hence, to process the printed material, the medical industry has had a very quick expansion into the use of OCR technology. With OCR, all the relevant information regarding a patient can be easily stored in an electronic database, which then allows quick and efficient extraction of the important and relevant information of patients in a medical institute.

Maintaining Accounts: Every day in an individuals day to day activated, there come purchases and receipts for those purchases. It is imperative to maintain an "account" of the receipts, both in personal life as well as for a business. Using OCR technology, the receipts can be processed and stored in an electronic database which can then be accessed to create an electronic account balance which can be updated and modified easily based on the expenditure and income of an individual or business.

CAPTCHA Bypassing: CAPTCHA is a technology that is used to determine if the test is being accessed by a human or by a software. The CAPTCHA consists of tests that normal humans can solve easily but the current technology cannot. For example, it may contain a picture with an assorted font and size and arrangement of letters with distracting elements which so that text cannot be read via OCR. Hence, even dictionary attack, where an attacker uses information of the victim to brute force attempt to determine the password, CAPTCHA will be able to stop it from happening. But current OCR frameworks are capable of filtering the noise and hence resulting in a clearer picture that will make the CAPTCHA picture be traceable by the malicious users, hence allowing them to bypass the CAPTCHA.

VI. Results and discussion

There have been multiple algorithms, methods and techniques proposed for the optical character recognition for text in images, however there are still not an adequate amount of literature surveys in this field. In this paper, we hope to further broaden the literature survey content available in the OCR field, with the hopes that our work may further inspire others to research for advancements in the field. In out paper, we begin with the discussion of the challenges that are commonly faced in OCR, following that there were points made regarding the important phases, the architecture along with the techniques and proposed algorithms of OCR. We also discuss how that any design application related to OCR we need to get a highly accurate character recognition rate, which can be obtained with detail to attention in the image, although after all this we still cannot propose a comprehensive algorithm for all the contained phases as it depends on our data sets, application specifications and more. After this, we have listed the most impactful OCR applications and a brief history of OCR in neural networks. In the current state, the

highest level of OCR has a very high text recognition accuracy, but we still believe that there are many more

References

1. Robust Text Detection and Recognition in Blurred Images
Sonia George and Noopa Jagadeesh

2. A Font style classification system for English OCR Bharath V Department of Computer science Amrita University Mysore, Karnataka, India. Bharath. 03.yadav@gmail.com N. Shobha Rani Department of Computer science Amrita University Mysore, Karnataka, India.

3. Customised OCR Correction for Historical Medical Text Paul Thompson, John McNaught and Sophia Ananiadou National Centre for Text Mining, School of Computer Science University of Manchester

4. Rafael C. Gonzalez, and Richard E. Woods, Digital Image Processing, 2nd edition, Prentice-Hall of India: New Delhi, 2002.

[5] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," IEEE Trans.

6 G. Caner and I. Haritaoglu, "Shape-DNA: Effective character restoration and enhancement for Arabic text documents," in Proc. IEEE Int. Conf. Pattern Recognit., 2010, pp. 2053–2056.

7 Y. Lou, A. L. Bertozzi, and S. Soattom, "Direct sparse deblurring," J. Math. Imag. Vis., vol. 39, no. 1, pp. 1–12, 2011.

8 C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Rotationinvariant features for multi-oriented text detection in natural images," PLoS ONE, vol. 8, no. 8, p. e70173, 2013